UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

DIEGO RAPHAEL AMANCIO

Classificação de textos com redes complexas

São Carlos
2013

DIEGO RAPHAEL AMANCIO

Classificação de textos com redes complexas

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de Doutor em Ciências.

*Área de concentração: Física Aplicada
Opção: Física Computacional*

Orientador: Prof. Dr. Luciano da Fontoura Costa

Co-orientador: Prof. Dr. Osvaldo Novais de Oliveira Jr.

Versão corrigida
(versão original na unidade que aloja o programa)

São Carlos
2013

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESSE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Ficha catalográfica elaborada pelo Serviço de Biblioteca e Informação IFSC/USP

Amancio, Diego R.

Classificação de textos com redes complexas / Diego Raphael Amancio; orientador Luciano da Fontoura Costa; co-orientador Osvaldo Novais de Oliveira Jr. – versão corrigida – São Carlos, 2013.

300 p.

Tese (Doutorado – Programa de Pós-graduação em Física Aplicada Computacional) – Instituto de Física de São Carlos, Universidade de São Paulo, 2013.

1. Redes complexas. 2. Classificação textual. 3. Processamento de textos. 4. Reconhecimento de padrões. I. Título.

Aos meus queridos pais.

AGRADECIMENTOS

Primeiramente agradeço a Deus pela vida e pela força para superar todas as dificuldades.

Aos meus pais, Marcelo Amancio e Adivaldina Amancio e irmãos Marcelo A. Amancio e Rodrigo Amancio pelo carinho, paciência e incentivo durante os anos de estudo.

Aos professores Luciano da Fontoura Costa e Osvaldo Novais Oliveira Jr. pela amizade, otimismo e competência na orientação do trabalho.

Aos amigos Thiago Silva, Ester Silva, Aquila Alexandrino e Yasmin pelos bons momentos compartilhados e também pelo carinho e amizade constantes mesmo à distância.

Aos colegas e amigos da universidade Alceu Costa, André Barbieri, Antonio Pessoti, Breno Ruy, Bruno Lopes, Camilo Akimushkin, César Comin, Cynthia Martins, Débora Gonçalves, Filipe de Andrade, Filipi Nascimento, Gustavo Vrech, Henrique Arruda, Lucas Antiqueira, Marcos Moura, Mariela Sasaki, Mary Montes, Matheus Viana, Mauro Miazaki, Paulino Villas Boas, Renato Fabbri, Renato Pimentel, Thomas Peron, Vilson Vieira e muitos outros pela amizade.

A Eduardo Altmann e Diego Rybski pela hospitalidade e convívio durante o período de estágio em Dresden e Potsdam (Alemanha).

Aos vários amigos e familiares que contribuíram diretamente ou indiretamente na minha formação profissional e pessoal.

Finalmente agradeço à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (processos 2009/02941-1 e 2010/00927-9) pelo apoio financeiro indispensável para a realização desta pesquisa.

Language is the blood of the soul into which
thoughts run and out of which they grow.
Oliver Wendell Holmes

A word fitly spoken is like apples of
gold in a setting of silver.
Proverbs 25:11 (ESV)

RESUMO

AMANCIO, D R. **Classificação de textos com redes complexas**. 2013. 300p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2013.

A classificação automática de textos em categorias pré-estabelecidas tem despertado grande interesse nos últimos anos devido à necessidade de organização do número crescente de documentos. A abordagem dominante para classificação é baseada na análise de conteúdo dos textos. Nesta tese, investigamos a aplicabilidade de atributos de estilo em tarefas tradicionais de classificação, usando a modelagem de textos como redes complexas, em que os vértices representam palavras e arestas representam relações de adjacência. Estudamos como métricas topológicas podem ser úteis no processamento de línguas naturais, sendo a tarefa de classificação apoiada por métodos de aprendizado de máquina, supervisionado e não supervisionado. Um estudo detalhado das métricas topológicas revelou que várias delas são informativas, por permitirem distinguir textos escritos em língua natural de textos com palavras distribuídas aleatoriamente. Mostramos também que a maioria das medidas de rede depende de fatores sintáticos, enquanto medidas de intermitência são mais sensíveis à semântica. Com relação à aplicabilidade da modelagem de textos como redes complexas, mostramos que existe uma dependência significativa entre estilo de autores e topologia da rede. Para a tarefa de reconhecimento de autoria de 40 romances escritos por 8 autores, uma taxa de acerto de 65% foi obtida com métricas de rede e intermitência de palavras. Ainda na análise de estilo, descobrimos que livros pertencentes ao mesmo estilo literário tendem a possuir estruturas topológicas similares. A modelagem de textos como redes também foi útil para discriminar sentidos de palavras ambíguas, a partir apenas de informação topológica dos vértices, evidenciando uma relação não trivial entre sintaxe e semântica. Para algumas palavras, a discriminação com redes complexas foi ainda melhor que a estratégia baseada em padrões de recorrência contextual de palavras polissêmicas. Os estudos desenvolvidos nesta tese confirmam que aspectos de estilo e semânticos influenciam na organização estrutural de conceitos em textos modelados como rede. Assim, a modelagem de textos como redes de adjacência de palavras pode ser útil não apenas para entender mecanismos fundamentais da linguagem, mas também para aperfeiçoar aplicações reais quando combinada com métodos tradicionais de processamento de texto.

Palavras-chave: Redes complexas. Classificação textual. Processamento de texto. Reconhecimento de padrões.

ABSTRACT

AMANCIO, D R. **Using complex networks to classify texts**. 2013. 300p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2013.

The automatic classification of texts in pre-established categories is drawing increasing interest owing to the need to organize the ever growing number of electronic documents. The prevailing approach for classification is based on analysis of textual contents. In this thesis, we investigate the applicability of attributes based on textual style using the complex network (CN) representation, where nodes represent words and edges are adjacency relations. We studied the suitability of CN measurements for natural language processing tasks, with classification being assisted by supervised and unsupervised machine learning methods. A detailed study of topological measurements in texts revealed that several measurements are informative in the sense that they are able to distinguish meaningful from shuffled texts. Moreover, most measurements depend on syntactic factors, while intermittency measurements are more sensitive to semantic factors. As for the use of the CN model in practical scenarios, there is significant correlation between authors' style and network topology. We achieved an accuracy rate of 65% in discriminating eight authors of novels with the use of network and intermittency measurements. During the stylistic analysis, we also found that books belonging to the same literary movement could be identified from their similar topological features. The network model also proved useful for disambiguating word senses. Upon employing only topological information to characterize nodes representing polysemous words, we found a strong relationship between syntax and semantics. For several words, the CN approach performed surprisingly better than the method based on recurrence patterns of neighboring words. The studies carried out in this thesis confirm that stylistic and semantic aspects play a crucial role in the structural organization of word adjacency networks. The word adjacency model investigated here might be useful not only to provide insight into the underlying mechanisms of the language, but also to enhance the performance of real applications implementing both CN and traditional approaches.

Keywords: Complex networks. Text classification. Text processing. Pattern recognition.

LISTA DE SÍMBOLOS

$\mathcal{G}$	Grafo
$\mathcal{G}^w$	Grafo ponderado
$\mathcal{G}'$	Subgrafo
$\mathcal{V}$	Conjunto de vértices
$\mathcal{E}$	Conjunto de arestas
$\mathcal{A}$	Matriz de adjacências
$\mathcal{W}$	Conjunto de pesos da rede ponderada
a_{ij}	Elemento da matriz $\mathcal{A}$
w_{ij}	Elemento da matriz $\mathcal{W}$
M	Tamanho do vocabulário
k	Grau
L	Comprimento dos caminhos mínimos
C	Coefficiente de aglomeração
γ	Expoente da lei de potência
γ_k	Expoente da lei livre de escala do grau
γ_N	Expoente da lei de Zipf para a frequência
N	Frequência absoluta das palavras no texto
s	<i>Strength</i>
d	Diâmetro
B	<i>Betweenness</i>
r	Assortatividade
E	Número total de arestas
S	Seletividade
b	Número de bigramas repetidos
T	Tempo de recorrência de uma palavra no texto
I	Intermitência

div	Diversidade
acc	Acessibilidade
$\langle \rangle$	Média calculada para palavras no livro
$\langle \rangle_2$	Média $\langle \rangle$ modificada
$\mathcal{H}$	Entropia
$\mathcal{J}$	Tamanho da janela na rede de adjacências
v	Coeficiente de variação
m_Y	Frequência do motivo Y
$\langle k_n \rangle$	Grau médio dos vizinhos
$\langle s_n \rangle$	<i>Strength</i> médio dos vizinhos
Δk_n	Desvio de grau dos vizinhos
Δs_n	Desvio de <i>strength</i> dos vizinhos
μ_{ij}	Força do objeto j na classe i
$\vec{\Gamma}_h$	Raio à distância h
R_h	Conjunto de vértices a uma distância h
$\vec{\rho}_h$	Vetor representando o conjunto de vértices a uma distância menor ou igual a h
$X^{(h)}$	Medida calculada na h -ésima hierarquia
$p(X)$	Distribuição de probabilidade da variável aleatória X
$p^{(d)}$	Proporção de <i>links</i> sintáticos errados capturados na modelagem de co-ocorrência com janela de ligação de tamanho d
$\mathcal{S}_{(\text{tr})}$	Conjunto de instâncias de treinamento
$\beta_{(\text{tr})}$	Instância de teste $\in \mathcal{S}_{(\text{tr})}$
$\mathcal{S}_{(\text{ts})}$	Conjunto de instâncias de teste
$\beta_{(\text{ts})}$	Instância de teste $\in \mathcal{S}_{(\text{ts})}$
$\mathcal{M}$	Quantidade de atributos usados na classificação
$\mathcal{F}$	Conjunto de atributos usados na classificação
$\beta^{(k)}$	Valor assumido pelo k -ésimo atributo da instância (de teste ou treinamento) β
$\mathcal{C}$	Conjunto de classes em um problema de classificação
c_i	Classe de um problema de classificação

Ω	Divergência Kullback-Leibler
ϵ_r	Raio de conexão do classificador híbrido
λ	Peso associado à abordagem de classificação topológica
P_T	Probabilidade de uma instância de teste pertencer a uma dada classe considerando a abordagem de classificação tradicional no classificador híbrido
P_H	Probabilidade de uma instância de teste pertencer a uma dada classe considerando a abordagem de classificação topológica no classificador híbrido
$\alpha(u)$	Peso associado à medida u na classificação híbrida
$f_{c_i}^{(\beta)}$	Variação normalizada observada após a inclusão da instância β no componente referente à classe c_i
$\Delta G^{(\beta)}$	Variação normalizada observada após a inclusão da instância β no componente referente à classe c_i
$\mathcal{K}$	Conjunto de <i>clusters</i>
$\mathcal{N}(c_i)$	Número de instâncias da classe c_i
$\mathcal{N}(\kappa)$	Número de instâncias no <i>cluster</i> κ
$\Lambda(\beta)$	Dissimilaridade média do objeto β ao centróide de seu <i>cluster</i>
$\zeta(\beta)$	Dissimilaridade entre o objeto β e o <i>cluster</i> mais próximo
$\rho(\kappa)$	Agrupamento mais próximo ao <i>cluster</i> κ
χ	Autovalor
ν	Autovetor
Υ	Matriz de transformação nos métodos de projeção
ϑ	Variabilidade dos dados mantida nos métodos de projeção
Φ	Dispersão dos dados no método de variáveis canônicas
W_{ij}	Oposição entre movimentos literários
skw	<i>Skewness</i> entre os movimentos literários i e j
cd	Contra-dialética
Δt	Tempo de publicação de um artigo científico
π	Distribuição de acesso no estado estacionário
τ	Rede de citação para caminhadas aleatórias
$n^{(i)}$	Número de <i>tokens</i> na partição i

$n^{(i)}(w)$	Frequência absoluta da palavra w na parte i
$f^{(i)}(w)$	Frequência relativa da palavra w
$p_i(w)$	Probabilidade de aparição da palavra w na partição i

LISTA DE ABREVIATURAS

ACC1	Sumarizador baseado na acessibilidade de primeiro nível
ACC2	Sumarizador baseado na acessibilidade de segundo nível
ACC3	Sumarizador baseado na acessibilidade de terceiro nível
ACCM	Sumarizador baseado na acessibilidade dos três primeiros níveis
AUC	Relevância de atributos baseada na área abaixo da curva
BA	Modelo Barabási-Albert
BLEU	<i>Bilingual evaluation understudy</i>
BTW	Sumarizador baseado no <i>betweenness</i>
CI	Consistência de citações
CI'	Consistência ponderada de citações
CVA	Análise de variáveis canônicas
C4.5	Classificador baseado em árvores de indução
DBPREF	Modelo de conexão preferencial ao grau e à comunidade
DN	Índice Dunn
EAT	<i>Edinburgh associative thesaurus</i>
EM	<i>Expectation maximization</i>
ER	Modelo Erdős-Rényi
GE	Modelo geográfico
κ -NN	Classificador de vizinhos mais próximos
<i>knn</i>	Regra de conexão de vizinhos mais próximos no classificador híbrido
KL	Divergência Kullback-Leibler
LHN	Leicht-Holme-Newman
LIHLA	Alinhador léxico baseado em uma heurística independente de língua
NIST	Instituto Nacional de Padrões e Tecnologia
PCA	Análise de componentes principais
PLN	Processamento de línguas naturais

PREF	Modelo de conexão preferencial à comunidade
PREFC	Modelo de conexão preferencial ao grau
SP	Sumarizador baseado nos caminhos mínimos
SWC	Silhueta simplificada
tf-idf	<i>Term frequency-inverse document frequency</i>
UNI	Modelo de conexão uniforme
UPGMA	<i>Unweighted pair group method with arithmetic mean</i>
VMS	Manuscrito Voynich
VUL	Sumarizador baseado na vulnerabilidade
WS	Modelo Watts-Strogatz

LISTA DE FIGURAS

Figura 2.1 -	Exemplos de motivos utilizados na tese	50
Figura 2.2 -	Exemplo de hierarquia em redes complexas	52
Figura 2.3 -	Exemplificação da medida de acessibilidade	54
Figura 2.4 -	Ilustração da caminhada do turista	55
Figura 3.1 -	Exemplo de espaço vetorial contendo quatro documentos descritos por duas palavras	58
Figura 3.2 -	Exemplo de modelagem de um texto como rede complexa	67
Figura 3.3 -	Exemplo de subrede semântica associativa	69
Figura 3.4 -	Histograma do tempo de retorno de palavras	76
Figura 3.5 -	Distribuição de palavras no texto	76
Figura 3.6 -	Função densidade de probabilidade $P(X)$ obtida do livro “ <i>The Adventures of Sally</i> ”	79
Figura 4.1 -	Exemplo de classificação realizada pelo algoritmo κ -NN	83
Figura 4.2 -	Relação entre a distância euclidiana e a distância Mahalanobis . .	85
Figura 4.3 -	Exemplo de árvore de decisão para classificar uma instância . . .	86
Figura 4.4 -	Exemplo de decisão baseada no algoritmo <i>Naive Bayes</i>	88
Figura 4.5 -	Fase de treinamento do algoritmo híbrido	89
Figura 4.6 -	Quantificação de relevância dos atributos usando o método AUC	95
Figura 5.1 -	Distribuição de medidas computadas para 15 diferentes versões do Novo Testamento	102
Figura 5.2 -	Exemplo de verificação da representividade das medidas	107
Figura 5.3 -	Palavras-chave para o Novo testamento em Português, Inglês, e Alemão e para o manuscrito Voynich	109
Figura 5.4 -	Visualização de algumas páginas do manuscrito Voynich	111
Figura 5.5 -	Rede obtida para o manuscrito Voynich	112
Figura 5.6 -	Distribuição das medidas para o Novo Testamento e comparação com o valor obtido para o manuscrito Voynich	117
Figura 6.1 -	Projeção CVA representando o estilo de livros	132
Figura 6.2 -	Projeção CVA ilustrando o melhor agrupamento para livros . . .	133
Figura 6.3 -	Teste de significância realizado para a medida de silhueta simplifi- cada e para o índice Dunn	133
Figura 6.4 -	Relacionamento hierárquico entre períodos literários	134
Figura 6.5 -	Dinâmica do coeficiente de aglomeração médio $\langle C \rangle$, do tamanho do vocabulário M e do coeficiente de lei de potência da distribuição de frequência γ_N	136
Figura 6.6 -	Ilustração das quantidades empregadas para definir os conceitos de oposição, <i>skewness</i> e contra-dialética	137

Figura 7.1 -	Árvore geradora mínima para a palavra ambígua <i>bar</i>	145
Figura 7.2 -	Exemplo de desambiguação para uma dada palavra ambígua . . .	147
Figura 7.3 -	Análise de variáveis canônicas para a palavra <i>ring</i>	148
Figura 7.4 -	<i>Scatter plot</i> para desambiguação da palavra <i>save</i>	150
Figura 7.5 -	Classificação com o algoritmo híbrido	152
Figura 8.1 -	Exemplo de derivação de rede de autores	164
Figura 8.2 -	Distribuição da consistência para 2.880 palavras.	170
Figura 8.3 -	Correlação de índices de consistência	172
Figura 8.4 -	Vizinhança de palavras consistentes	175
Figura 8.5 -	Comparação da consistência média para os índices de consistência em diferentes classes	177
Figura 8.6 -	Agrupamento hierárquico usando a consistência média	178
Figura 9.1 -	Coefficientes para o caminho mínimo	185
Figura 9.2 -	Análise multivariada com o PEx	186
Figura 9.3 -	Agrupamento hierárquico de traduções em espanhol	186
Figura 9.4 -	Distinção de traduções em espanhol	187
Figura 9.5 -	Agrupamento de Ward para traduções em espanhol	187
Figura 9.6 -	Acerto no mapeamento entre traduções do espanhol	190
Figura 9.7 -	Acerto no mapeamento entre traduções do inglês	191
Figura 10.1 -	Extrato de textos originais e simplificados	196
Figura 10.2 -	Distribuição de diferenças relativas	197
Figura 10.3 -	Correlação de métricas locais para caminhos mínimos	198
Figura 10.5 -	Análise de componentes principais de uma rede	200
Figura 10.6 -	Estimação de componentes de uma rede	201
Figura 10.7 -	Comparação do número de <i>clusters</i>	202
Figura 10.8 -	Análise de variáveis canônicas para simplificações naturais	203
Figura 10.9 -	Análise de variáveis canônicas para simplificações fortes	204
Figura 10.10 -	Coefficiente de Spearman entre os sumarizadores	212
Figura 11.1 -	Comparação do modelo para a rede CN	221
Figura 11.2 -	Comparação do modelo para a rede GF	222
Figura 11.3 -	Dinâmica do índice-h para modelos UNI e PREF	225
Figura 11.4 -	Dinâmica do índice-h para modelos PREFC e DBPREF	226
Figura 12.1 -	Exemplo de rede de citação	231
Figura 12.2 -	Visualização de uma rede de similaridades	234
Figura 12.3 -	Caminhadas aleatórias para redes complexas	240
Figura 12.4 -	Caminhadas aleatórias para grafenos	243
Figura 13.1 -	Rede colaborativa fictícia	247
Figura 13.2 -	Componente gigante do arXiv	248
Figura 13.3 -	Medida-f obtida para a tarefa de desambiguação hierárquica . . .	250
Figura 13.4 -	Desempenho da terceira hierarquia	250
Figura 13.5 -	Medida-f para a tarefa de desambiguação baseada na abordagem topológica	252
Figura 13.6 -	Variação da taxa de acerto Γ_H em função de λ	257
Figura 13.7 -	Distribuição de λ^* para redes colaborativas	257

Figura 13.8 - Previsão de λ^* a partir de Γ_C	261
Figura 14.1 - Rede complexa da tese	264

LISTA DE TABELAS

Tabela 2.1 -	Características das medidas de distribuição de grau $P(k)$, grau médio $\langle k \rangle$, coeficiente de aglomeração $\langle C \rangle$ e comprimento médio dos caminhos mínimos $\langle L \rangle$	41
Tabela 2.2 -	Comparação do comprimento médio dos caminhos mínimos L para palavras com a mesma frequência N no livro <i>The adventures of Sally</i>	45
Tabela 2.3 -	Comparando o <i>betweenness</i> B de palavras com frequência similares no livro <i>The adventures of Sally</i>	47
Tabela 2.4 -	Coeficiente de aglomeração de palavras com $N = 5$	48
Tabela 3.3 -	Passos de pré-processamento textual	65
Tabela 3.4 -	Comparação de intermitência de palavras com frequência similares no livro <i>The Adventures of Sally</i>	77
Tabela 4.1 -	Exemplo de computação de relevância de atributos em um problema com três atributos	94
Tabela 5.1 -	Livros em Inglês para análise de variabilidade de métricas	99
Tabela 5.2 -	Livros em Português para análise de variabilidade de métricas . .	100
Tabela 5.3 -	Valores normalizados de métricas para o conjunto de textos do Novo Testamento e romances do Português e Inglês	101
Tabela 5.4 -	Lista de medidas e propriedades	104
Tabela 5.5 -	Lista de palavras chaves obtidas para o Novo Testamento	108
Tabela 5.6 -	Condições que precisam ser satisfeitas pelas medidas para responder cada uma das questões de interesse em textos	111
Tabela 5.7 -	Lista de livros em Inglês empregados para validar a metodologia .	113
Tabela 5.8 -	Lista de livros em Português usados para validar a metodologia .	113
Tabela 5.9 -	Distância entre textos em Inglês e textos aleatórios (parte 1) . . .	113
Tabela 5.10 -	Distância entre textos em Inglês e textos aleatórios (parte 2) . . .	114
Tabela 5.11 -	Distância entre textos em Português e textos aleatórios (parte 1)	114
Tabela 5.12 -	Distância entre textos em Português e textos aleatórios (parte 2)	114
Tabela 5.13 -	Valores de medidas informativas para o manuscrito Voynich . . .	115
Tabela 5.14 -	Valores de motivos informativos para o manuscrito Voynich . . .	115
Tabela 5.15 -	Compatibilidade dos textos em Inglês com textos naturais	116
Tabela 5.16 -	Compatibilidade dos textos em Português com textos naturais . .	116
Tabela 5.17 -	Compatibilidade do VMS com línguas naturais	117
Tabela 5.18 -	Lista de palavras-chave para o manuscrito Voynich	119
Tabela 6.1 -	Corpus para os experimentos em reconhecimento de autoria . . .	123
Tabela 6.2 -	Taxa de acerto com os três melhores classificadores	124
Tabela 6.3 -	<i>p-values</i> referentes às taxas de acerto no reconhecimento de autoria	124

Tabela 6.4 -	<i>Rank</i> de atributos segundo a taxa de acerto dos classificadores . . .	125
Tabela 6.5 -	Livros usados nos experimentos de análise de variação de estilo . . .	130
Tabela 6.6 -	Livros usados nos experimentos de análise de variação de estilo com a época de publicação (continuação)	131
Tabela 6.7 -	Relação entre o melhor agrupamento baseado em fatores topológicos de redes complexas e os estilos tradicionais.	132
Tabela 6.8 -	Lista de mudanças mais significativas	135
Tabela 6.9 -	Relevância de métricas no eixo principal da técnica CVA	135
Tabela 6.10 -	Relevância de métricas no eixo secundário da técnica CVA	136
Tabela 6.11 -	Índices de oposição e <i>skewness</i>	138
Tabela 6.12 -	Índice de contra-dialética	138
Tabela 7.1 -	Lista de livros para análise das palavras ambíguas	146
Tabela 7.2 -	Lista de palavras usadas na desambiguação de sentido de palavras	146
Tabela 7.3 -	Resultados da tarefa de desambiguação de palavras	148
Tabela 7.4 -	Resultados da caracterização de palavras ambíguas usando a abordagem tradicional e com a abordagem baseada em métricas de redes complexas	149
Tabela 7.5 -	Relevância de métricas obtidas com os métodos KL e AUC	151
Tabela 7.6 -	Abordagem estrutural para discriminar sentidos de palavras	153
Tabela 7.7 -	Abordagem baseada na recorrência de palavras vizinhas para a discriminar sentidos de palavras ambíguas	154
Tabela 7.8 -	Desambiguação semântica com a caminhada do turista	156
Tabela 7.9 -	Desambiguação semântica com a caminhada do turista	157
Tabela 8.1 -	Banco de dados para a análise de consistência	163
Tabela 8.2 -	Parâmetros de ajuste da curva log-normal	169
Tabela 8.3 -	Correlação de índices de consistência	171
Tabela 8.4 -	Caracterização de classes de consistência	173
Tabela 8.5 -	Correlação entre atributos linguísticos e consistência	174
Tabela 8.6 -	Relação entre consistência e semântica	176
Tabela 9.1 -	Taxa de acerto na distinção de traduções	185
Tabela 9.2 -	Matriz de confusão para a classificação de textos traduzidos.	188
Tabela 9.3 -	Matriz de confusão para a classificação de textos traduzidos em inglês.	188
Tabela 9.4 -	Matriz de confusão para método hierárquico	190
Tabela 10.1 -	Índices de complexidade textual	195
Tabela 10.2 -	Operações de simplificação	195
Tabela 10.3 -	Estatística do <i>corpus</i> de simplificação	196
Tabela 10.4 -	Análise unidimensional de regularidade de simplificações	199
Tabela 10.5 -	Acertos para textos originais e simplificações fortes	205
Tabela 10.6 -	Acertos para textos originais e simplificações naturais	205
Tabela 10.7 -	Lista de atributos linguísticos	206
Tabela 10.8 -	Atributos linguísticos para identificar simplificação natural	207
Tabela 10.9 -	Atributos linguísticos para identificar simplificação forte	208
Tabela 10.10 -	ROUGE-1 para os métodos baseados em adjacência	210
Tabela 10.11 -	ROUGE-1 para os métodos baseados em ligações sintáticas	211

Tabela 10.12 - Correlação de Spearman entre modelos	212
Tabela 10.13 - Palavras-chave de acordo com a acessibilidade	213
Tabela 10.14 - Palavras com menor relevância para a acessibilidade	214
Tabela 11.2 - Estratégias de modelagem de redes de citações	220
Tabela 12.1 - Consistência de citações para duas áreas	238
Tabela 12.2 - Consistência modificada para duas áreas	238
Tabela 13.1 - Significância estatística da terceira hierarquia	251
Tabela 13.2 - Significância da abordagem topológica	251
Tabela 13.3 - Avaliação das medidas $\langle k_n \rangle$ e $\langle s_n \rangle$ com KL	253
Tabela 13.4 - Avaliação das medidas $\langle k_n \rangle$ e $\langle s_n \rangle$ com AUC	253
Tabela 13.5 - Acerto da desambiguação intrínseca	255
Tabela 13.6 - Comparação da estratégia híbrida com a tradicional	258
Tabela 13.7 - Relevância de medidas em redes colaborativas	259

SUMÁRIO

1	Introdução	35
1.1	Motivação e objetivos	36
1.2	Organização da tese	37
2	Redes complexas	39
2.1	Definições, história e breve revisão	39
2.2	Modelos de rede	40
2.2.1	Modelo Erdős-Rényi (ER)	42
2.2.2	Modelo Watts-Strogatz (WS)	42
2.2.3	Modelo Barabási-Albert (BA)	42
2.2.4	Modelo Geográfico (GE)	43
2.3	Redes aplicadas ao estudo da linguagem	43
2.4	Métricas de redes complexas	44
2.4.1	Grau e <i>strength</i>	44
2.4.2	Caminhos mínimos e diâmetro	45
2.4.3	<i>Betweenness</i>	46
2.4.4	Coefficiente de aglomeração	46
2.4.5	Assortatividade	48
2.4.6	Motivos	49
2.5	Medidas hierárquicas	49
2.5.1	Acessibilidade	51
2.6	Caminhada do turista	53
3	Processamento de textos	57
3.1	O modelo de espaço vetorial	57
3.2	Modelos ocultos de Markov	59
3.3	<i>Part-of-speech tagging</i>	62
3.4	Pré-processamento	65
3.5	Redes da linguagem escrita	66
3.5.1	Redes de co-ocorrência	66
3.5.2	Redes sintáticas	68

3.5.3	Redes semânticas	69
3.6	Medidas básicas	69
3.7	Entropia das palavras	70
3.8	Intermitência de palavras	73
3.9	Caracterização da estrutura textual	77
4	Reconhecimento de padrões	81
4.1	Técnicas de classificação	82
4.1.1	Vizinhos mais próximos (κ -NN)	82
4.1.2	Árvores de decisão - C4.5	85
4.1.3	<i>Naive-Bayes</i>	87
4.1.4	Classificação híbrida	88
4.2	Avaliação da classificação supervisionada	91
4.3	Avaliação da classificação não supervisionada	91
4.3.1	Silhueta simplificada (SWC)	91
4.3.2	Índice Dunn (DN)	92
4.4	Relevância de atributos	93
4.4.1	Método KL	93
4.4.2	Método AUC	94
4.5	Métodos de projeção	94
4.5.1	Análise de componentes principais (PCA)	95
4.5.2	Análise de variáveis canônicas (CVA)	96
5	Análise das propriedades estatísticas de textos	97
5.1	Metodologia de análise	98
5.1.1	Normalização das medidas	98
5.1.2	Variabilidade entre línguas e textos	99
5.2	Distinguindo livros de sequências aleatórias	102
5.3	Dependência com o estilo e com a língua	103
5.4	Representatividade das medidas	105
5.5	Deteção de palavras-chave	107
5.6	Estudo de caso: o manuscrito Voynich (VMS)	108
5.6.1	O VMS é compatível com textos aleatórios?	112
5.6.2	O VMS é compatível com textos escritos em línguas naturais? . . .	115
5.6.3	Qual língua é mais similar ao VMS?	117
5.6.4	Palavras-chave do manuscrito VMS	119
5.7	Conclusões	119
6	Análise de estilo de textos	121
6.1	Reconhecimento de autoria	122

6.1.1	Resultados da classificação	124
6.1.2	Relevância de atributos	124
6.1.3	Comparação com métodos tradicionais	127
6.2	Estilos literários de livros	128
6.2.1	Agrupamento de livros por época	129
6.2.2	Relevância de atributos	131
6.2.3	Análise de evolução do estilo	134
6.3	Conclusões	138
7	Desambiguação de palavras	141
7.1	Técnicas tradicionais de desambiguação	142
7.1.1	Abordagens tradicionais com <i>bag-of-words</i>	143
7.1.2	Abordagens tradicionais com grafos	144
7.2	Relacionamento entre topologia de redes e sentidos das palavras	145
7.3	Análise de relevância dos atributos	149
7.4	Desambiguação baseada na classificação híbrida	151
7.5	Conclusões	158
8	Consistência de palavras	161
8.1	Rede dos autores	162
8.2	Índices de consistência	162
8.2.1	Histograma de co-ocorrência de vizinhos	165
8.2.2	Similaridade cosseno	165
8.2.3	Correlação de Pearson	166
8.2.4	Índice Leicht-Holme-Newman	167
8.2.5	Índice Sørensen	167
8.2.6	Frequência de vizinhos compartilhados	168
8.3	Distribuição e inter correlação de índices de consistência	169
8.4	Correlação entre consistência e atributos linguísticos	171
8.5	Usando a consistência para identificar autoria	175
8.6	Conclusões	178
9	Tradução automática	181
9.1	Distinção de diferentes tipos de traduções	182
9.2	Avaliação de traduções automáticas com métricas hierárquicas	188
9.3	Conclusões	191
10	Complexidade textual	193
10.1	Abordagens tradicionais	193
10.2	Projeto PorSimples	195

10.3	Diferenças entre textos originais e simplificados	196
10.4	Simplicidade textual e regularidade topológica	199
10.5	Classificação de textos complexos	201
10.6	Sumarização extrativa	204
10.7	Sumarizadores extrativos automáticos	205
10.7.1	Acessibilidade	205
10.7.2	Acessibilidade e <i>strength</i>	206
10.7.3	Caminhos mínimos	207
10.7.4	<i>Betweenness</i>	208
10.7.5	Vulnerabilidade	209
10.8	Avaliação da qualidade de sumários	209
10.9	Avaliação dos sumarizadores propostos	210
10.10	Identificação de palavras-chave	212
10.11	Conclusões	213
11	Modelagem de redes de citações	217
11.1	Modelo de rede de citações	218
11.2	Efeito da visibilidade na evolução do índice-h	223
11.3	Conclusões	225
12	Cienciometria virtual	229
12.1	Redes de citação e similaridade	230
12.2	Similaridade baseada na topologia e na consistência	232
12.3	Buscas sistemáticas em redes de citações	236
12.4	Análise dos índices de consistência	237
12.5	Análise da busca sistemática	238
12.6	Cienciometria virtual	240
12.7	Conclusões	242
13	Desambiguação em redes de colaboração	245
13.1	Redes de colaboração	246
13.2	Desambiguação hierárquica	249
13.3	Desambiguação topológica	251
13.4	Desambiguação híbrida intrínseca	254
13.5	Desambiguação híbrida extrínseca	254
13.5.1	Relevância das medidas topológicas	258
13.5.2	É possível prever o valor de λ^* ?	260
13.6	Conclusões	261
14	Conclusões	263

14.1 Principais contribuições	263
14.2 Trabalhos publicados em periódicos internacionais	268
14.3 Publicações em congressos internacionais	270
14.4 Limitações das abordagens e trabalhos futuros	270
Referências	273
Apêndice A Lista de <i>stopwords</i> para o Inglês	297
Apêndice B Lista de <i>stopwords</i> para o Português	299

Capítulo 1

Introdução

O surgimento da linguagem representa uma das maiores transições na história da evolução humana. A importância da habilidade comunicativa se torna evidente com as especificidades que diferenciam humanos de outros seres vivos. Embora o processo de comunicação seja realizado de forma quase imediata, a linguagem não é um processo simples e direto. Regras de formação de estruturas em vários níveis são seguidas implicitamente para gerar uma quantidade imensa de combinações a partir de um conjunto limitado de unidades básicas, tais como letras, sílabas, palavras ou sentenças. Para entender os mecanismos de comunicação, uma grande quantidade de estudos em várias áreas da ciência já foi conduzida. Apesar do imenso esforço nessas investigações, várias questões referentes, por exemplo, à origem e à evolução da linguagem se mantêm indecifráveis. A dificuldade em estudos deste tipo se deve, provavelmente, à falta de evidências diretas para suportar uma teoria ou modelo específico. Há, por isso, vários modelos na literatura, sendo aqueles baseados em sistemas dinâmicos (1) e complexos (2) os de interesse para esta tese.

Redes complexas são usadas para modelar sistemas reais (3), cujas propriedades topológicas são caracterizadas por padrões de conectividade que não são nem totalmente aleatórios nem totalmente regulares (4). Apesar da natureza distinta dos sistemas complexos, eles compartilham propriedades topológicas. A heterogeneidade da densidade de conexões devido à distribuição livre de escala (5) e a presença de caminhos relativamente curtos entre entidades (6) representam dois dos padrões mais recorrentes nesses sistemas. Também são frequentes os padrões de organização modular (7) e hierárquica (8). Dentre os sistemas complexos com várias dessas características (3), estão os textos. A teoria de redes proposta na última década foi fundamental para consolidar a verificação da natureza complexa da linguagem, mas as primeiras indicações quantitativas comprovando tal característica datam da primeira metade do século XX. Em 1949, Zipf observou que a relação entre o *ranking* de frequência das palavras de um texto em prosa e a frequência absoluta segue uma lei de potência, independentemente do texto ou do idioma analisado (9).

Os modelos de redes complexas (2, 3, 10) representam uma ferramenta adicional para

o estudo da linguagem humana. A emergência da lei de Zipf (9), por exemplo, pôde ser explicada por uma lei de conexão preferencial (11) com os vértices (palavras) mais frequentes, de modo similar ao modelo tradicional proposto por Simon em 1955 (12). Esta mesma lei pôde ser entendida como um mecanismo de mínimo esforço (13) para comunicação entre locutores e ouvintes. Outras propriedades de sistemas complexos têm sido empregadas para estudar a linguagem, considerando fatores que vão além da frequência das palavras. Além das propriedades tradicionais de grafo, como a análise espectral (14), a estrutura de redes passou a ser utilizada. Resultados interessantes incluem a análise do *British National Corpus* em termos de rede, em que as propriedades de pequeno mundo e livre de escala emergem ao conectar palavras que aparecem próximas (15). Outro estudo refere-se à modelagem da Wordnet (16) como rede, onde conceitos são conectados de acordo com relações semânticas de antonímia e hiperonímia. Neste caso, as mesmas características livre de escala e pequeno mundo foram encontradas como consequência das relações de polissemia (17). Isso também se aplicou ao *thesaurus* da língua inglesa quando palavras com significado semelhante foram conectadas (18).

Além de resultados teóricos, o uso de redes tem se mostrado útil para aplicações de processamento de línguas naturais. Aplicações neste sentido incluem a análise de sentimentos (19–21), tradução automática (22–24) e estratégias de sumarização (25, 26). Apesar do amplo uso do modelo de redes para estudos linguísticos, uma quantidade relativamente pequena tem sido dedicada ao processamento de línguas naturais, em que as propriedades estruturais dos textos são muitas vezes desconsideradas. Diante desta lacuna, nesta tese aplicamos conceitos e metodologias de redes complexas para tarefas de classificação textual que incluem, por exemplo, a análise de estilo, o reconhecimento de autoria (27) e desambiguação lexical (28, 29). Embora aparentemente distintas, estas tarefas incorporam algum tipo de classificação textual. Além deste estudo prático, analisamos as principais medidas para identificar os fatores que cada uma das métricas das redes capta em textos. Como aplicação e generalização dos métodos desenvolvidos, mostramos que é possível aplicar a teoria de redes aliada ao processamento de línguas naturais e tarefas de classificação para o estudo de problemas cienciométricos.

1.1 Motivação e objetivos

A motivação para esta tese é introduzir conceitos e metodologias de redes complexas e séries temporais em tarefas de classificação em processamento de línguas naturais (PLN), tais como reconhecimento de autoria e desambiguação de sentido de palavras. Como a maioria das abordagens tradicionais classifica textos levando em consideração o conteúdo das palavras (30), aqui apresenta-se uma abordagem complementar, já que o tratamento

baseado em redes complexas enfatiza o modo de organização estrutural de palavras e não o seu conteúdo. O objetivo principal é verificar a existência de padrões evidenciando a dependência de características textuais (como estilo) com a forma de combinar palavras. Também verificamos até que ponto a abordagem apresentada é comparável à abordagem tradicional. Adicionalmente, aplicamos as técnicas desenvolvidas no estudo de redes de citação e colaboração de cientistas.

1.2 Organização da tese

Esta tese está organizada da seguinte maneira. No capítulo 2 apresentamos os conceitos e metodologias de redes complexas e aplicações em PLN. O capítulo 3 traz a metodologia básica para processamento de textos com redes complexas, enquanto as técnicas de reconhecimento de padrão são descritas no capítulo 4. O capítulo 5 apresenta um estudo sobre as principais medidas de redes complexas aplicadas a textos, incluindo um estudo de caso com um manuscrito cujo significado é desconhecido. O capítulo 6 trata da análise de estilo em textos, com aplicações em reconhecimento de autoria. O capítulo 7 estuda o relacionamento entre estrutura topológica de palavras e sentidos na aplicação de desambiguação de sentido de palavras. O capítulo 8 estuda a propriedade de consistência das palavras e o capítulo 9 é dedicado ao problema de avaliação de qualidade de traduções automáticas. O capítulo 10 introduz a metodologia para classificação de textos de acordo com a complexidade léxico-sintática. Aplicações das modelagens em problemas de cienciometria são apresentadas nos capítulos 11, 12 e 13, que tratam da análise de redes de citações, estratégias de citações e desambiguação de redes colaborativas. A tese se encerra no capítulo 14 com conclusões e perspectivas futuras.

Capítulo 2

Redes complexas

2.1 Definições, história e breve revisão

Desde os primórdios da ciência, a metodologia científica tem sofrido mudanças de paradigmas. Um dos primeiros modelos de pesquisa científica baseado na dúvida sistemática e na redução da investigação em partes menores foi proposto por René Descartes. Este modelo, denominado reducionista, imperou durante séculos. Foi usado com sucesso, por exemplo, para decifrar a estrutura da matéria, nos seus constituintes fundamentais. Para tratar sistemas complexos, empregam-se novos paradigmas no método científico, principalmente com a integração de ciências para resolver questões que demandam a introdução de relações não-lineares nos modelos (31). Neste contexto, a pesquisa com redes complexas (2, 3, 10) visa estudar como o relacionamento entre as partes de um sistema pode dar origem a uma funcionalidade específica.

O estudo de redes (ou grafos) remonta aos trabalhos de Leonhard Euler para a solução do famoso problema das pontes de Königsberg. Por muitos anos, acreditava-se que o relacionamento entre vértices da rede era governado por padrões de conectividade aleatórios (32, 33). Por este motivo, a pesquisa em redes foi inicialmente baseada em modelos de conexão aleatória (33). Notou-se mais tarde que esse modelo era incapaz de representar a natureza complexa de sistemas reais. Diferentemente das redes aleatórias, redes reais são preponderantemente heterogêneas quanto aos seus constituintes básicos (3). Além disso, redes reais como mapas de *software*, genomas, redes cerebrais e arquiteturas de Internet apresentam padrões que independem da natureza de suas ligações (3). Um dos comportamentos compartilhados por redes complexas modelando sistemas reais é a estrutura pequeno mundo (4) (ou *small world*), onde vértices são conectados por um número relativamente pequeno de passos. Em redes sociais, tal propriedade remete ao famoso experimento que comprova que qualquer habitante do planeta está separado por apenas seis indivíduos intermediários (6). Redes reais também são caracterizadas pela heterogeneidade

de suas conexões. Enquanto muitos vértices são fracamente conectados, outros vértices, os *hubs*, são fortemente conectados (5). Por interligarem vários vértices da rede, os *hubs* são fundamentais para garantir a resiliência da estrutura (2). Em outras palavras, quando *hubs* são eliminados, o desempenho do sistema é fortemente comprometido. Em redes sociais, por exemplo, os *hubs* desempenham papel essencial na transmissão de informação ou doenças (2).

Matematicamente, uma rede $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ é definida por um conjunto $\mathcal{V} = \{v_1, v_2, \dots, v_M\}$ de vértices e por um conjunto $\mathcal{E} = \{e_1, e_2, \dots, e_E\}$ de arestas. A conectividade é representada por uma matriz simétrica não ponderada $\mathcal{A}$. Por construção, $\mathcal{A}$ é uma matriz quadrada de tamanho M . Os elementos a_{ij} de $\mathcal{A}$ indicam se o vértice v_i está conectado ao vértice v_j quando $a_{ij} = 1$. Uma caminhada de comprimento L corresponde a uma sequência alternada $v_0 e_0 v_1 e_1 \dots e_{L-1} v_{L-1}$ de vértices e arestas de forma que $e_i = \{v_i, v_{i+1}\}$ para $i < L$. Um caminho conectando dois vértices corresponde a uma caminhada sem repetição de vértices. Neste caso, v_j é dito sucessor de v_i se existe um caminho através da rede levando v_i a v_j . Um ciclo é caracterizado por um caminho fechado onde arestas não são repetidas. Uma subrede $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ de $\mathcal{G}$ corresponde a um conjunto de vértices $\mathcal{V}'$ e arestas $\mathcal{E}'$ tal que $\mathcal{V}' \subseteq \mathcal{V}$, $\mathcal{E}' \subseteq \mathcal{E}$ e as arestas de $\mathcal{G}'$ conectam os mesmos vértices de $\mathcal{G}$. Uma rede é considerada conectada quando existe pelo menos um caminho conectando todos os pares de vértices.

Redes ponderadas são caracterizadas pela presença de arestas com pesos. A representação matricial deste tipo de rede é feita com a matriz de pesos $\mathcal{W}$. Uma rede ponderada é definida matematicamente por $\mathcal{G}^w = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, onde $\mathcal{W} = \{w_1, w_2, \dots, w_E\}$ é o conjunto de valores reais representando a intensidade do peso das arestas. Além das ponderações, as arestas podem também ter direções, dando origem a redes direcionadas. Este tipo de rede pode ser diretamente representado por matrizes $\mathcal{A}$ e $\mathcal{W}$. Assim, o elemento a_{ij} informa se a conexão $v_i \rightarrow v_j$ existe e o elemento w_{ij} armazena o peso da conexão $v_i \rightarrow v_j$.

2.2 Modelos de rede

Os principais modelos de rede e suas características estão descritos a seguir. As características de distribuição de grau $P(k)$, grau médio $\langle k \rangle$, coeficiente de aglomeração $\langle C \rangle$ e comprimento dos caminhos mais curtos $\langle L \rangle$ para os modelos Erdős-Rényi (ER), Watts-Strogatz (WS) e Barabási-Albert (BA) estão resumidos na Tabela 2.1 (ver Ref. (10)).

Tabela 2.1 – Características das medidas de distribuição de grau $P(k)$, grau médio $\langle k \rangle$, coeficiente de aglomeração $\langle C \rangle$ e comprimento médio dos caminhos mínimos $\langle L \rangle$. M é o número de vértices e p representa a probabilidade de conexão. Nas redes BA, m é o número de novas arestas adicionadas ao vértice recém introduzido. Nas redes WS, κ representa o número de vizinhos de cada vértice. A função $f(u)$ é constante se $u \ll 1$. Caso contrário, se $u \gg 1$, $f(u) = u^{-1} \ln(u)$.

Medida	Erdős-Rényi	Watts-Strogatz	Barabási-Albert
$P(k)$	$\frac{e^{-\langle k \rangle} \langle k \rangle^k}{k!}$	$\sum_{i=1}^{\min(k-\kappa, k)} \frac{k!}{i!(k-i)!} (1-p)^i p^{\kappa-i} \frac{(pk)^{k-\kappa-i}}{(k-\kappa-i)!} e^{-\kappa}$	$\sim k^{-3}$
$\langle k \rangle$	$p(M-1)$	2κ	$2m$
$\langle C \rangle$	p	$\frac{3(\kappa-1)(1-p)^3}{2(2\kappa-1)}$	$\sim M^{-0,75}$
$\langle L \rangle$	$\frac{\ln M}{\ln \langle k \rangle}$	$p^\tau f(Mp^\tau)$	$\sim \frac{\log M}{\log(\log(M))}$

2.2.1 Modelo Erdős-Rényi (ER)

O modelo de rede aleatória concebido por Erdős e Rényi (33) é um dos mais simples, em que a rede é construída da seguinte maneira. Inicialmente existem m_0 vértices, todos desconectados. Arestas conectando dois vértices distintos são adicionadas aleatoriamente de forma a não haver repetição na conexão dos vértices. Este modelo também pode ser construído com o mesmo conjunto inicial de m_0 vértices com probabilidade p de conexão entre vértices distintos. Neste caso, o número médio de conexões de cada vértice é dado por $\langle k \rangle = p(M - 1)$ e a conectividade $P(k)$ segue uma distribuição de Poisson (ver Tabela 2.1).

2.2.2 Modelo Watts-Strogatz (WS)

O processo de construção de uma rede Watts-Strogatz inicia-se com um anel contendo m_0 vértices, sendo que cada vértice é conectado com os seus κ vizinhos mais próximos à esquerda e com seus κ vizinhos mais próximos à direita. No próximo passo, cada aresta é modificada com probabilidade p para variar as propriedades da rede entre um grafo regular (no caso extremo em que $p = 0$) e um grafo aleatório (quando $p \rightarrow 1$) com poucos ciclos e comprimentos curtos para caminho entre dois vértices. O número médio de conexões é $\langle k \rangle = 2\kappa$.

2.2.3 Modelo Barabási-Albert (BA)

Muitos sistemas reais são caracterizados por uma distribuição do tipo lei de potência $P(k) \sim k^{-\gamma_k}$, evidenciado pela presença de *hubs* ou vértices com grande número de conexões. A fim de incorporar este tipo de distribuição na formação das redes, o modelo Barabási-Albert sugere que o crescimento das redes é baseado numa regra de conexão preferencial. A rede é gerada iniciando-se com um conjunto inicial de m_0 vértices. Para cada novo vértice adicionado, m novas arestas ligando vértices já existentes e o novo vértice são criadas. A conexão entre o novo vértice v_i e o vértice v_j já existente na rede ocorre com a seguinte probabilidade:

$$P(v_i \rightarrow v_j) = k_j / \sum_{v_m \in \mathcal{V}} k_m. \quad (2.1)$$

De acordo com a equação 2.1, vértices fortemente conectados são mais prováveis de receber novas conexões.

2.2.4 Modelo Geográfico (GE)

Na maioria das redes complexas, os vértices não ocupam uma posição especial no espaço. Este é o caso por exemplo das redes de palavras ou das redes de interação de proteínas. No entanto, em alguns casos, como nas redes de transporte ou na Internet, cada vértice ocupa uma posição especial na rede. O modelo geográfico é útil, portanto, para modelar estes tipos de redes onde existe uma informação espacial relevante associada a cada vértice.

Um modelo simples para geração de redes geográficas consiste em distribuir vértices aleatoriamente num espaço bidimensional conectando-os com uma probabilidade que decai com a distância entre eles, como por exemplo

$$P(v_i \rightarrow v_j) \sim e^{-\theta_{ij}/\theta_0}, \quad (2.2)$$

onde θ_{ij} é a distância geográfica dos vértices e θ_0 representa um fator de escala para o comprimento das arestas. Este modelo, que é útil para modelar, por exemplo, redes de rodovias, gera redes com distribuição Poisson.

2.3 Redes aplicadas ao estudo da linguagem

Um dos resultados estatísticos mais conhecidos para a linguagem é a lei de Zipf (9), que relaciona a frequência N das palavras com o seu *rank* i através de uma lei de potência do tipo $N \sim i^{-\gamma_N}$. Apesar da generalidade desta lei, ela tem sido pouco útil para auxiliar na investigação das origens da linguagem, já que vários mecanismos de formação de sistemas levam a uma dependência de Zipf (12, 34, 35). Além disso, nenhuma informação a respeito da organização das palavras no texto pode ser inferida, pois um texto sem estrutura sintática mantém a distribuição de frequência das palavras (36). Devido a esta e outras limitações, as diferentes propriedades da linguagem passaram a ser estudadas sob o ponto de vista de sistemas complexos, como redes. Neste tipo de modelagem, usualmente os vértices representam palavras que são conectadas de acordo com um critério particular. Dentre os critérios mais usados estão as relações de adjacência (redes de co-ocorrência), as relações sintáticas (redes de dependência sintática) e as relações semânticas (redes semânticas). Todas essas redes apresentam a característica pequeno mundo e a propriedade de livre de escala na distribuição de grau, a despeito de as ligações terem propriedades distintas. Isso sugere que a linguagem escrita apresenta organização flexível, descrita por princípios universais. Maiores detalhes a respeito dos principais tipos de redes utilizadas na linguagem escrita são fornecidos na Seção 3.5.

2.4 Métricas de redes complexas

A caracterização da topologia de redes complexas é geralmente feita a partir de medidas estruturais. Existem mais de uma centena de medidas de rede (10). Nesta tese, utilizamos algumas das mais comuns, descritas a seguir.

2.4.1 Grau e *strength*

O grau (*degree*) é uma das mais simples medidas de centralidade, já que vértices importantes (ou centrais) são aqueles com alto grau. O grau de um dado nó v_i corresponde ao número de arestas a ele associado. Em termos da matriz de adjacência $\mathcal{A}$, o grau do vértice i é

$$k_i = \sum_{j=1}^M a_{ij} = \sum_{i=1}^M a_{ij}. \quad (2.3)$$

Uma quantidade importante relacionada à medida de grau é o expoente da lei de potência γ_k da distribuição livre de escala. Se k_{min} for o menor grau para o qual a lei de potência é obedecida, então γ_k e o respectivo erro $\epsilon(\gamma_k)$ podem ser calculados analiticamente (2):

$$\gamma_k = 1 + M \left[\ln \prod_{i=1}^M \frac{k_i}{k_{min} - \frac{1}{2}} \right]^{-1}, \quad (2.4)$$

$$\epsilon(\gamma_k) = \sqrt{M} \left[\ln \prod_{i=1}^M \frac{k_i}{k_{min} - \frac{1}{2}} \right]^{-1} = \frac{\gamma_k - 1}{\sqrt{M}}. \quad (2.5)$$

Note que neste caso os pesos das arestas não são considerados. Quando a matriz $\mathcal{W}$ é usada na equação 2.3, a medida de grau é denominada como *strength*, que é dada por

$$s_i = \sum_{j=1}^M w_{ij} = \sum_{i=1}^M w_{ij}. \quad (2.6)$$

Exemplos de aplicações desta medida ocorrem em redes sociais (para verificar o grau de influência de cada indivíduo) e em redes de citações (para verificar o número de citações recebidos por um artigo científico).

2.4.2 Caminhos mínimos e diâmetro

A medida de caminhos mínimos está relacionada com a medida de centralidade (*close-ness centrality* (10)). Seja $\text{dist}(i, j)$ o comprimento do menor caminho ligando dois vértices v_i e v_j . O comprimento médio dos caminhos mínimos do vértice v_i , denotado como L_i , é calculado como

$$L_i = \frac{1}{M-1} \sum_{j=1}^M \text{dist}(i, j). \quad (2.7)$$

O fator $M-1$ é usado no denominador, uma vez que a influência do próprio vértice nos caminhos mínimos é desconsiderada ($\text{dist}(i, i) = 0$). Além da média obtida na equação 2.7, outra medida importante referente às distâncias geodésicas é o diâmetro d , definido como $d = \max \text{dist}(i, j)$.

A interpretação desta medida em nível textual pode ser realizada comparando-se as palavras com altos e baixos valores de caminhos mínimos. A Tabela 2.2 compara a média dos caminhos mínimos para palavras de mesma frequência no livro “*The adventures of Sally*”. As palavras com menores valores de L incluem os personagens “Sally” ($L = 2, 35$ e $N = 347$) e “Fillmore” ($L = 2, 45$ e $N = 349$), além das palavras de alta frequência, tais como “say” ($L = 2, 45$ e $N = 349$), “good” ($L = 2, 46$, $N = 107$) e “man” ($L = 2, 50$, $N = 193$). As palavras assumindo os maiores valores de L foram: “white-clad” ($L = 6, 33$ e $N = 1$), “affability” ($L = 6, 31$ e $N = 1$), “whirl” ($L = 5, 89$ e $N = 1$), “jazz” ($L = 5, 87$ e $N = 1$) e “war-aims” ($L = 5, 84$ e $N = 1$). Interessantemente, todas estas cinco palavras apareceram apenas uma vez no texto, indicando que uma das razões pela qual uma palavra assume um alto valor de L é a baixa frequência N . No entanto, L não é uma consequência apenas da frequência N das palavras, uma vez que palavras pouco frequentes também podem assumir baixos valores de L . De acordo com os resultados da Tabela 2.2, a frequência não possui uma alta influência em L , já que a correlação $\text{Corr}(L, N) = -0,36$.

Tabela 2.2 – Comparação do comprimento médio dos caminhos mínimos L para palavras com a mesma frequência N no livro “*The adventures of Sally*”. Para um dado valor de N , L pode assumir uma ampla gama de valores como consequência da dependência de L na conectividade dos vizinhos.

Palavra	N_i	L_i	Palavra	N_i	L_i
<i>red</i>	5	3,71	<i>earth</i>	5	2,99
<i>shudder</i>	4	3,97	<i>lucky</i>	4	3,00
<i>Maxwell</i>	3	5,55	<i>funny</i>	3	3,10
<i>dark</i>	2	5,15	<i>kiss</i>	2	3,08
<i>affability</i>	1	6,34	<i>repose</i>	1	3,11

2.4.3 *Betweenness*

Outro conceito de centralidade, diferente do grau, é denominado *betweenness* (B). Esta quantidade associa altos valores de centralidade para vértices bastante acessados por caminhos mínimos. Vértices com altos valores de *betweenness* possuem considerável influência na rede devido ao controle da distribuição de informação. Em sistemas de informação, vértices com altos valores de *betweenness* são fundamentais para manter a estrutura da rede, pois sua remoção causa grande impacto na eficiência de distribuição de informação (10). Para calcular o *betweenness* quando há mais de um caminho mínimo entre dois vértices, associa-se um peso inversamente proporcional ao número de caminhos mínimos entre eles. Deste modo, se existem n_L caminhos mínimos entre um par de vértices, então cada um dos caminhos é ponderado com um peso igual a n_L^{-1} . O *betweenness* é então calculado como a soma dos pesos associados aos caminhos mínimos. Nesta tese, a medida de *betweenness* é definida de maneira normalizada de modo a eliminar correlações com outras medidas (p.e. o grau k). Se η_{sit} é o número de caminhos geodésicos de v_s a v_t que passam por v_i , e η_{st} é o número de caminhos mínimos que conectam v_s a v_t , então o *betweenness* é dado por:

$$B_i = \frac{1}{M^2} \sum_{s=1}^M \sum_{t=1}^M \frac{\eta_{sit}}{\eta_{st}}. \quad (2.8)$$

Com esta normalização o valor de B fica limitado ao intervalo $[0, 1]$.

No contexto de análise textual, palavras de alta frequência tendem a possuir altos valores de B . No entanto, algumas palavras podem desempenhar um papel de ponto de articulação ligando conceitos relacionados a comunidades distintas. Para ilustrar este conceito, a Tabela 2.3 mostra que as palavras com valores similares de N podem assumir diferentes valores de B . Uma comparação entre as colunas da esquerda e da direita sugere que as palavras com altos valores de B conectam conceitos de comunidades semânticas distintas por causa da sua alta probabilidade em aparecer em vários contextos. Portanto, de modo análogo ao coeficiente de aglomeração C , o *betweenness* parece quantificar a variedade de contextos em que uma palavra pode aparecer. No entanto, note que B utiliza os padrões de conectividade global, enquanto que C utiliza apenas informação local.

2.4.4 Coeficiente de aglomeração

Vértices de redes também são caracterizados por atributos da sua vizinhança. O coeficiente de aglomeração (*clustering coefficient*) quantifica como o subgrafo formado pela vizinhança de um vértice se aproxima de um clique (i.e., um subgrafo totalmente conectado). Vértices com máximo valor de aglomeração satisfazem a propriedade de

Tabela 2.3 – Comparando o *betweenness* B de palavras com frequência similares no livro “*The adventures of Sally*”. Para um dado valor de frequência N , o *betweenness* pode variar bastante. Mesmo palavras de baixa frequência podem assumir altos valores de *betweenness* caso elas apareçam em vários contextos.

Palavra	N	B	Palavra	N	B
<i>say</i>	349	745.634	<i>Sally</i>	347	1.192.881
<i>know</i>	143	243.357	<i>Fillmore</i>	138	393.955
<i>tell</i>	65	53.904	<i>Gerald</i>	62	108.528
<i>allow</i>	20	15.816	<i>Roville</i>	21	32.449
<i>heaven</i>	10	1.147	<i>second</i>	10	22.004
<i>rugger</i>	5	855	<i>worthy</i>	5	10.503
<i>fish</i>	4	174	<i>spectator</i>	4	14.746
<i>paper-knife</i>	3	233	<i>group</i>	3	8.320
<i>worship</i>	2	44	<i>sell</i>	2	8.346
<i>thaw</i>	1	11	<i>price</i>	1	8.295

transitividade de vizinhos: se dois vértices v_j e v_k são vizinhos de v_i , então v_j e v_k também são vizinhos entre si. Matematicamente, o coeficiente de aglomeração C_i do vértice v_i é definido como a razão entre a quantidade Ψ_i de ligações entre vizinhos de v_i , dado por

$$\Psi_i = \sum_{j=1}^M \sum_{l=1}^M a_{ij} a_{jl} a_{li} \quad (2.9)$$

e o máximo número possível de *links* entre eles, quando todos estão ligados entre si, formando um clique de $k_i(k_i - 1)$ conexões. Por definição, quando o vértice possui menos de dois vizinhos, seu coeficiente de aglomeração é igual a zero. Em resumo:

$$C_i = \begin{cases} 2\Psi_i/(k_i^2 - k_i) & \text{para } k_i > 1, \\ 0 & \text{para } k_i \leq 1. \end{cases} \quad (2.10)$$

Em várias redes, existe clara relação entre o coeficiente de aglomeração C e o grau k . As formas mais comuns de correlações ocorrem com $C \sim k^{-3/4}$ (2).

A fim de verificar a relação do coeficiente de aglomeração com a função das palavras em textos, isolamos as palavras (e seus vértices vizinhos) com os maiores e menores valores de coeficiente de aglomeração no livro “*The Adventures of Sally*”, de P. G. Wodehouse. Estas palavras estão mostradas na Tabela 2.4. A partir da definição, espera-se que as palavras assumindo os maiores valores de C possuam vários vizinhos conectados entre si. Este é o caso por exemplo das palavras “*sand*” e “*excitement*”. Por outro lado, palavras cujos vizinhos não estão relacionados entre si apresentam baixos valores de C (p.e., não existe ligação entre os vizinhos de “*full*” ou entre os vizinhos de “*high*”. Qualitativamente, o coeficiente de aglomeração verifica como as palavras estão conectadas a contextos específicos. De fato, as palavras “*sand*” e “*excitement*” tendem a aparecer em contextos mais restritos, enquanto que “*full*” tende a ocorrer em vários contextos. Portanto, a

medida de aglomeração parece ser útil para quantificar a especificidade de contexto das palavras.

Tabela 2.4 – Palavras do livro “*The adventures of Sally*” que apresentam os maiores e menores valores de coeficiente de aglomeração (o coeficiente de aglomeração médio $\langle C \rangle = 0,085$ no livro), para palavras com $N_i = 5$. As cinco palavras com $C_i = 0$ foram selecionadas aleatoriamente dentre as 18 palavras com $N_i = 5$ e $C_i = 0$.

Palavra	Vizinhos	C
<i>shortly</i>	<i>twelve, see, say, Sally, news</i> <i>never, heaven, find, enter e Carmylle</i>	0,27
<i>excitement</i>	<i>thing, suppressed, Sally, mince, last,</i> <i>can, come, bristle, brief e apart</i>	0,25
<i>sand</i>	<i>watch, want, sit, shuffling, seat,</i> <i>here, golden, first, dard e Roville</i>	0,18
<i>nose</i>	<i>voice, tip, sort, smut, smooth,</i> <i>Sally, oh, glance, tell e come</i>	0,18
<i>country</i>	<i>time, still, somewhere, say, place</i> <i>may, happen, great e glorious</i>	0,18
<i>startle</i>	<i>shy, seem, mill, little, gratify, gather,</i> <i>first, everyday, displeased e conderably</i>	0,00
<i>high</i>	<i>recess, mouth, motive, lapse, figure</i> <i>even, disposal, critical, collar e check</i>	0,00
<i>gold</i>	<i>voice, spin, pencil, loan, knob</i> <i>information, high, heavy, frame e buy</i>	0,00
<i>gift</i>	<i>tongue, take, sort, potential e mean</i> <i>few, easily, compensating, blessing e acquire</i>	0,00
<i>full</i>	<i>tuesday, peal, later, home e happy</i> <i>gratitude, gleaming, glance, color e battle</i>	0,00

2.4.5 Assortatividade

Em algumas redes vértices possuem uma classificação natural. Por exemplo, em redes sociais pessoas são classificadas por idade, sexo ou raça. Na análise dos padrões de conectividade das redes muitas vezes há interesse em saber se as ligações são estabelecidas preferencialmente entre vértices de uma mesma classe ou entre vértices de classes distintas. Este tipo de análise é feita com a medida de assortatividade. Um dos tipos de assortatividade mais estudados em redes complexas quantifica a correlação dos graus de vértices conectados. Esta correlação de grau pode ser expressa em termos da probabilidade condicional de que um vértice com grau k esteja conectado com um vértice de grau k' :

$$P(k'|k) = \frac{\langle k \rangle P(k, k')}{kP(k)}, \quad (2.11)$$

onde $P(k, k')$ corresponde à probabilidade de que uma aresta conecte vértices com graus k e k' . Em redes direcionadas, k denota o grau do vértice fonte da aresta e k' é o grau do vértice alvo. Neste caso, o grau considerado pode ser o grau de entrada, ou grau de saída ou a soma dos anteriores. Estendendo o conceito para grafos com arestas ponderadas, o grau k pode ser substituído pelo *strength* s (definido na equação 2.11).

A correlação de grau também pode ser calculada considerando o coeficiente de correlação de Pearson dos graus dos vértices conectados:

$$r = \frac{E^{-1} \sum_{j>i} k_i k_j a_{ij} - \left[E^{-1} \sum_{j>i} \frac{1}{2} (k_i + k_j) a_{ij} \right]^2}{E^{-1} \sum_{j>i} \frac{1}{2} (k_i^2 + k_j^2) a_{ij} - \left[E^{-1} \sum_{j>i} \frac{1}{2} (k_i + k_j) a_{ij} \right]^2}, \quad (2.12)$$

onde $E = 1/2 \sum k_i$ é o número total de arestas da rede. Quando $r < 0$, a rede é denominada não assortativa. Caso contrário, se $r > 0$, a rede é denominada assortativa, i.e., vértices de grau alto tendem a ser conectar com outros vértices de grau elevado.

2.4.6 Motivos

A topologia das redes complexas também pode ser caracterizada em termos de motivos, que correspondem a padrões de conectividade expressos em pequenos subgrafos chamados motivos (10). Em vários sistemas reais, cada motivo está geralmente relacionado com uma função do sistema. Este tipo de representação é útil na descoberta de padrões em modelos de redes complexas, mas sua detecção é computacionalmente cara. Por esta razão, restringimos a análise de motivos para os subgrafos ilustrados na Figura 2.1. Para fins de notação, a frequência do motivo Y será denotada como m_Y .

2.5 Medidas hierárquicas

As medidas hierárquicas representam uma extensão das medidas convencionais. Ao contrário destas últimas, as medidas hierárquicas analisam vizinhanças mais distantes. Seja $R_1(i)$ o conjunto de vizinhos imediatos de um dado vértice v_i . Ou seja, $R_1(i)$ representa o conjunto de vértices intermediários na transmissão da informação do vértice v_i a um dado vértice v_j , $v_j \notin R_1(i)$. Portanto, pode-se dizer que tal conjunto é capaz de formar “arestas virtuais” entre v_i e v_j . Matematicamente, define-se $R_1(i)$ a partir do vetor $\vec{\nu}_1(i)$

$$\vec{\nu}_1(i) = \mathcal{W} \cdot \vec{\nu}(i),$$

Figura 2.1 – Exemplo de motivos utilizados na tese. A análise com motivos restringiu-se apenas aos subgrafos contendo três vértices devido à alta complexidade de detecção de motivos contendo mais vértices.

onde

$$\begin{aligned}\vec{\nu}(\mathbf{1}) &= \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \end{pmatrix} \\ \vec{\nu}(\mathbf{2}) &= \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \end{pmatrix} \\ \vec{\nu}(\mathbf{3}) &= \begin{pmatrix} 0 & 0 & 1 & \cdots & 0 \end{pmatrix}.\end{aligned}$$

Usando a função delta de Kronecker sobre o vetor $\vec{\nu}_1$, definimos a quantidade $\vec{\rho}_1(i)$ como:

$$\vec{\rho}_1(i) = \delta(\vec{\nu}_1(i)). \quad (2.13)$$

Se um vértice $v_j \in R_1(i)$, então $\vec{\rho}_1(i) = 1$. Caso contrário, $\vec{\rho}_1(i) = 0$. De maneira análoga ao cálculo de $\vec{\nu}_1(i)$, pode-se estender seu conceito para o cálculo de $\vec{\nu}_h(i)$, que representa o conjunto de vértices dentro do raio de distância h :

$$\vec{\nu}_h(i) = \mathcal{W}^h \vec{\nu}(i). \quad (2.14)$$

Assim, analogamente à equação 2.13, pode-se definir $\vec{\rho}_h(i)$ como o conjunto de vértices a uma distância máxima de h arestas do vértice i , como descrito na equação 2.15.

$$\vec{\rho}_h(i) = \delta\left(\sum_{k=1}^h \vec{\rho}_k(i) + \vec{\nu}(i)\right). \quad (2.15)$$

Com estas definições, é possível conhecer o raio à distância h do vértice v_i pelo cálculo do vetor

$$\vec{\Gamma}_h(i) = \vec{\rho}_h(i) - \vec{\rho}_{h-1}(i), \quad (2.16)$$

cujos elementos não nulos representam o conjunto $v_i \in R_h(i)$. Dessa forma, as medidas convencionais são calculadas colapsando os vértices dentro do raio R_h como um único vértice, desprezando as arestas e vértices internos. Um exemplo de cálculo de hierarquias é mostrado na Figura 2.2. Nesta tese, usamos a segunda, terceira e quarta hierarquias para as medidas de grau ($k^{(2)}$, $k^{(3)}$ e $k^{(4)}$), coeficiente de aglomeração ($C^{(2)}$, $C^{(3)}$ e $C^{(4)}$) e *strength* ($s^{(2)}$, $s^{(3)}$ e $s^{(4)}$).

2.5.1 Acessibilidade

Diferentemente de outras medidas de centralidade (grau e *betweenness*), a acessibilidade não é obtida a partir do número de caminhos ou ligações que passam por um dado vértice. Esta medida é definida como uma normalização da medida de diversidade *div* (37). Para definir esta quantidade, considere que $P_h(i, j)$ representa a probabilidade de se alcançar um vértice j através de uma caminhada aleatória auto-excludente (37) de comprimento h

Figura 2.2 – Exemplo de cálculo de hierarquias para o vértice central em amarelo na rede em (a). Os vértices em vermelho pertencem ao conjunto R_1 (“amarelo”), enquanto os vértices em azul pertencem ao conjunto R_2 (“amarelo”). Para o cálculo das medidas hierárquicas, vértices dentro do raio considerado são colapsados num único vértice mantendo as arestas externas. Em (b), os vizinhos do vértice amarelo são colapsados em um único vértice, enquanto em (c) os vizinhos dos vizinhos são colapsados num único vértice. Desta forma, os graus hierárquicos no segundo e terceiro níveis são $k^{(2)}(\text{“amarelo”}) = 9$ e $k^{(3)}(\text{“amarelo”}) = 11$, respectivamente.

a partir do vértice de referência i . Neste caso, são considerados os caminhos a partir de do vértice analisado para cada um dos $n_h(i)$ vértices no anel concêntrico de distância h . Considerando esta notação, a medida de diversidade é dada por

$$\text{div}^{(h)}(i) = - \sum P_h(i, j) \ln P_h(i, j). \quad (2.17)$$

A partir da equação 2.17 a acessibilidade pode ser obtida diretamente de

$$\text{acc}^{(h)}(i) = \exp(\text{div}^{(h)}(i)). \quad (2.18)$$

Com esta definição, $0 \leq \text{acc}^{(h)}(i) \leq n_h(i)$, onde $n_h(i)$ é o número de vértices no anel h . Um exemplo de cálculo desta medida é fornecido na Figura 2.3. Além de ser útil para melhorar a caracterização da rede, esta medida têm sido útil para detectar bordas de redes complexas (37).

2.6 Caminhada do turista

A caminhada do turista pode ser caracterizada por uma partícula (turista) visitando vértices em uma mapa n -dimensional. A cada intervalo de tempo discreto, o turista segue uma regra simples determinística: ele visita o vizinho mais próximo ainda não visitado nos μ passos anteriores. Em outras palavras, o turista realiza uma caminhada determinística parcialmente auto-excludente, onde o fator autor-excludente é limitado pela janela de memória $\mu - 1$. Esta quantidade pode ser entendida como uma força repulsiva emitida pelos vizinhos visitados, prevenindo o turista de visitá-los neste intervalo. Portanto, a trajetória nunca intersecciona ela mesma dentro da janela de memória. Apesar de representar uma regra simples, alguns estudos mostram que este movimento dinâmico apresenta comportamento complexo sempre que $\mu > 1$ (38). Cada caminhada do turista pode ser decomposta em dois termos: (i) a *parte inicial transiente* de comprimento t , e (ii) um *ciclo* (atrator) com período c . A Figura 2.4, ilustra uma caminhada do turista com $\mu = 1$, i.e., o turista sempre visita o vizinho mais próximo. Neste caso, podemos perceber que o comprimento transiente é $t = 2$ e o comprimento do ciclo é $c = 8$.

Figura 2.3 – Probabilidade de transições de comprimento $h = 2$ considerando como referência o vértice 1 para duas configurações. Na primeira, consideram-se apenas as arestas em vermelho e na segunda consideram-se ambas as arestas azuis e vermelhas. A primeira configuração apresenta a maior diversidade de distribuição de transições de probabilidade $P^{(\text{verm})}_2(1, j)$, resultando num maior valor de acessibilidade $\text{acc}^{(h, \text{verm})}_2(1) = 5$. Quando as arestas azuis são adicionadas na segunda configuração, observa-se uma tendência aparente de acesso aos vértices 8 e 9, resultando num número menor de vértices efetivamente acessados $\text{acc}^{(h, \text{azul})}_2(1)$.

Figura 2.4 – Ilustração de uma caminhada do turista com $\mu = 1$. Os vértices em amarelo representam os vértices não visitados. Os pontos em vermelho e verde indicam os vértices visitados. O verde representa a parte transiente da caminhada e o vermelho representa o ciclo da caminhada.

Capítulo 3

Processamento de textos

Este capítulo trata essencialmente do processo de transformação de textos em redes complexas. Embora nesta tese empregemos essencialmente redes de co-ocorrência de palavras, apresentamos também outros tipos de redes modelando características distintas da linguagem. Também neste capítulo apresentamos medidas básicas de textos. Por fim, descrevemos métodos para caracterizar livros a partir das métricas para palavras.

3.1 O modelo de espaço vetorial

Um dos modelos mais simples de representação textual é o chamado modelo de espaço vetorial para similaridade semântica. Documentos completos ou porções dos documentos são representados em um espaço em que cada dimensão representa uma palavra. Desta forma, dois documentos são considerados semelhantes semanticamente se eles compartilham várias palavras. Para ilustrar o modelo, considere a Figura 3.1. Nesta figura, quatro documentos são representados com as seguintes coordenadas $\vec{d}_1 = [0, 25 \ 0, 98]^T$, $\vec{d}_2 = [0, 75 \ 0, 60]^T$, $\vec{d}_3 = [0, 95 \ 0, 25]^T$ e $\vec{d}_4 = [0, 70 \ 0, 75]^T$. Os documentos d_1 e d_3 são semanticamente pouco similares já que, enquanto d_1 é bem representado pela “PALAVRA 2” e possui poucas ocorrências da “PALAVRA 1”, o documento d_3 apresenta o padrão inverso. Para quantificar a similaridade semântica entre documentos utiliza-se geralmente a medida cosseno (ou coeficiente de correlação normalizado):

$$\cos(\vec{d}_1, \vec{d}_2) = \sum_{i=1}^M d_1(i)d_2(i) / \sqrt{\sum_{i=1}^M d_1^2(i)d_2^2(i)}, \quad (3.1)$$

onde M representa o tamanho do vocabulário utilizado para descrever o espaço vetorial (no exemplo da Figura 3.1, $M = 2$). Note que, de acordo com a equação 3.1, quanto maior o valor de $\cos(\vec{d}_1, \vec{d}_2)$, maior é a similaridade semântica entre os documentos. No

Figura 3.1 – Exemplo de espaço vetorial contendo duas dimensões (palavras). As dimensões correspondem às palavras “PALAVRA 1” e “PALAVRA 2”. A posição dos documentos d_1 , d_2 , d_3 e $d_?$ é apresentada em notação vetorial.

exemplo da Figura 3.1, o documento mais similar a $d_?$ é d_2 pois $\cos(\vec{d_?}, \vec{d_2}) > \cos(\vec{d_?}, \vec{d_1})$ e $\cos(\vec{d_?}, \vec{d_2}) > \cos(\vec{d_?}, \vec{d_3})$.

Para fins de comparação, a distância euclidiana para vetores normalizados pode ser considerada equivalente à medida cosseno em sistemas de *rank*. Isto pode ser concluído observando-se que a distância euclidiana entre dois vetores $\vec{x}$ e $\vec{y}$ pode ser escrita como

$$\begin{aligned}
 \|\vec{x} - \vec{y}\|^2 &= \sum_i (x_i - y_i)^2 \\
 &= \sum_i x_i^2 - 2 \sum_i x_i y_i + \sum_i y_i^2 \\
 &= 1 - 2 \sum_i x_i y_i + 1 \\
 &= 2 \left(1 - \sum_i x_i y_i \right) \\
 &= 2 \left(1 - \cos(\vec{x}, \vec{y}) \right).
 \end{aligned} \tag{3.2}$$

Consequentemente

$$\cos(\vec{x}, \vec{y}) > \cos(\vec{x}, \vec{z}) \Leftrightarrow \|\vec{x} - \vec{y}\| < \|\vec{x} - \vec{z}\|. \tag{3.3}$$

3.2 Modelos ocultos de Markov

Um modelo oculto de Markov é uma função probabilística de um processo de Markov em que um dado sistema é modelado com parâmetros desconhecidos a serem determinados pela análise de parâmetros observáveis (39). Modelos de Markov são especialmente utilizados em aplicação de processamento de áudio e *part-of-speech tagging*, que é discutido na Seção 3.3.

Modelos de Markov modelam sequências de variáveis aleatórias independentes cujo valor atual depende apenas do valor assumido no estado anterior. Um modelo oculto de Markov é especificado pela quintúpla (S, K, Π, A, G) , cujos significados são apresentados na Tabela 3.1.

Tabela 3.1 – Definição dos elementos de um modelo oculto de Markov.

Definição	Notação
Conjunto de estados	$S = \{\Sigma_1, \dots, \Sigma_E\}$
Alfabeto de saída	$K = \{\mathcal{K}_1, \dots, \mathcal{K}_M\}$
Probabilidades dos estados iniciais	$\Pi = \{\pi_i\}, i \in S$
Probabilidades de transição	$A = \{A_{ij}\}, i, j \in S$
Probabilidades de emissão de símbolos	$G = \{g_{ijk}\}, i, j \in S, k \in K$
Sequência de estados	$\mathbb{X} = (\mathbb{X}_1, \dots, \mathbb{X}_{T+1}) \quad \mathbb{X}_t : S \mapsto \{1 \dots E\}$
Sequência de saída	$O = (o_1, \dots, o_T) \quad o_t \in K$

Considere $\mathbb{X} = (\mathbb{X}_1, \dots, \mathbb{X}_T)$ uma sequência de variáveis aleatórias assumindo valores no conjunto finito $\Sigma = \{\Sigma_1, \dots, \Sigma_E\}$. As seguintes propriedades são satisfeitas por um sistema Markoviano:

- **Dependência de curto alcance:** a probabilidade de transição do estado atual (instante t) para o próximo estado (instante $t + 1$) depende apenas do estado visitado anteriormente. Em outras palavras, $P(\mathbb{X}_{t+1} = \Sigma_k | \mathbb{X}_1, \dots, \mathbb{X}_t) = P(\mathbb{X}_{t+1} = \Sigma_k | \mathbb{X}_t)$.
- **Invariância no tempo:** $P(\mathbb{X}_{t+1} = \Sigma_k | \mathbb{X}_1, \dots, \mathbb{X}_t) = P(\mathbb{X}_2 = \Sigma_k | \mathbb{X}_1)$.

A transição de estados é descrita por uma matriz estocástica $A : A_{ij} = P(\mathbb{X}_{t+1} = \Sigma_j | \mathbb{X}_t = \Sigma_i)$. Neste caso, $\sum_j A_{ij} = 1$. Dada uma sequência observada $O = (o_1, \dots, o_T)$ e um modelo $\mu = (A, G, \Pi)$, é possível calcular a probabilidade $P(O | \mu)$ da seguinte maneira. Dada a sequência de estados $\mathbb{X} = (\mathbb{X}_1, \dots, \mathbb{X}_{T+1})$, $P(O | \mathbb{X}, \mu)$ e $P(\mathbb{X} | \mu)$ podem ser escritos como:

$$\begin{aligned}
 P(O | \mathbb{X}, \mu) &= \prod_{t=1}^T P(o_t | \mathbb{X}_t, \mathbb{X}_{t+1}, \mu) \\
 &= g_{\mathbb{X}_1 \mathbb{X}_2 o_1} g_{\mathbb{X}_2 \mathbb{X}_3 o_2} \dots g_{\mathbb{X}_T \mathbb{X}_{T+1} o_T}
 \end{aligned} \tag{3.4}$$

$$P(\mathbb{X} | \mu) = \pi_{\mathbb{X}_1} A_{\mathbb{X}_1 \mathbb{X}_2} A_{\mathbb{X}_2 \mathbb{X}_3} \dots A_{\mathbb{X}_T \mathbb{X}_{T+1}} \tag{3.5}$$

Uma vez que $P(O, \mathbb{X}|\mu) = P(O|\mathbb{X}, \mu)P(\mathbb{X}|\mu)$, $P(O|\mu)$ pode ser obtida em função dos parâmetros do sistema:

$$\begin{aligned} P(O|\mu) &= \sum_{\mathbb{X}} P(O|\mathbb{X}, \mu)P(\mathbb{X}|\mu) \\ &= \sum_{\mathbb{X}_1 \dots \mathbb{X}_{T+1}} \pi_{\mathbb{X}_1} \prod_{t=1}^T a_{\mathbb{X}_t \mathbb{X}_{t+1}} b_{\mathbb{X}_t \mathbb{X}_{t+1} o_t}. \end{aligned} \quad (3.6)$$

Embora a avaliação da quantidade obtida na equação 3.6 possa ser computada diretamente dos parâmetros do sistema markoviano, sua computação é ineficiente. Para evitar esta complexidade, uma técnica de programação dinâmica (40) deve ser utilizada. A forma de armazenamento da computação anterior por programação dinâmica pode ser realizada através de dois métodos: o ascendente e o descendente. No primeiro caso, é necessário definir uma variável auxiliar α da seguinte maneira:

$$\alpha_i(t) = P(o_1 o_2 o_3 \dots o_{t-1}, \mathbb{X}_t = i | \mu). \quad (3.7)$$

Esta variável, que é armazenada na matriz (Σ_i, t) , modela a probabilidade de o sistema estar no estado Σ_i no instante t , dada as observações $o_1 o_2 o_3 \dots o_{t-1}$. Para fins de programação dinâmica, a inicialização, a indução e a probabilidade de alcançar o estado final são dadas por:

- **Inicialização:** definição dos valores da variável α_i para $t = 1$:

$$\alpha_i(1) = 1, \text{ para } 1 \leq i \leq E. \quad (3.8)$$

- **Indução:** definição dos valores de $\alpha(t+1)$ em função de $\alpha(t)$:

$$\alpha_j(t+1) = \sum_{i=1}^E \alpha_i(t) A_{ij} g_{ij o_t}. \quad (3.9)$$

- **Probabilidade final:** probabilidade de alcançar o estado final dada as observações $o_1 o_2 o_3 \dots o_{t-1}$:

$$P(O|\mu) = \sum_{i=1}^E \alpha_i(T+1). \quad (3.10)$$

Outra possibilidade é usar um procedimento descendente em contraposição ao procedimento ascendente. Neste caso, define-se a variável $\xi_i(t)$ que representa a probabilidade de observar o resto da sequência assumindo que no instante t o sistema encontra-se no estado Σ_i . Para isto, a variável ξ é definida como:

$$\xi_i(t) = P(o_t \dots o_T | \mathbb{X}_t = i, \mu). \quad (3.11)$$

As fases de inicialização, indução e probabilidade final são calculadas respectivamente com as seguintes equações

$$\xi_i(T+1) = 1, \quad 1 \leq E \quad (3.12)$$

$$\xi_i(t) = \sum_{j=1}^E A_{ij} g_{ij o_t} \xi_j(t+1), \quad 1 \leq t \leq T \text{ e } 1 \leq i \leq E \quad (3.13)$$

$$P(O|\mu) = \sum_{i=1}^E \pi_i \xi_i(1). \quad (3.14)$$

Alternativamente, é possível combinar as duas abordagens anteriores notando-se a seguinte relação entre $P(O, \mathbb{X}_t = i|\mu)$, α e ξ :

$$\begin{aligned} P(O, \mathbb{X}_t = i|\mu) &= P(o_1 \dots o_T, \mathbb{X}_t = i|\mu) \\ &= P(o_1 \dots o_{t-1}, \mathbb{X}_t = i, o_t \dots o_T|\mu) \\ &= P(o_1 \dots o_{t-1}, \mathbb{X}_t = i|\mu) P(o_t \dots o_T | o_1 \dots o_{t-1}, \mathbb{X}_t = i, \mu) \\ &= P(o_1 \dots o_{t-1}, \mathbb{X}_t = i|\mu) P(o_t \dots o_T | \mathbb{X}_t = i, \mu). \end{aligned} \quad (3.15)$$

Portanto,

$$P(O|\mu) = \sum_{i=1}^E \alpha_i(t) \xi_i(t), \quad 1 \leq t \leq T+1. \quad (3.16)$$

Um segundo problema muito comum na análise de modelos ocultos de Markov refere-se à estimação dos parâmetros que melhor explica as observações. Uma abordagem simples para resolver este problema seria definir a quantidade $\varphi_i(t)$ como

$$\begin{aligned} \varphi_i(t) &= P(\mathbb{X}_t = i | O, \mu) \\ &= P(\mathbb{X}_t = i, O | \mu) / P(O | \mu) \\ &= \alpha_i(t) \xi_i(t) / \sum_{j=1}^E \alpha_j(t) \xi_j(t) \end{aligned} \quad (3.17)$$

e encontrar o conjunto de estados que maximiza $\varphi_i(t)$. Em outras palavras, o conjunto de estados mais provável é

$$\tilde{\mathbb{X}}_t = \arg \max_{1 \leq i \leq N} \varphi_i(t), \quad 1 \leq t \leq T+1. \quad (3.18)$$

Infelizmente, $\tilde{\mathbb{X}}_t$ maximiza apenas a quantidade de estados visitados corretamente. A ordem de passagem pelos estados não é necessariamente a ordem correta. Para calcular a sequência de estados mais prováveis, o algoritmo Viterbi (41) é utilizado. O objetivo deste algoritmo é maximizar a quantidade

$$\arg \max_{\mathbb{X}} P(\mathbb{X} | O, \mu),$$

o que equivale a maximizar

$$\arg \max_{\mathbb{X}} P(\mathbb{X}, O|\mu).$$

A variável auxiliar no processo de programação dinâmica é definida como:

$$\Lambda_j(t) = \max_{\mathbb{X}_1 \dots \mathbb{X}_{t-1}} P(\mathbb{X}_1 \dots \mathbb{X}_{t-1}, o_1 \dots o_{t-1}, \mathbb{X}_t = j|\mu). \quad (3.19)$$

A quantidade $\Lambda_j(t)$ armazena, para cada passo do cálculo, a probabilidade do caminho mais provável que leva a um dado nó. Sua inicialização é calculada como $\Lambda_j(1) = \pi_j, 1 \leq j \leq E$ e sua atualização é dada por

$$\Lambda_j(t+1) = \max_{1 \leq i \leq E} \Lambda_i(t) A_{ij} g_{ijo_t} \quad 1 \leq j \leq E. \quad (3.20)$$

A variável $\psi_j(t)$ armazena o vértice do arco de entrada que levou ao caminho mais provável. Os passos de programação dinâmica utilizados neste caso são dados por

$$\psi_j(t+1) = \arg \max_{1 \leq i \leq E} \Lambda_i(t) A_{ij} g_{ijo_t} \quad 1 \leq j \leq E. \quad (3.21)$$

Finalmente, a sequência de estados mais prováveis seguida pelo algoritmo é calculada com as seguintes equações

$$\tilde{\mathbb{X}}_{T+1} = \arg \max_{1 \leq i \leq E} \Lambda_i(T+1) \quad (3.22)$$

$$\tilde{\mathbb{X}}_t = \psi_{\tilde{\mathbb{X}}_{t+1}}(t+1) \quad (3.23)$$

$$P(\tilde{\mathbb{X}}) = \max_{1 \leq i \leq E} \Lambda_i(T+1) \quad (3.24)$$

3.3 *Part-of-speech tagging*

Part-of-speech tagging corresponde à tarefa de rotulação de uma palavra em um texto com uma determinada classe gramatical (p.e., substantivo, verbo, adjetivo, advérbio, etc). Este processo de rotulação de palavras é inerentemente ambíguo, já que muitas palavras podem ser classificadas com mais de uma categoria. Deste modo, rotuladores automáticos selecionam as categorias mais prováveis para uma dada palavra em um dado contexto. Esta tarefa serve como um pré-processamento em aplicações de extração de informações e *parsing*.

Uma das abordagens mais tradicionais de rotulação é conhecida como rotulador baseado no modelo de Markov (ver Seção 3.2). Neste modelo, uma sequência de rótulos em um texto é considerada uma cadeia de Markov, definida por um conjunto de estados S , um alfabeto

de saída K , probabilidades de transição $\mathbb{A}$ e probabilidades de emissão de símbolos $\mathbb{G}$ (ver Tabela 3.1). Por se tratar de um modelo markoviano, as propriedades de dependência de curto alcance e invariância no tempo devem ser satisfeitas. Neste modelo, cada rótulo $t^{(j)}$ representa um estado. Portanto, a propriedade de dependência de curto alcance, por exemplo, pode ser entendida da seguinte maneira:

- **Dependência de curto alcance:** a probabilidade de mudança para o próximo rótulo depende apenas do rótulo observado anteriormente. Em outras palavras, $P(\mathbb{X}_{i+1} = t^{(j)} | \mathbb{X}_1, \dots, \mathbb{X}_i) = P(\mathbb{X}_{i+1} = t^{(j)} | \mathbb{X}_i)$.

Para derivar a estratégia de rotulamento baseada em modelos de Markov, considere a notação apresentada na Tabela 3.2.

Tabela 3.2 – Tabela de notação e definições para o rotulador baseado em modelos de Markov.

Notação	Definição
w_i	palavra na posição i do corpus
t_i	rótulo da palavra i
$w_{i,i+m}$	palavras ocorrendo na posição i até $i + m$
$t_{i,i+m}$	rótulos para $w_{i,i+m}$
$w^{(l)}$	l -ésima palavra do vocabulário
$t^{(j)}$	j -ésimo rótulo do conjunto de rótulos
$\mathbb{C}(w^{(l)})$	número de ocorrências de $w^{(l)}$ no conjunto de treinamento
$\mathbb{C}(t^{(j)})$	número de ocorrências de $t^{(j)}$ no conjunto de treinamento
$\mathbb{C}(t^{(j)}, t^{(k)})$	número de ocorrência de $t^{(j)}$ seguido de $t^{(k)}$
$\mathbb{C}(w^{(l)} : t^{(j)})$	número de ocorrência de $w^{(l)}$ rotulado como $t^{(j)}$
T	quantidade de rótulos no conjunto de rótulos
n	tamanho da sentença analisada

Para estimar as probabilidades, utiliza-se um conjunto de treinamento, isto é, um conjunto de texto cujos rótulos das palavras já foram fornecidas de antemão. A probabilidade do rótulo $t^{(k)}$ aparecer imediatamente após o rótulo $t^{(j)}$ é estimada a partir de:

$$P(t^{(k)} | t^{(j)}) = \frac{\mathbb{C}(t^{(j)}, t^{(k)})}{\mathbb{C}(t^{(j)})}. \quad (3.25)$$

O cálculo de $P(t^{(k)} | t^{(j)})$ torna possível a estimação de $P(t_{i+1} | t_i)$, isto é, a probabilidade de uma sequência particular ocorrer. Em termos práticos, o objetivo da rotina de rotulação é encontrar a sequência de rótulos (estados da cadeia de Markov) mais prováveis para uma sequência de palavras (símbolos emitidos pela cadeia). Portanto, a tarefa de rotulação pode ser entendida como a tarefa de estimação de

$$P(O_n = k | \mathbb{X}_n = \Sigma_i, \mathbb{X}_{n+1} = \Sigma_j) = g_{ijk}. \quad (3.26)$$

A probabilidade de l -ésima palavra do vocabulário ($w^{(l)}$) ser emitida por um dado estado

(ou rótulo) é dada por:

$$P(w^{(l)}|t^{(j)}) = \frac{\mathbb{C}(w^{(l)} : t^{(j)})}{\mathbb{C}(t^{(l)})}. \quad (3.27)$$

Finalmente, a melhor rotulação da sentença pode ser encontrada a partir do teorema de Bayes:

$$\begin{aligned} \arg \max_{t_{1,n}} P(t_{1,n}|w_{1,n}) &= \arg \max_{t_{1,n}} \frac{P(w_{1,n}|t_{1,n})P(t_{1,n})}{P(w_{1,n})} \\ &= \arg \max_{t_{1,n}} P(w_{1,n}|t_{1,n})P(t_{1,n}). \end{aligned} \quad (3.28)$$

Esta expressão pode ser reduzida a parâmetros que podem ser estimados a partir de um corpus de treinamento. Se assumimos que as palavras são independentes uma das outras, a expressão na equação 3.28 pode ser simplificada para

$$P(w_{1,n}|t_{1,n})P(t_{1,n}) = \prod_{i=1}^n P(w_i|t_{1,n}) \prod_{i=1}^n P(t_i|t_{1,i-1}). \quad (3.29)$$

Assumindo também que a identidade de uma palavra só depende do seu rótulo (ou seja $P(w_i|t_{1,n}) = P(w_i|t_i)$ e $P(t_i|t_{1,i-1}) = P(t_i|t_{i-1})$), a equação 3.29 pode ser reescrita como

$$\begin{aligned} P(w_{1,n}|t_{1,n})P(t_{1,n}) &= \prod_{i=1}^n P(w_i|t_i) \prod_{i=1}^n P(t_i|t_{i-1}) \\ &= \prod_{i=1}^n \left[P(w_i|t_i)P(t_i|t_{i-1}) \right]. \end{aligned} \quad (3.30)$$

Finalmente, a equação que determina o rotulamento ótimo para uma dada sentença é:

$$\tilde{t}_{1,n} = \arg \max_{t_{1,n}} P(t_{1,n}|w_{1,n}) = \arg \max_{t_{1,n}} \prod_{i=1}^n \left[P(w_i|t_i)P(t_i|t_{i-1}) \right] \quad (3.31)$$

Embora a princípio a equação 3.31 possa ser avaliada para todos os pares possíveis de rótulos $t_{1,n}$, esse procedimento daria origem a um algoritmo com complexidade exponencial no tamanho da sentença a ser rotulada. Para evitar esta complexidade, a estratégia de programação dinâmica do algoritmo Viterbi é utilizada (ver Seção 3.2). Neste algoritmo, as funções $\Lambda_i(j)$ e $\psi_{i+1}(j)$ são definidas. A primeira quantidade representa a probabilidade de estar no estado (rótulo) j na palavra i . Já $\psi_{i+1}(j)$ representa o estado (rótulo) mais provável na palavra i supondo que o modelo de Markov encontra o rótulo j na palavra $i + 1$. Assume-se que sentenças são delimitadas por pontos. Portanto $\Lambda_1(.) = 1$. O passo de indução é baseado na equação 3.31. Em termos da notação usada na Seção 3.2 para descrever o algoritmo Viterbi, $A_{kj} = P(t^{(j)}|t^{(k)})$ e $g_{jkw^{(l)}} = P(w^{(l)}|t^{(j)})$:

$$\Lambda_{t+1}(t^{(j)}) = \max_{1 \leq k \leq T} \left(\Lambda_i(t^{(k)})P(t^{(j)}|t^{(k)})P(w_{i+1}|t^{(j)}) \right), \quad 1 \leq j \leq T \quad (3.32)$$

$$\psi_{t+1}(t^j) = \arg \max_{1 \leq k \leq T} \left(\Lambda_i(t^k) P(t^j | t^k) P(w_{i+1} | t^j) \right), \quad 1 \leq j \leq T \quad (3.33)$$

Os rótulos escolhidos $\mathbb{X}_1, \mathbb{X}_2, \dots, \mathbb{X}_n$ para as palavras $w_1, w_2, \dots, w_n$ são obtidos a partir de:

$$\mathbb{X}_n = \arg \max_{i \leq j \leq T} \Lambda_n(t^j) \quad (3.34)$$

$$\mathbb{X}_i = \psi_{i+1}(\mathbb{X}_{i+1}), \quad 1 \leq i \leq n-1 \quad (3.35)$$

$$P(\mathbb{X}_1, \dots, \mathbb{X}_n) = \max_{1 \leq j \leq T} \Lambda_{n+1}(t^j). \quad (3.36)$$

3.4 Pré-processamento

Antes de transformar um texto em uma rede, alguns passos de pré-processamento são aplicados. Inicialmente, retiram-se do texto sinais de pontuação e uma série de palavras de pouco conteúdo semântico^a (*stopwords*) uma vez que estamos interessados na inter-relação entre palavras com pronunciado conteúdo semântico. Este procedimento é tipicamente adotado em muitos trabalhos relacionados (veja por exemplo (24, 26–28)). Em seguida, uma etapa de lematização é aplicada nas palavras remanescentes com auxílio de um rotulador. Palavras no plural são transformadas para sua forma singular, verbos são transformados para sua forma infinitiva e nomes são convertidos para sua forma masculina. Com esta normalização, todas as palavras referentes a um mesmo conceito são associadas a um mesmo vértice, independentemente das diferenças de flexão. Para exemplificar o pré-processamento, a Tabela 3.3 ilustra os respectivos passos para um extrato em inglês do livro “*The Adventures of Sally*”, de P. G. Wodehouse.

Tabela 3.3 – Exemplo de aplicação dos passos de pré-processamento para modelagem de textos como redes. Um extrato (primeira coluna) obtido do livro “*The Adventures of Sally*”, de P. G. Wodehouse está ilustrado após a remoção das *stopwords* (segunda coluna) e consequente lematização (terceira coluna).

Original	Sem Stopwords	Após Lematização
<i>What’s that? asked Sally.</i>	<i>asked Sally</i>	<i>ask Sally</i>
<i>Pay my bill for last week,</i>	<i>pay bill last week</i>	<i>pay bill last week</i>
<i>due this morning. Sally got up</i>	<i>morning Sally got</i>	<i>morning Sally get</i>
<i>quickly, and flitting down the</i>	<i>quickly flitting</i>	<i>quickly flit</i>
<i>table, put her arm round her</i>	<i>table put arm</i>	<i>table put arm</i>
<i>friend’s shoulder and whispered</i>	<i>friend shoulder whispered</i>	<i>friend shoulder whisper</i>
<i>in her ear.</i>	<i>ear</i>	<i>ear</i>

^a As classes de palavras eliminadas compreendem os artigos, preposições e os advérbios. A lista de palavras eliminadas está disponível nos Apêndices A e B.

3.5 Redes da linguagem escrita

Três tipos de redes predominam na modelagem da linguagem como redes complexas baseadas em relações de co-ocorrência, sintáticas e semânticas. Estas estão descritas a seguir.

3.5.1 Redes de co-ocorrência

A linguagem escrita é composta de cadeias lineares de palavras. Dessa forma, o modo mais natural de se associar palavras é conectar palavras adjacentes. Este tipo de rede, denominada rede de co-ocorrência, é amplamente difundido na literatura (15, 22, 27, 29). Na verdade, a modelagem de co-ocorrência pode ser entendida como uma aproximação das ligações sintáticas estabelecidas, já que a maioria destas acontece na primeira vizinhança (15, 42). Nesta rede, vértices representam palavras e arestas são estabelecidas entre palavras vizinhas. A rede obtida para o exemplo da Tabela 3.3 está na Figura 3.2.

Quando o pré-processamento descrito na Seção 3.4 não é aplicado neste tipo de rede, os *hubs* representam palavras de baixo conteúdo semântico (*stopwords*). Este é o caso das palavras de alta frequência em textos como artigos e preposições. O tamanho típico destas redes para um idioma é $M \sim 10^3 - 10^6$, com grau médio $\langle k \rangle \sim 4 - 8$ (43). Outro fator importante nesta modelagem refere-se à escolha do tamanho da janela de conexão das palavras vizinhas. Esta abordagem de conexão representa uma simplificação considerável, já que correlações de longo alcance de fato ocorrem na linguagem humana (44) e por este motivo deveriam a princípio ser consideradas no modelo. A proporção $p^{(1)}$ e $p^{(2)}$ de *links* sintáticos captados de forma equivocada na modelagem de co-ocorrência com janelas de comprimento igual a 1 e 2 são dados por:

$$p^{(1)} = 1 - p_1 \quad (3.37)$$

$$p^{(2)} = \frac{(n-1)(1-p_1) + (n-2)(1-p_2)}{2n-3}, \quad (3.38)$$

onde n é o tamanho das sentenças e p_1 e p_2 representam as probabilidades de que palavras à distância 1 e 2 estejam ligadas (42). Neste caso, considerando que $n \rightarrow \infty$ e $p_1 = 0,70$ e $p_2 = 0,17$ (valores de probabilidade referentes ao idioma inglês (42)), obtemos $p^{(1)} = 0,30$ e $p^{(2)} = 0,56$. Apesar destes erros, a maioria dos trabalhos com redes de adjacência conecta apenas palavras adjacentes (15) pelos motivos listados a seguir:

- a escolha de *links* distantes relevantes necessita um método automático, o que ainda

Figura 3.2 – Exemplo de (a) subgrafo não orientado obtido para a sentença “*What’s that? asked Sally. Pay my bill for last week, due this morning. Sally got up quickly, and flitting down the table, put her arm round her friend’s shoulder and whispered in her ear.*” obtido do livro “*The Adventures of Sally*”, de P.G. Wodehouse; e (b) rede global obtida para as primeiras 1000 associações do mesmo livro. Os passos de pré-processamento, incluindo a remoção de *stopwords*, estão ilustrados na Tabela 3.3.

não existe para realizar operações em vários contextos.

- embora a metodologia de conexão não possa capturar todos os tipos de conexões possíveis, ela é capaz de capturar as principais relações entre palavras, já que poucos tipos de *links* sintáticos são observados a longas distâncias. De fato, existem evidências que os mecanismos de linguagens naturais tendem a minimizar ou restringir a distância euclidiana entre palavras ligadas sintaticamente em sentenças (45).
- o objetivo da metodologia é capturar a maior quantidade possível de *links* relevantes entre palavras.
- uma vez que *links* sintáticos não se cruzam (46) a possível existência de *links* conectando palavras distantes sugere a existência de *links* de curta distância. A afirmação contrária, no entanto, não é verdadeira.

Redes de co-ocorrência apresentam a característica de escala no grau, com dois coeficientes típicos (15) na lei de potência. Por exemplo, na rede de co-ocorrência do *British National Corpus* dois regimes são evidentes com coeficientes $\gamma_N = -1,50$ e $\gamma_N = -2,70$ (15). Além disso, redes de adjacência também são pequeno-mundo. No caso do *British National Corpus*, embora a rede seja formada de $4,7 \times 10^5$ palavras, a distância média entre dois vértices é $\langle L \rangle < 3$ (15). Em outras palavras, para alcançar quaisquer vértices é necessário menos de 3 passos.

3.5.2 Redes sintáticas

Enquanto redes de co-ocorrência (42) são construídas de acordo com relações de adjacência de palavras, as redes de dependência sintática associam com uma aresta apenas as palavras que apresentam alguma relação de dependência na sintaxe da língua. As dependências sintáticas são dadas por gramáticas de dependência, formalismos que definem a estrutura de uma sentença por meio de grafos semelhantes a árvores. Apesar da distinção conceitual entre a modelagem de co-ocorrência e a modelagem sintática, algumas peculiaridades as relacionam. Por exemplo, 90% das ligações da rede sintática acontecem entre palavras imediatamente vizinhas no texto. Desta forma, pode-se imaginar a modelagem por co-ocorrência de palavras como uma aproximação da modelagem sintática (42). O tamanho típico de redes sintáticas na literatura é $M \sim 10^3 - 10^4$, com grau médio $\langle k \rangle \sim 5 - 10$ (43). Assim como nas redes de co-ocorrência, as redes sintáticas apresentam as propriedades livre de escala e pequeno mundo.

3.5.3 Redes semânticas

Redes semânticas são geralmente construídas a partir de um conjunto inicial de palavras mapeando conceitos que são associados de acordo com certas regras semânticas. Dentre as principais regras de associação, podem-se citar as relações “é-um”, “parte-todo” ou “oposição” (17, 43, 47). Uma rede semântica associativa está ilustrada na Figura 3.3. Estudos sobre a organização de tais redes revelam que o número de conexões de uma palavra é proporcional ao número de sentidos que ela pode assumir (43), sugerindo que as palavras polissêmicas são responsáveis pela organização estrutural do sistema complexo de conceitos. Em geral, redes semânticas apresentam alto coeficiente de aglomeração, o que favorece a busca da informação realizada por meio de associação. O tamanho típico para estas na literatura é $M \sim 10^4 - 10^5$, com grau médio $\langle k \rangle \sim 2 - 4$ (43).

Figura 3.3 – Exemplo de subrede semântica associativa. Vértices representam conceitos e arestas representam relações semânticas entre conceitos. Note, por exemplo, que “*beer*” está conectado com “*lager*” pois esta última é um tipo de cerveja.

3.6 Medidas básicas

As medidas estatísticas mais simples para caracterizar texto são o tamanho do vocabulário M , que corresponde ao número de palavras distintas no texto, e a frequência

das palavras, denotado por N . A heterogeneidade de contextos ao redor das palavras é quantificado com a medida de seletividade S (36). Se uma palavra é fortemente seletiva então os seus vizinhos mais próximos tendem a se repetir nas diferentes ocorrências desta palavra. Matematicamente, a seletividade de uma palavra i é dada por $S_i = 2N_i/\mathcal{U}_i$, onde $\mathcal{U}_i$ representa o número de palavras distintas que aparecem imediatamente antes ou depois da palavra i no texto. Outra medida básica é o número de palavras repetidas adjacentes. Em algumas línguas, esta repetição é bastante rara, mas pode ter frequência não desprezível em outras (ver Capítulo 5). Nesta tese, o número de palavras ou bigramas que aparecem de forma adjacente são denotadas como b .

3.7 Entropia das palavras

Um dos tópicos mais importantes na área de recuperação da informação refere-se à identificação de palavras-chave. Uma das abordagens mais simples é verificar as propriedades estatísticas das palavras relevantes através de um método de análise de frequência baseada na distribuição Zipf. Esta análise, desenvolvida por Luhn (48), consiste na contagem da frequência de cada palavra distinta no texto a fim de gerar uma lista ordenada por esta quantidade. Através de uma análise empírica desta distribuição, Luhn propôs a eliminação das palavras situadas em ambas as extremidades da distribuição Zipf. Desta forma, eliminam-se as palavras pouco frequentes (irrelevantes para o documento) e as palavras de alta frequência, que são formadas em muitos casos de *stopwords*.

Estratégias baseadas em conceitos provenientes da Física também foram propostas na literatura. Uma das principais abordagens, definida por Herrera e Pury (49), é baseada na entropia de informação da Shannon, definida num conjunto discreto de probabilidades $p_1, p_2, \dots, p_{\mathcal{P}}$ como:

$$\mathcal{H}(p_1, p_2, \dots, p_{\mathcal{P}}) = -K \sum_{i=1}^{\mathcal{P}} p_i \log p_i, \quad (3.39)$$

onde K representa uma constante positiva. Algumas propriedades importantes da definição na equação 3.39 são:

- A entropia $\mathcal{H}$ é contínua em $\{p_i\}$.
- Se o conjunto original é dividido em dois conjuntos com probabilidades $\varrho'_1 = p_1 + \dots + p_k$ e $\varrho_2 = p_{k+1} + \dots + p_n$, então a entropia $\mathcal{H}(w_1, w_2)$ pode ser reescrita a partir dos parâmetros ϱ_1, ϱ_2 e da entropia total:

$$\mathcal{H}(\varrho_1, \varrho_2) = \mathcal{H}(p_1, \dots, p_n) - \varrho_1 \mathcal{H}(p_1/\varrho_1, \dots, p_k/\varrho_1) - \varrho_2 \mathcal{H}(p_1/\varrho_2, \dots, p_k/\varrho_2). \quad (3.40)$$

- Para $p_1 = p_2 = \dots = p_n = \mathcal{P}^{-1}$, a entropia $\mathcal{H}$ é uma função monotônica e crescente em $\mathcal{P}$.

A idéia básica por trás deste método é que as palavras relevantes se concentram em seções específicas do texto e a entropia de informação pode ser usada para captar esta distribuição irregular das palavras. Para definir esta medida de relevância, divide-se o documento em partições naturais (como parágrafos, capítulos ou seções). Seja $n^{(i)}$ o número total de *tokens* na parte i e $n^{(i)}(w)$ a frequência de ocorrência da palavra w nesta parte. A fração $f_i(w) \equiv n^{(i)}(w)/n^{(i)}$, $1 \leq i \leq \mathcal{P}$ representa a frequência relativa da palavra w na partição i . Neste caso, as quantidades

$$\sum_{i=1}^{\mathcal{P}} n^{(i)} \quad \text{e} \quad \sum_{i=1}^{\mathcal{P}} n^{(i)}(w) = N_w$$

são respectivamente o número total de *tokens* no corpus e o número total de *tokens* da palavra w . A medida que quantifica a chance de a palavra w aparecer na partição i é:

$$p_i(w) = \frac{f_i(w)}{\sum_{j=1}^{\mathcal{P}} f_j(w)}. \quad (3.41)$$

A aplicação da entropia definida em 3.39 à equação 3.41 resulta na fórmula da entropia associada à palavra w :

$$\mathcal{H}(w) = -\frac{1}{\ln(\mathcal{P})} \sum_{i=1}^{\mathcal{P}} p_i(w) \ln(p_i(w)), \quad (3.42)$$

onde $K = \ln^{-1}(\mathcal{P})$ foi selecionado para satisfazer a condição $\max \mathcal{H} = 1$. Portanto, quando uma dada palavra está uniformemente distribuída nas partições do documento a entropia de informação de w assume o valor $\mathcal{H}(w) = 1$. Por outro lado, se a palavra w aparece apenas em uma partição ($p_j = 1$ e $p_i = 0$, $1 \leq i \leq \mathcal{P}$ e $i \neq j$) então $\mathcal{H}(w) = 0$. Portanto, *stopwords* tendem a apresentar alto valor de entropia pois estas palavras são distribuídas uniformemente em todas as partições do documento.

Em geral, é desejável normalizar a quantidade definida na equação 3.42 pois esta quantidade cresce com a frequência da palavra (50). Neste caso, é necessário conhecer a entropia esperada de uma palavra ocorrendo N_w vezes no texto. O valor esperado pode ser obtido analiticamente da seguinte maneira. Considerando um corpus de $N = n^{(i)}$ tokens dividido em $\mathcal{P}$ partes, com $n^{(i)}$ tokens na i -ésima parte, onde $i = 1, 2, \dots, \mathcal{P}$. Em um *corpus* aleatório, a probabilidade de que uma palavra w apareça na parte j é N_j/N . Dessa forma, a probabilidade de que w apareça $n^{(1)}(w)$ vezes na parte 1, $n^{(2)}(w)$ vezes na parte 2 e

assim por diante segue uma distribuição multinomial

$$p(n^{(1)}, n^{(2)}, \dots, n^{(\mathcal{P})}) = n! \prod_{j=1}^{\mathcal{P}} \frac{1}{n^{(j)}!} \left(\frac{N_j}{N} \right)^{n^{(j)}}, \quad (3.43)$$

onde

$$n = \sum_{j=1}^{\mathcal{P}} n^{(j)} \quad (3.44)$$

é a frequência absoluta (número de tokens) da palavra w . Para simplificar o cálculo, consideramos o caso em que todas as partes possuem o mesmo comprimento, isto é, $n^{(i)} = n/\mathcal{P}$. Dessa forma, a medida de probabilidade definida na equação 3.41 pode ser definida como $p_i = n^{(i)}/n$ e a entropia é definida por:

$$\mathcal{H} = \frac{1}{\ln \mathcal{P}} \sum_{i=1}^{\mathcal{P}} \frac{n^{(i)}}{n} \ln \left(\frac{n^{(i)}}{n} \right). \quad (3.45)$$

Agora é necessário avaliar o valor médio da entropia na distribuição dada pela equação 3.43. Para isto, precisamos calcular a média de cada termo da equação 3.45 usando a distribuição marginal, $p(n^{(i)})$, obtida da equação 3.43. Todas as distribuições marginais geram binomiais com média $n\mathcal{P}^{-1}$ e variância $n\mathcal{P}^{-1}(1 - \mathcal{P}^{-1})$. Dessa forma, a entropia média é dada por

$$\langle \mathcal{H} \rangle = -\frac{\mathcal{P}}{\ln \mathcal{P}} \sum_{m=0}^n \frac{m}{n} \ln \left(\frac{m}{n} \right) \left(\frac{n!(1 - \mathcal{P}^{-1})^{n-m}}{m!(n-m)!\mathcal{P}^m} \right). \quad (3.46)$$

Para palavras com frequência absoluta alta ($n \gg 1$) é possível aproximar a distribuição binomial para uma gaussiana ($G(x; \mu, \sigma)$) com média $\mu = \mathcal{P}^{-1}$ e variância $\sigma^2 = n^{-1}(\mathcal{P} - 1)\mathcal{P}^{-2}$. Dessa forma, a equação 3.46 pode ser reescrita como:

$$\langle \mathcal{H} \rangle \approx -\frac{\mathcal{P}}{\ln \mathcal{P}} \int_0^1 x \ln x G(x; \mu, \sigma) dx. \quad (3.47)$$

No limite em que $n \gg 1$, $\sigma \rightarrow 0$ e a gaussiana se concentra na sua média μ . Usando a expansão de $x \ln x$ em torno de μ ,

$$x \ln x \approx \mu \ln \mu + (1 + \ln \mu)(x - \mu) + \frac{1}{2\mu}(x - \mu)^2, \quad (3.48)$$

na equação 3.47 e fazendo uso da propriedade

$$\int_{-\infty}^{\infty} (x - \mu)^2 G(x; \mu, \sigma) dx = \sigma^2, \quad (3.49)$$

recuperamos o resultado de (49)

$$\mathcal{H}_{rand}(w) \approx 1 - \frac{\mathcal{P} - 1}{2N_w \ln(\mathcal{P})}. \quad (3.50)$$

Desta forma, define-se o índice de entropia normalizado para a palavra w como:

$$\mathcal{H}_{norm}(w) = \frac{1 - \mathcal{H}(w)}{1 - \mathcal{H}_{rand}(w)} = \frac{2N_w \ln \mathcal{P}}{\mathcal{P} - 1} \left[1 + \frac{1}{\ln \mathcal{P}} \sum_{i=1}^{\mathcal{P}} p_i(w) \ln p_i(w) \right]. \quad (3.51)$$

Embora esta abordagem baseada na teoria de informação forneça palavras-chave razoáveis (49), ela pressupõe a existência de partições naturais. Em muitos casos, no entanto, este particionamento não está bem definido. Por este motivo, outras metodologias para reconhecimento de palavras-chave em documentos têm sido desenvolvidas. Uma delas é a intermitência (ou *burstiness*) das palavras, que é descrita na próxima seção.

3.8 Intermitência de palavras

A distribuição irregular de palavras em diferentes documentos representa uma característica textual útil para reconhecimento de padrões. Por exemplo, a análise de palavras recorrentes em documentos específicos pode ser útil para detectar palavras-chave, tópicos e até mesmo a autoria de documentos (39, 51). Esta ideia básica é utilizada na medida *term frequency - inverse document frequency (tf-idf)* (39). O termo relativo à frequência é geralmente calculado como

$$\text{tf}(i) = N_i / \sum_{j=1}^M N_j. \quad (3.52)$$

Note que $\text{tf}(i)$ é definido de forma normalizada a fim de evitar um viés favorecendo uma dada palavra em documentos longos. Já o termo *idf* é dado por

$$\text{idf}(i, t) = \log \left(\frac{\|\mathcal{T}\|}{\|\{\tau \in \mathcal{T} : i \in \tau\}\|} \right), \quad (3.53)$$

onde $\|\mathcal{T}\|$ representa o número de documentos na coleção e $\|\{\tau \in \mathcal{T} : i \in \tau\}\|$ quantifica o número de documentos no qual a palavra i ocorre. Infelizmente, há situações em que a comparação com um corpus $\mathcal{T}$ não é possível, como no problema de reconhecimento de autoria sem a utilização de exemplos de textos escritos pelos potenciais autores. Problemas deste tipo são tratados fazendo uso do fato de que palavras são distribuídas irregularmente não apenas entre documentos, mas também dentro de um mesmo documento (49, 50, 52–55).

A quantificação da irregularidade da distribuição de palavras em textos é feita com medidas normalmente usadas por físicos (50, 55). Seguindo as Refs. (49, 53, 55), usamos a estatística de tempo de recorrência, uma medida padrão de intermitência em séries temporais (52). Em textos, o tempo é avaliado de acordo com o número de palavras. Mais especificamente, para cada palavra i , o tempo de recorrência T_j é definido como o número de palavras entre duas ocorrências sucessivas de i (i.e., ocorrência j e a ocorrência $j + 1$) acrescentado de 1. Por exemplo, o tempo de recorrência para a palavra ‘de’ na sentença anterior é $T_1 = 8$ e $T_2 = 6$. Uma palavra que aparece N_i vezes num texto de tamanho N_T gera uma sequência de $N_i - 1$ tempos de recorrência $\{T_1, T_2, \dots, T_{N_i-1}\}$. A fim de incorporar também o tempo T_f até a primeira ocorrência da palavra e depois da última ocorrência T_l , consideramos $T_N = T_f + T_l$. Neste caso $\langle T \rangle = N_T/N_i$, onde $\langle \cdot \rangle$ é a média sobre os distintos T_j ’s. Note que o tempo de recorrência médio $\langle T \rangle$ não fornece informação adicional além da frequência N_i . A intermitência da palavra aparece na variância dos T_j ’s em torno de $\langle T \rangle$ e pode ser quantificada por

$$I \equiv \sigma_T / \langle T \rangle, \quad (3.54)$$

onde $\sigma_T = \langle T^2 \rangle - \langle T \rangle^2$. Palavras intermitentes possuem $I > 1$, e palavras aparecendo em intervalos regulares apresentam $I < 1$. Para calcular o valor da intermitência de palavras com distribuição aleatória no texto, i.e., com distribuição de tempos de recorrência T seguindo uma distribuição do tipo

$$P(t = T) = \frac{\lambda^T e^{-\lambda}}{T!}, \quad (3.55)$$

considere o seguinte teorema a respeito da esperança $E[X]$ de uma variável aleatória X .

Teorema 1. Propriedade de média e variância de variáveis aleatórias: *Seja X uma variável aleatória. Considere a função $\phi(t) = E[e^{tX}]$. Os valores de $E[X]$ e $E[X^2]$ podem ser calculados como $E[X] = \dot{\phi}(0)$ e $E[X^2] = \ddot{\phi}(0)$.*

Demonstração. Seja

$$\phi(t) = E[e^{tX}] = \int_{-\infty}^{+\infty} e^{tX} dX. \quad (3.56)$$

Calculando a primeira derivada de $\phi(t)$, obtemos

$$\dot{\phi}(t) = \frac{d}{dt} E[e^{tX}] = E \left[\frac{d}{dt} (e^{tX}) \right] = E[X e^{tX}]. \quad (3.57)$$

Portanto $\dot{\phi}(0) = E[X]$. Analogamente,

$$\ddot{\phi}(t) = \frac{d}{dt} \dot{\phi}(t) = \frac{d}{dt} E[X e^{tX}] = E \left[\frac{d}{dt} (X e^{tX}) \right] = E[X^2 e^{tX}], \quad (3.58)$$

e portanto $\ddot{\phi}(0) = E[X^2]$. □

A partir do teorema 1, é possível calcular o valor da intermitência de palavras com distribuição aleatória no texto conforme demonstração do seguinte teorema:

Teorema 2. Intermitência de palavras apresentando distribuição aleatória no texto: *palavras cujo tempo de recorrência segue uma distribuição Poisson apresentam intermitência $I = 1$ (ver exemplo na Figura 3.4).*

Demonstração. Aplicando as equações 3.56 e 3.57 na equação 3.55, obtemos

$$\phi(t) = E[e^{tX}] = \sum_{n=0}^{\infty} \frac{e^{tn} e^{-\lambda} \lambda^n}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} = e^{-\lambda} e^{\lambda e^t} = \exp\{\lambda(e^t - 1)\} \quad (3.59)$$

$$\dot{\phi}(t) = \lambda e^t \exp\{\lambda(e^t - 1)\}, \quad (3.60)$$

$$\ddot{\phi}(t) = (\lambda e^t)^2 \exp\{\lambda(e^t - 1)\} + \lambda e^t \exp\{\lambda(e^t - 1)\}. \quad (3.61)$$

Portanto, a esperança e a variância calculada da distribuição Poisson são dadas por:

$$E[X] = \dot{\phi}(0) = \lambda, \quad (3.62)$$

$$E[X^2] = \ddot{\phi}(0) = \lambda^2 + \lambda. \quad (3.63)$$

Como é dada pela razão entre $E[X^2]$ e $E[X]$, a intermitência pode ser calculada como

$$I = \frac{\text{Var}(T)}{E[T]} = \frac{E[T^2] - E^2[T]}{E[T]} = \frac{\lambda^2 + \lambda - \lambda^2}{\lambda} = 1. \quad (3.64)$$

Assim, palavras distribuídas aleatoriamente no texto possuem intermitência $I = 1$. □

Para verificar como a intermitência varia de palavra para palavra, comparamos na Tabela 3.4 as palavras com frequências similares mas com diferentes valores de intermitência em um livro. Claramente é possível notar que as palavras mais intermitentes são as palavras que representam os tópicos abordados (por exemplo, nomes dos personagens e lugares), independentemente da frequência. Um exemplo da distribuição dos tempos de retorno de palavras tópicos é mostrada na Figura 3.5. De fato, 15 das 16 palavras mais intermitentes estão ligadas a personagens específicos. Portanto, a medida de intermitência caracteriza bem as palavras específicas de um dado contexto.

Além da intermitência, correlações de longo alcance também têm sido utilizadas para caracterizar as propriedades temporais de textos e sistemas complexos em geral (56, 57). Nesta tese, usamos apenas a medida de intermitência uma vez que nossas análises se concentram em palavras, enquanto correlações de longo alcance tipicamente utilizam letras (58).

Figura 3.4 – Histograma do tempo de retorno para a palavras *ginger* (a) no livro “*The Adventures of Sally*” e em (b) um texto aleatório. Note que, em textos aleatórios, a probabilidade de curtos espaços de recorrência é bem maior no texto aleatório.

Figura 3.5 – Posição absoluta no texto contada a partir do início do livro “*The Adventures of Sally*” para as palavras (a) *Carmyle* ($N_i = 54$) e (b) *feel* ($N_i = 54$). Note que a distribuição em (a) é mais irregular que a distribuição em (b), indicando que a palavra representada em (a) representa melhor o tópico abordado no livro do que a palavra em (b).

Tabela 3.4 – No livro *The Adventures of Sally* existem $N_T = 15\,173$ palavras (tokens), 3 657 tipos (palavras distintas) e 716 palavras com $N_i \geq 5$. As cinco palavras com os maiores valores de intermitência são mostradas na parte esquerda da tabela. Para comparação, no lado direito é mostrado para cada uma destas palavras outra palavra com a frequência mais próxima.

Palavra	N_i	I_i	Palavra	N_i	I_i
<i>Jules</i>	26	4,31	<i>Turn</i>	26	1,55
<i>Hobson</i>	31	4,09	<i>Here</i>	31	1,35
<i>Ginger</i>	115	3,86	<i>Get</i>	117	1,24
<i>Carmyle</i>	54	3,60	<i>Feel</i>	53	0,87
<i>Bunbury</i>	20	3,59	<i>People</i>	20	1,39

3.9 Caracterização da estrutura textual

A caracterização textual é feita através do uso de medidas básicas, medidas de redes complexas e medidas de intermitência de aparição de palavras. Estas medidas estão definidas respectivamente nas Seções 3.6, 3.8 e 2.4. Note que a maioria destas medidas é relativa às palavras e não ao texto como um todo. Portanto, é necessário adotar algum procedimento para caracterizar textos a partir da caracterização das palavras. A escolha mais natural é o valor médio $\langle X \rangle$, onde $\langle \dots \rangle \equiv \frac{1}{M} \sum_{i=1}^M \dots$ representa a média considerando as M palavras distintas. Também é interessante tomar o desvio padrão $\sigma(X) = \frac{1}{M} \sum_{i=1}^M (X_i - \langle X \rangle)^2$. Tais escolhas, no entanto, podem não considerar as especificidades das distribuições de cada medida. Por este motivo, um estudo mais detalhado sobre o comportamento das medidas para palavras é necessário.

Iniciamos a análise das medidas observando a distribuição de cada uma delas no livro “*The Adventures of Sally*”. A distribuição destas medidas para todas as palavras deste livro foi estimada pela densidade de probabilidade empírica $P(X)$. A Figura 3.6 exemplifica dois tipos de distribuições obtidas, que independem do livro analisado (distribuições similares foram observadas para outros livros). Medidas como L , C e I apresentaram picos e larguras bem definidos (similar a uma distribuição Gaussiana). Por outro lado, as medidas N , B , S , k e s parecem seguir uma distribuição do tipo lei de potência $p(X) \sim X^{-\alpha}$ (9, 27, 36, 59). A cauda em N ocorre devido à lei de Zipf (9), que também aparece em B , k e s devido à alta correlação com N . A distribuição de S segue uma lei de potência de acordo com a Ref. (36). Devido à observação destes dois comportamentos distintos, duas caracterizações adicionais foram propostas: uma para $X = \{L, C, I\}$ e outra para $X = \{N, B, S, k, s\}$.

Para $X = \{N, B, S, k, s\}$, as palavras mais frequentes contribuem fortemente para $\langle X \rangle$. Para compensar este efeito, consideramos também uma média modificada, que é definida como $\langle X \rangle_2 \equiv \langle \log X \rangle$. Para $X = \{L, C, I\}$, o oposto é verdadeiro. A média é dominada pelo grande número de palavras pouco frequentes. Por este motivo, a seguinte média modificada é introduzida: $\langle X \rangle_2 \propto \sum_i X_i \log N_i$. Na verdade, $\langle X \rangle_2$ representa uma

média ponderada com pesos proporcionais ao logaritmo da frequência. Alternativamente, calculamos a média $\|F\|^{-1} \sum_{i \in F} X_i$, onde F é o conjunto das 50 palavras mais frequentes. Espera-se, portanto, que as quantidades $\langle X \rangle$ e $\langle X \rangle_2$ forneçam boa estimativa dos valores típicos de X .

Para as medidas $X = \{N, B, S, k, s\}$, a informação relevante é fornecida pelas palavras com altos valores de X , i.e., pelas palavras nas caudas das distribuições do tipo $P(X) \sim X^{-\alpha}$. A fim de caracterizar as distribuições destas medidas, utilizamos o coeficiente α da lei de potência^b. Os altos valores de $X = \{L, C, I\}$ foram caracterizados calculando-se o terceiro momento da distribuição de $P(X)$, que é uma medida de assimetria da distribuição. Resumindo, as medidas das palavras foram obtidas da seguinte maneira:

$$\text{Valor médio: } \langle X \rangle \quad \text{para } X = \{N, k, L, C, S, B, I, s\}. \quad (3.65)$$

$$\text{Desvio: } \sigma_X \quad \text{para } X = \{N, k, L, C, S, B, I, s\}. \quad (3.66)$$

$$\text{Média modificada: } \langle X \rangle_2 = \begin{cases} \langle \log(X) \rangle & \text{para } X = \{N, B, S, k, s\}, \\ \langle X \log N \rangle / \langle \log N \rangle & \text{para } X = \{L, C, I\}, \\ \|F\|^{-1} \sum_{i \in F} X_i & \text{para } X = \{L, C, I\}. \end{cases} \quad (3.67)$$

$$\text{Cauda: } \gamma_X = \begin{cases} \alpha \text{ in } X^{-\alpha} & \text{para } X = \{N, B, S, k, s\}, \\ \text{assimetria}(X) \equiv \langle \left(\frac{X - \langle X \rangle}{\sigma_X} \right)^3 \rangle & \text{para } X = \{L, C, I\}. \end{cases} \quad (3.68)$$

Obviamente, a escolha destas quantidades é arbitrária. No entanto, guiamos esta escolha na tentativa de capturar características das distribuições que possam ser úteis para aplicações de processamento de línguas naturais.

^b O *fitting* foi realizado usando a distribuição acumulada com espaçamento logarítmico, como sugerido na Ref. (60). O limiar $X > 3 \times 10^3$ foi usado para $X = B$ (veja Fig. 3.6e). Para as outras medidas, nenhum limiar foi usado.

Figura 3.6 – Função densidade de probabilidade $P(X)$ obtida a partir de palavras distintas do livro “*The Adventures of Sally*”. (a) $X = L$ caminho mínimo, (b) $X = I$ intermitência, (c) $X = N$ frequência, e (d) $X = B$ betweenness. Em (c) e (d) a distribuição cumulativa $\rho(x \geq X) \equiv \int_X^\infty \rho(x)dx$ está ilustrada, sendo que $\rho(X)$ é mostrado no gráfico inferior. As legendas indicam que os atributos definidos nas equações (3.66)-(3.68) obtidos para tais distribuições, e $\text{Corr}(X, N)$ indica o coeficiente de correlação de Pearson entre X e N calculado para todas as palavras

Capítulo 4

Reconhecimento de padrões

O problema de classificação (ou categorização) consiste na tarefa de associar categorias a elementos de acordo com suas propriedades ou atributos (61). Esta tarefa é essencial para descobrir padrões em grandes quantidades de dados. Existem diversas técnicas de categorização automática, tradicionalmente subdivididas em:

- **Classificação supervisionada:** uma relação binária mapeando a entrada para uma saída é gerada pelo algoritmo.
- **Classificação não supervisionada:** agrupa os dados com características semelhantes entre si. Esta categoria é frequentemente denominada na literatura como *clustering*.
- **Classificação semi-supervisionada:** o conjunto inicial de dados é composto apenas por uma pequena entrada rotulada. Grande parte dos dados da entrada não está rotulada, i.e., a classe associada a eles é desconhecida. Neste caso o objetivo é similar à classificação supervisionada: gerar uma relação binária mapeando a entrada para saída.

Antes de descrever as técnicas de classificação, é necessário apresentar alguns conceitos e nomenclatura utilizada nesta tese. Dado um conjunto de exemplos, a seguinte distinção pode ser feita:

- **Conjunto de treinamento:** representa o conjunto de exemplos que é utilizado como entrada aos métodos de reconhecimento de padrões. Em outras palavras, refere-se ao conjunto de exemplos conhecidos pelo algoritmo de aprendizado responsável pela inferência do classificador. O conjunto de treinamento, também conhecido como *seen cases*, será representado neste capítulo como $\mathcal{S}_{\text{tr}} = \{\beta_{(\text{tr},1)}, \beta_{(\text{tr},2)}, \beta_{(\text{tr},3)} \dots\}$.
- **Conjunto de teste:** é o conjunto utilizado para avaliar o desempenho do modelo inferido pelo algoritmo indutor. Este conjunto é denotado por $\mathcal{S}_{\text{ts}} = \{\beta_{(\text{ts},1)}, \beta_{(\text{ts},2)}, \beta_{(\text{ts},3)} \dots\}$.

Geralmente, para uma avaliação de desempenho mais robusta, deve-se garantir que $\mathcal{S}_{\text{tr}} \cap \mathcal{S}_{\text{ts}} = \emptyset$.

Um exemplo β pode ser caracterizado por $\mathcal{M}$ atributos através do vetor $\vec{\beta} = (F_1 = \beta^{(1)}, F_2 = \beta^{(2)}, \dots, F_{\mathcal{M}} = \beta^{(\mathcal{M})})$, onde $\mathcal{F} = \{F_1, F_2, \dots\}$ representa o conjunto de atributos descrevendo o exemplo β . Em outras palavras, o valor do k -ésimo atributo para um dado exemplo β é representado por $\beta^{(k)}$. No caso da classificação supervisionada, um dado exemplo pode assumir um conjunto finito de classes $\mathcal{C} = \{c_1, c_2, c_3, \dots, c_l\}$.

Nesta tese tratamos principalmente problemas de classificação supervisionada. O objetivo é encontrar automaticamente padrões relacionando a estrutura do texto verificada pelas métricas de redes complexas com suas características de estilo, por exemplo. Os algoritmos de classificação utilizados foram: vizinhos mais próximos (61, 62) (κ -NN), C4.5 (61, 63), *Naive Bayes* (61) e classificação híbrida (64). Os detalhes destes algoritmos, métodos de avaliação e técnicas de redução de dimensionalidade são apresentados nas próximas seções.

4.1 Técnicas de classificação

4.1.1 Vizinhos mais próximos (κ -NN)

Um dos algoritmos mais simples de classificação é o algoritmo de vizinhos mais próximos (κ *nearest neighbours*). Este algoritmo classifica um exemplo $\beta_{(\text{ts})} \in \mathcal{S}_{\text{ts}}$ de acordo com um processo de voto realizado em $\mathcal{S}_{\text{tr}}$ (i.e., no conjunto de dados cuja classe é conhecida de antemão). Se a maioria dos κ vizinhos mais próximos de $\beta_{(\text{ts})}$ pertence à classe c_i , então c_i é associada ao exemplo $\beta_{(\text{ts})}$. A Figura 4.1 ilustra o processo de classificação.

O algoritmo κ -NN pode ser estendido para uma versão *fuzzy*. Diferentemente do método tradicional, a versão *fuzzy* associa uma probabilidade u_{ik} que representa a força com que a instância β_k pertence à classe c_i . Por definição, u_{ik} é restrito às seguintes condições:

$$\sum_{i=1}^{|\mathcal{C}|} u_{ik} = 1, \quad (4.1)$$

$$0 \leq \sum_{k=1}^n u_{ik} \leq n, \quad (4.2)$$

onde n representa o número total de elementos. Note que, de acordo com a equação 4.1, u_{ik} pode ser visto como uma probabilidade. Dessa forma, se $|\mathcal{C}| = 2$ e $u_{1k} = 0,95$ então existe uma grande possibilidade de que a instância β_k pertença à classe “1”. Por outro

Figura 4.1 – Exemplo de classificação com o classificador induzido pelo algoritmo κ -NN. Note que a decisão se baseia num processo de votação realizado nos κ vizinhos mais próximos da instância de teste (quadrado verde). Para $\kappa = 5$, a classe associada é a classe “vermelho”, já que esta classe domina o círculo pontilhado mais interno. Por outro lado, quando $\kappa = 13$ e portanto todos os objetos dentro do círculo central pontilhado passam a ser considerados, a classe associada é a classe “azul” pois esta classe se torna a classe majoritária. Finalmente, quando $k = 19$, a classe “azul” continua sendo associada à instância de teste porque a maioria dos objetos mais próximos pertence à classe “azul”.

lado, se $u_{1k} = 0,55$, então a instância β_k está situada na borda da classe “1”. A equação 4.2, por sua vez, impõe que cada classe não pode conter mais do que n objetos.

Para determinar a probabilidade de um objeto desconhecido pertencer a uma dada classe, é necessário calcular sua distância às instâncias de treinamento. A distância não normalizada $\tilde{u}_i(\beta)$ de uma instância de teste β à i -ésima classe é calculada em função da distância de β aos κ vizinhos mais próximos no conjunto de treinamento:

$$\tilde{u}_i(\beta) = \sum_{j=1}^{\kappa} u_{ij} \|\beta - \beta_{(\text{tr},j)}\|^{2/(1-m)}. \quad (4.3)$$

Neste caso, $\tilde{u}_i(\beta)$ depende não apenas da distância dos κ vizinhos mais próximos, mas também da força com que estes vizinhos pertencem à classe c_i . Portanto, $\tilde{u}_i(\beta)$ assume altos valores sempre que a distância entre β e $\beta_{(\text{tr},j)}$ for pequena e $\beta_{(\text{tr},j)}$ ser um objeto bastante representativo da classe c_i (isto é, u_{ij} é alto). O parâmetro m é responsável pelo peso associado à distância $\|\beta - \beta_{(\text{tr},j)}\|$. Se m assumir altos valores, $\|\beta - \beta_{(\text{tr},j)}\|$ pouco influenciará na computação de $\tilde{u}$. Por outro lado, sempre que $m \rightarrow 1$ a influência dos vizinhos mais próximos se torna tão forte que a contribuição dos vizinhos mais distantes praticamente desaparece. Finalmente, para assegurar que $u_i(\beta)$ represente um valor de probabilidade é necessário normalizar $\tilde{u}_i(\beta)$ de modo que $\sum_i \alpha \tilde{u}_i(x) = 1$. Neste caso, α

pode ser obtido a partir de

$$\begin{aligned}
\alpha^{-1} &= \sum_i \tilde{u}_i(\beta) \\
&= \sum_i \sum_j u_{ij} \|\beta - \beta_{(\text{tr},j)}\|^{2/(1-m)} \\
&= \sum_j \left[\sum_i u_{ij} \|\beta - \beta_{(\text{tr},j)}\|^{2/(1-m)} \right] \\
&= \sum_j \|\beta - \beta_{(\text{tr},j)}\|^{2/(1-m)} \sum_i u_{ij} \\
&= \sum_j \|\beta - \beta_{(\text{tr},j)}\|^{2/(1-m)}
\end{aligned} \tag{4.4}$$

Portanto, $u_i(\beta)$ é dado por:

$$u_i(\beta) = \frac{\sum_{j=1} u_{ij} \|\beta - \beta_{(\text{tr},j)}\|^{2/(1-m)}}{\sum_{j=1} \|\beta - \beta_{(\text{tr},j)}\|^{2/(1-m)}}. \tag{4.5}$$

Para definir a força u_{ij} com que cada objeto $\beta_{(\text{tr},j)}$ pertence à classe c_i , considere as seguintes definições.

$$\mu_i = \frac{1}{n_i} \sum_{\beta_{(\text{tr},j)} \in c_i} \beta_{(\text{tr},j)} \tag{4.6}$$

representa o centróide da classe c_i composta de n_i instâncias de treinamento. $\text{cov}(a, b)$ é a covariância entre os atributos a e b :

$$\text{cov}(a, b) = \frac{1}{n-1} \sum_{v=1}^n (\beta_v^{(a)} - \langle \beta^{(a)} \rangle) (\beta_v^{(b)} - \langle \beta^{(b)} \rangle), \tag{4.7}$$

onde $\langle \dots \rangle$ é a média sobre as n observações das duas variáveis. A matriz de covariância entre todos os atributos é definida como $\text{Cov} = \{\text{cov}(a, b)\}$. A similaridade entre a instância $\beta_{(\text{tr},j)}$ e a classe c_i é dada por:

$$\psi(i, \beta_{(\text{tr},j)}) = \left[(\beta_{(\text{tr},j)} - \mu_i)^T \text{Cov}_i^{-1} (\beta_{(\text{tr},j)} - \mu_i) \right]^{-1/2}, \tag{4.8}$$

que representa o inverso da distância Mahalanobis (65). Esta medida é útil porque, ao contrário da distância euclidiana, ela considera a dispersão Cov_i das instâncias na classe c_i para calcular a distância de $\beta_{(\text{tr},j)}$ na classe c_i . Dessa forma, se $\beta_{(\text{tr},j)}$ está distante do centróide da classe c_i mas está incluído dentro da dispersão da região de i , então a distância entre β_j e a classe c_i será relativamente pequena. Por outro lado, se $\beta_{(\text{tr},j)}$ está próxima do centróide da classe c_i mas a dispersão de c_i é pequena de forma que $\beta_{(\text{tr},j)}$ se localize distante da região de dispersão típica, então a distância entre $\beta_{(\text{tr},j)}$ e c_i será razoavelmente grande (ver exemplo na Figura 4.2).

Finalmente, a centralidade do elemento $\beta_{(\text{tr},j)}$ do conjunto de treinamento com relação à classe c_i é dada por

$$u_{ij} = \frac{\psi(i, \beta_{(\text{tr},j)})}{\sum_i \psi(i, \beta_{(\text{tr},j)})}. \tag{4.9}$$

Figura 4.2 – Relação entre a distância euclidiana $d^{(euc)}$ e a distância Mahalanobis $d^{(mah)}$ entre o ponto p e as classes “vermelho” e “azul”. $d^{(euc)}(p, \text{vermelho}) / d^{(euc)}(p, \text{azul}) = 1,15$ e $d^{(mah)}(p, \text{vermelho}) / d^{(mah)}(p, \text{azul}) = 0,62$. Embora as distâncias euclidianas de p aos centróides dos dois grupos sejam similares, p é mais similar ao grupo vermelho segundo a medida Mahalanobis porque este último possui uma maior variação.

Na equação 4.9, $\psi(i, \beta_{(tr,j)})$ é normalizado pela soma das similaridades entre $\beta_{(tr,j)}$ para satisfazer a equação 4.1.

4.1.2 Árvores de decisão - C4.5

Para problemas de decisão usando árvores, o objetivo é determinar uma estrutura de árvore capaz de classificar instâncias desconhecidas conforme a árvore é percorrida da raiz até a folha. Diferentemente do algoritmo κ -NN, árvores de decisão não apenas classificam instâncias desconhecidas de dados, mas também sumarizam os padrões presentes nos dados de treinamento. Vértices representam questões do tipo “SE condição ENTÃO” e arestas direcionam a resposta à condição. Quando um vértice folha é alcançado, um exemplo desconhecido $\beta_{(ts)} \in \mathcal{S}_{ts}$ é classificado com a classe $c_i \in \mathcal{C}$ armazenada naquele vértice. Em resumo, uma árvore de decisão é caracterizada pelas seguintes características:

1. Cada vértice interno corresponde a um teste realizado no atributo F_k da instância $\beta_{(ts)} \in \mathcal{S}_{ts}$.
2. Arestas representam possíveis valores dos atributos.
3. Um vértice folha especifica a categoria c_i atribuída ao exemplo desconhecido $s \in \mathcal{S}_{ts}$.

O processo de classificação de $\beta_{(ts)}$ é exemplificado na Figura 4.3.

Para construir uma árvore de decisão, o algoritmo C4.5 escolhe inicialmente quais atributos devem ser testados com maior precedência no momento da classificação. A ideia

Figura 4.3 – Exemplo de árvore de decisão para classificar uma instância $\beta_{(ts)} \in \mathcal{S}_{ts}$ em uma das classes do conjunto $\mathcal{C} = \{c_1, c_2, c_3\}$. Suponha que a instância seja composta dos seguintes atributos: $\beta_{(ts)}^{(1)} = -0,23$, $\beta_{(ts)}^{(2)} = +0,29$ e $\beta_{(ts)}^{(3)} = +0,38$. O primeiro teste leva à aresta “NÃO” e o segundo teste leva à aresta “SIM”. Por este motivo a instância desconhecida é classificada como pertencendo à classe c_3 .

é selecionar a medida que fornece a melhor separação dos dados no princípio da construção das árvores. Tendo selecionado a melhor medida, o processo é repetido iterativamente até que o aparecimento de um subconjunto do conjunto de treinamento onde existe apenas uma classe remanescente. Para selecionar a melhor medida, o conjunto de treinamento $\mathcal{S}_{tr}$ é avaliado através da medida de ganho de informação (61) (ou divergência Kullback-Leibler (66)), denotado por Ω . Seja $\Omega(\mathcal{S}_{tr}, F_k)$ o ganho de informação do atributo $F_k \in \mathcal{F}$ relativo ao conjunto $\mathcal{S}_{tr}$. Para definir Ω , considere que $V(F_k)$ representa o conjunto de todos os valores possíveis do atributo F_k no conjunto de treinamento. Em outras palavras,

$$V(F_k) = \bigcup_{i=1}^{\|\mathcal{S}_{tr}\|} \beta_{(tr,i)}^{(k)}. \quad (4.10)$$

Se $\mathcal{H}(\mathcal{S}_{tr})$ é a entropia total de $\mathcal{S}_{tr}$, então o ganho de informação do atributo F_k é

$$\Omega(\mathcal{S}_{tr}, F_k) = \mathcal{H}(\mathcal{S}_{tr}) - \mathcal{H}(\mathcal{S}_{tr}|F_k), \quad (4.11)$$

onde

$$\mathcal{H}(\mathcal{S}_{tr}|F_k) = \sum_{v \in V(F_k)} \frac{\|\beta_{(tr)} \in \mathcal{S}_{tr} | \beta_{(tr)}^{(k)} = v\|}{\|\mathcal{S}_{tr}\|} \cdot \mathcal{H}(\{\beta_{(tr)} \in \mathcal{S}_{tr} | \beta_{(tr)}^{(k)} = v\}). \quad (4.12)$$

4.1.3 Naive-Bayes

Diferentemente dos algoritmos κ -NN e C4.5, esta técnica de reconhecimento de padrões é baseada em um paradigma estatístico. Para classificar uma nova instância, o algoritmo estima a distribuição de probabilidade de cada uma das classes no conjunto $\mathcal{C}$ e utiliza a estratégia de estimação conhecida como *Maximum a Posteriori* para inferir a classe correta. A probabilidade de uma dada classe $c_i \in \mathcal{C}$ ser a classe correta (a classe na qual o exemplo $\beta \in \mathcal{S}_{ts}$ deve ser classificado) é dada por:

$$P(c_i | \vec{\beta}) = \frac{P(\vec{\beta} | c_i) P(c_i)}{P(\vec{\beta})} = \frac{P(F_1 = \beta^{(1)}, F_2 = \beta^{(2)}, \dots, F_M = \beta^{(M)} | c_i) P(c_i)}{P(F_1 = \beta^{(1)}, F_2 = \beta^{(2)}, \dots, F_M = \beta^{(M)})}. \quad (4.13)$$

Na equação 4.13, $P(c_i)$ pode ser estimada como $\mathcal{N}(c_i) / \sum_{c_i \in \mathcal{C}} \mathcal{N}(c_i)$, onde $\mathcal{N}(c_i)$ representa o número de objetos em $\mathcal{S}_{tr}$ que estão rotulados com a classe c_i . Para fins de classificação, o valor $P(\vec{\beta})$ pode ser ignorado já que é constante para toda classe $c_i \in \mathcal{C}$. Finalmente, para estimar $P(\vec{\beta} | c_i)$, o algoritmo *Naive Bayes* presume que os atributos descrevendo as classes são independentes. Dessa forma, o processo de estimação de $P(\vec{\beta} | c_i)$ é calculado como

$$P(\vec{\beta} | c_i) = P(F_1 = \beta^{(1)}, F_2 = \beta^{(2)}, \dots, F_M = \beta^{(M)} | c_i) = \prod_{k=1}^M P(F_k = \beta^{(k)} | c_i). \quad (4.14)$$

Substituindo a equação 4.14 em 4.13 temos que

$$P(c_i | \vec{\beta}) = \frac{P(c_i)}{P(F_1 = \beta^{(1)}, \dots, F_M = \beta^{(M)})} \prod_{k=1}^M P(F_k = \beta^{(k)} | c_i). \quad (4.15)$$

Usando a regra *Maximum a Posteriori*, a classe c_s do exemplo é dada por

$$c_\beta = \arg \max_{c_i \in \mathcal{C}} P(c_i) \prod_{k=1}^M P(F_k = \beta^{(k)} | c_i). \quad (4.16)$$

A partir da análise da equação 4.16, percebe-se que o último parâmetro a ser estimado é $P(F_k | c_i)$. Especificamente nesta tese utilizamos uma estimativa de *kernel* (65), gerando gaussianas cujas médias e desvios correspondam às médias e desvios observados nos dados de cada classe. A fim de ilustrar o processo de decisão bayesiana com o método de estimativa de *kernel* (65), considere a Figura 4.4, sobre o problema de discriminação de duas classes caracterizadas por uma única medida. As classes consideradas são $\mathcal{C} = \{\text{"vermelho"}, \text{"azul"}\}$. Uma vez que a frequência de cada uma das classes no conjunto de treinamento é idêntica, o termo efetivo na equação 4.16 para classificação é $P(F_k = \beta^{(k)} | c_i)$, onde $\{c_i \in \mathcal{C}\}$. Dessa forma, as fronteiras de decisão são estabelecidas onde

$P(F|\text{“vermelho”}) = P(F|\text{“azul”})$. No exemplo da Figura 4.4, qualquer instância de teste satisfazendo a condição $-0,19 < \beta_{ts} < 0,00$ será classificada na classe $c_\beta = \text{“azul”}$ pois esta classe é a mais provável nesta região. Analogamente, qualquer instância de teste satisfazendo $0,00 < \beta_{ts} < 0,18$ será classificada com a classe $c_\beta = \text{“vermelho”}$ porque $P(F|\text{“vermelho”}) > P(F|\text{“azul”})$.

Figura 4.4 – Exemplo de decisão baseada no algoritmo *Naive Bayes*. Qualquer instância de teste satisfazendo a condição $P(F|\text{“vermelho”}) > P(F|\text{“azul”})$ será classificada com a classe $c_\beta = \text{“vermelho”}$. Analogamente, a condição $P(F|\text{“azul”}) > P(F|\text{“vermelho”})$ determina as fronteiras da classe $c_\beta = \text{“azul”}$.

O método de classificação *Naive Bayes* apresenta desempenhos razoáveis apesar da suposição de independência dos atributos^a. Provavelmente as taxas de classificação se mantêm altas mesmo na simplificação imposta porque os valores de $P(c_i|\vec{\beta})$, apesar de aproximados, são utilizados apenas para discriminar classes (67). Ressalta-se que este classificador pode ser visto como um classificador *fuzzy* (68) já que $P(c_s|\vec{\beta})$ pode ser utilizado como nível de confiança do resultado da classificação.

4.1.4 Classificação híbrida

Classificadores tradicionais são baseados nas relações de proximidade dos dados. O classificador híbrido (64) utiliza também a informação de organização estrutural dos dados pertencendo às classes. Na fase de treinamento, os dados no conjunto de treinamento $\mathcal{S}_{tr}$ são mapeados em um grafo $\mathcal{G}$ usando uma técnica de formação de rede $\mathcal{S}_{tr} \mapsto \mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, onde $\mathcal{V}$ representa o conjunto de vértices e $\mathcal{E}$ representa o conjunto de arestas. Cada

^a Para ilustrar como nem sempre os atributos são independentes, considere o problema de desambiguação de sentido de palavras onde o sentido é dado pela frequência de palavras vizinhas. Portanto cada atributo é representado pela frequência de uma palavra vizinha. Dessa forma, por exemplo, a palavra “sol” é mais provável de ocorrer num dado contexto com a palavra “calor” do que num contexto que contém a palavra “poeta”.

Figura 4.5 – Esquema de construção da rede na fase de treinamento exemplificado com o conjunto de classes $\mathcal{C} = \{\text{“azul”}, \text{“cinza”}, \text{“vermelho”}\}$. Nesta fase, é permitida apenas a conexão de vértices pertencendo à mesma classe. Por este motivo, a conexão entre os vértices **A** e **B** não foi concretizada, muito embora estejam próximos. Na fase de classificação, quando uma nova instância (quadrado verde) pertencente ao conjunto $\mathcal{S}_{ts}$ é inserida, ela é conectada com as instâncias do $\mathcal{S}_{tr}$ independentemente da classe. Como consequência da inserção da instância de teste (quadrado verde), a topologia das classes “azul” e “cinza” se alteram. Esta perturbação na topologia é utilizada para associar ao exemplo $\beta_{(ts)} \in \mathcal{S}_{ts}$ à classe cuja perturbação observada foi a mínima dentre todas as outras classes.

vértice em $\mathcal{V}$ é uma instância de treinamento. As arestas em $\mathcal{E}$ são criadas usando uma combinação de duas técnicas. Na primeira técnica, denominada ϵ_r , vértices dentro de um raio ϵ_r são conectados. Na segunda, denominada knn , todos os k vértices mais próximos são conectados, independentemente da distância. Ambas as abordagens possuem suas limitações quando os dados são distribuídos mais densamente ou de maneira mais esparsa. Em regiões esparsas, a técnica knn pode conectar vértices muito distantes enquanto a utilização de valores não adequados para ϵ_r pode gerar a criação de componentes desconectados. Para resolver esses problemas, a rede é construída combinando as abordagens, de tal forma que a técnica ϵ_r seja usada em regiões densas e a técnica knn seja empregada em regiões esparsas. Para gerar uma representação única de cada classe, na fase de treinamento permite-se apenas a conexão de vértices pertencendo à mesma classe. Com este mecanismo, cada classe é representada por um único componente do grafo. Para exemplificar o processo de construção da rede, a Figura 4.5 mostra um esquema da rede para um problema de três classes após a fase de treinamento. Note que cada classe $c_i \in \{\text{“azul”}, \text{“cinza”}, \text{“vermelho”}\}$ é representada como um único componente.

Durante a fase de classificação, os dados não classificados no conjunto de testes $\mathcal{S}_{ts}$ são

apresentados aos classificadores separadamente. Cada instância é inserida usando a técnica tradicional ϵ_r , ou seja, novas instâncias são conectadas a cada vértice dentro do raio, independentemente da classe do vértice. Após a inserção do dado, o impacto estrutural da inserção em cada uma das classes do problema é verificado. Mais especificamente, o impacto é analisado com métricas de redes complexas. Se pouca mudança ocorre, então a instância inserida é consistente com o padrão daquela classe e consequentemente haverá uma probabilidade maior de se associar a instância àquela classe. Por outro lado, se a inclusão modifica fortemente o padrão da classe, então a probabilidade de se associar aquela classe à instância será baixa. Para exemplificar o processo, a Figura 4.5 exhibe o esquema ilustrando como a classificação é feita.

Além da classificação baseada na organização topológica dos dados no espaço de atributos (classificação de alto nível), utiliza-se um classificador tradicional para fins de combinação de estratégias. As classificações realizadas pelas estratégias de alto nível e tradicional são combinadas através de uma combinação linear para gerar a classificação do arcabouço híbrido. A combinação de estratégias baseia-se na associação convexa de dois termos: (i) um classificador tradicional em sua versão *fuzzy* (k -NN, C4.5 ou *Naive Bayes*) e o classificador baseado na organização topológica dos dados. Matematicamente, a probabilidade $P(c_i|\beta)$ de que a instância de teste $\beta_{(ts)} \in \mathcal{S}_{ts}$ pertença à classe $c_i \in \mathcal{C}$ é:

$$P(c_i|\vec{\beta}) = (1 - \lambda)P_T(c_i|\vec{\beta}) + \lambda P_H(c_i|\vec{\beta}), \quad (4.17)$$

onde $P_T(c_i|\vec{\beta})$ representa a probabilidade associada por um classificador tradicional de que a instância de teste $\beta_{(ts)}$ pertença à classe c_i ; $P_H(c_i|\vec{\beta})$ representa a mesma quantidade calculada segundo os fatores topológicos do classificador de alto nível e $\lambda \in [0, 1]$ representa a ponderação associada à estratégia de alto nível. Note que quando $\lambda = 0$, a classificação é restrita à abordagem tradicional. Especificamente, $\mathcal{C}_i^{(j)}$ é calculado como

$$P_H(c_i|\vec{\beta}) = \frac{\sum_{u=1}^{\mathcal{M}} \alpha(F_u) [1 - f_{c_i}^{(\beta)}(F_u)]}{\sum_{g \in \mathcal{C}} \sum_{u=1}^{\mathcal{M}} \alpha(F_u) [1 - f_g^{(\beta)}(F_u)]}, \quad (4.18)$$

onde $\alpha(F_u) \in [0, 1], \forall u \in \{1, \dots, \mathcal{M}\}$, $\sum_{u=1}^{\mathcal{M}} \alpha(F_u) = 1$ é o peso associado a cada uma das medidas e $f_{c_i}^{(\beta)}(F_u)$ é uma função que depende da u -ésima medida de rede aplicada à instância β da classe c_i , sendo dada por

$$f_{c_i}^{(\beta)}(F_u) = \Delta G_{c_i}^{(\beta)}(F_u) p(c_i) \quad (4.19)$$

onde $\Delta G_{c_i}^{(\beta)}(F_u) \in [0, 1]$ é a variação da u -ésima medida no componente da classe c_i se β é incluído nesta classe e $p(c_i) \in [0, 1]$ é a proporção de dados pertencendo à classe c_i . De acordo com a definição da equação 4.19, se existe uma pequena mudança nas medidas de rede, presume-se que a instância de teste é consistente com todos os outros dados da

mesma classe.

As medidas que fazem parte do classificador de alto nível são a assortatividade r (equação 2.11), o coeficiente de aglomeração médio $\langle C \rangle$ (calculado a partir da equação 2.10) e o grau médio $\langle k \rangle$ (calculado a partir da equação 2.3). Enquanto a medida de grau fornece informação exclusivamente local, a medida de assortatividade analisa a rede de modo global. No meio termo, a medida de aglomeração fornece informação a respeito da conectividade dos vizinhos do vértice. Dessa forma, estas três medidas caracterizam a estrutura da rede tanto do ponto de vista local como do ponto de vista global.

4.2 Avaliação da classificação supervisionada

O desempenho de qualquer classificador depende fortemente das características dos dados a serem classificados. Não existe algoritmo que forneça desempenho otimizado para todos os tipos de distribuição de dados. Muitos métodos de avaliação têm sido utilizados para comparar o desempenho de classificadores. Por questão de simplicidade usamos como taxa de qualidade a porcentagem de acerto com a estratégia *10-fold cross-validation* (61). Uma das vantagens desta estratégia é que ela pode ser usada mesmo quando não existe um conjunto específico para avaliar qualidade. Neste método o conjunto de treinamento é dividido aleatoriamente em 10 partes, de forma que 9 partes sejam usadas para treinamento e 1 para avaliação. O processo é repetido 10 vezes iterativamente até que todas as partes tenham sido usadas para avaliação. No final, a taxa de acerto é computada como a taxa de acerto média obtida durante este processo iterativo.

4.3 Avaliação da classificação não supervisionada

A avaliação de qualidade da classificação não supervisionada é denominada na literatura como validação de agrupamento (ou *cluster validation*). Vários métodos têm sido descritos na literatura (69). Nas Seções 4.3.1 e 4.3.2 descrevemos dois dos critérios mais usados (69): silhueta simplificada (SWC) e índice Dunn (DN), respectivamente.

4.3.1 Silhueta simplificada (SWC)

Para definir este critério, considere as seguintes definições. Suponha a existência de um conjunto de agrupamentos $\mathcal{K}$. $\Lambda(\beta)$ representa a dissimilaridade da instância β ao

centróide de seu cluster $\kappa \in \mathcal{K}$ e é dado por

$$\Lambda(\beta) = \left\| \vec{\beta} - \frac{1}{\mathcal{N}(\kappa)} \sum_{\beta' \in \kappa} \vec{\beta}' \right\|, \quad (4.20)$$

onde $\mathcal{N}(\kappa)$ é o número de objetos no agrupamento κ . $\zeta(\beta)$, dado na equação 4.21, quantifica a dissimilaridade entre o objeto $\beta \in \kappa$ e o agrupamento mais próximo $\rho(\kappa)$.

$$\zeta(\beta) = \left\| \vec{\beta} - \frac{1}{\mathcal{N}(\rho(\kappa))} \sum_{\beta' \in \rho(\kappa)} \vec{\beta}' \right\|, \quad (4.21)$$

Na equação 4.21, $\mathcal{N}(\rho(\kappa))$ representa o número de objetos no agrupamento $\rho(\kappa)$, que é dado por:

$$\rho(\kappa) = \arg \min_{x \in \mathcal{K} - \{\kappa\}} \left\| \frac{1}{\mathcal{N}(\kappa)} \sum_{\beta \in \kappa} \vec{\beta} - \frac{1}{\mathcal{N}(x)} \sum_{\beta' \in x} \vec{\beta}' \right\| \quad (4.22)$$

Com estas definições, a qualidade do agrupamento segundo a métrica de silhueta simplificada SWC é dada por:

$$SWC = \left(\sum_{\kappa \in \mathcal{K}} \mathcal{N}_{\kappa} \right)^{-1} \sum_{i \in \cup_{\kappa \in \mathcal{K}} \kappa} \frac{\zeta(i) - \Lambda(i)}{\max\{\zeta(i), \Lambda(i)\}}. \quad (4.23)$$

De acordo com a equação 4.23, o índice SWC fica restrito ao intervalo $[-1, 1]$. Quanto maior o seu valor, maior será a qualidade do agrupamento.

4.3.2 Índice Dunn (DN)

O índice Dunn (DN) (69) é definido como a razão entre a maior distância inter-grupo δ e a menor distância intra-grupo Δ :

$$DN = \min_{\kappa, \kappa' \in \mathcal{K} \text{ e } \kappa \neq \kappa'} \left\{ \frac{\delta_{\kappa, \kappa'}}{\max_{\kappa'' \in \mathcal{K}} \Delta_{\kappa''}} \right\}. \quad (4.24)$$

A distância $\delta_{\kappa, \kappa'}$ entre os grupos κ e κ' é

$$\delta_{\kappa, \kappa'} = \left\| \frac{1}{\mathcal{N}(\kappa)} \sum_{\beta \in \kappa} \vec{\beta} - \frac{1}{\mathcal{N}(\kappa')} \sum_{\beta' \in \kappa'} \vec{\beta}' \right\| \quad (4.25)$$

e a distância intra-grupos Δ_{κ} para o grupo κ é:

$$\Delta_{\kappa} = \frac{2}{\mathcal{N}(\kappa)} \sum_{\beta \in \kappa} \left\| \vec{\beta} - \frac{1}{\mathcal{N}(\kappa)} \sum_{\beta' \in \kappa} \vec{\beta}' \right\|. \quad (4.26)$$

Similarmente à definição da silhueta simplificada, o valor do índice Dunn é diretamente proporcional à qualidade do agrupamento. No entanto, sua variação não é restrita a um intervalo limitado.

4.4 Relevância de atributos

No processo de reconhecimento de padrões muitas vezes é importante identificar a relevância dos atributos para a tarefa de classificação ou *clustering*. Nos casos em que há grande variedade de atributos, a remoção dos atributos menos relevantes pode ser fundamental para amenizar os problemas causados pela “maldição da dimensionalidade” (61). Além disso, a escolha de atributos relevantes pode aperfeiçoar a capacidade de generalização, aumentar a eficiência do processo de aprendizado e facilitar a interpretação do modelo gerado. Duas técnicas de quantificação de relevância de atributos foram usadas neste projeto. A primeira é baseada na divergência Kullback-Leibler (KL) e a segunda é baseada na frequência de aparecimento do atributo nos melhores classificadores.

4.4.1 Método KL

Neste método, os atributos são avaliados isoladamente, ou seja, a relevância dos atributos independe da sua eficiência quando combinado com outros atributos. A medida característica deste método, denominada divergência Kullback-Leibler (Ω), é usada em algoritmos de árvores de decisão (61) para auxiliar na escolha dos atributos ao longo do processo de formação da árvore. A ideia é que atributos relevantes fornecem um alto ganho de informação já que a especificação de seu valor diminui a incerteza referente à escolha das classes. Nesta tese, o método KL atribui relevância a um atributo F_k no conjunto de treinamento $\mathcal{S}_{tr}$ usando o ganho de informação $\Omega(\mathcal{S}_{tr}, F_k)$ definido na equação 4.11.

Aqui Ω assume altos valores quando F_k correlaciona bem com a função inferida. Portanto, a relevância do atributo F_k é proporcional a $\Omega(\mathcal{S}_{tr}, F_k)$. A quantificação da relevância de atributos com Ω tem a vantagem de a relevância independe do algoritmo para induzir classificadores. No entanto, como apontado anteriormente, ela não considera o relacionamento entre duas ou mais medidas. A combinação de atributos é considerada no método AUC, apresentado a seguir.

Tabela 4.1 – Exemplo de computação de relevância de atributos em um problema com os atributos **A**, **B** e **C**. Os classificadores estão organizados por ordem decrescente de precisão e os atributos usados em cada uma das $2^3 = 8$ combinações de classificadores possíveis com 3 atributos são mostrados em cada coluna. O classificador mais eficiente na segunda coluna da tabela mostra que apenas o atributo **B** foi utilizado. O atributo **B** aparece frequentemente entre os melhores classificadores, possuindo maior relevância no problema. A curva de crescimento de cada um dos atributos é mostrada na Figura 4.6.

Ranking	1	2	3	4	5	6	7	8
Atributo A		•	•		•			•
Atributo B	•	•	•				•	
Atributo C			•			•	•	•

4.4.2 Método AUC

O cálculo da relevância de um atributo no método AUC (*Area Under the Curve*) é baseado na sua frequência de aparição entre os classificadores mais precisos, considerando todas as combinações de atributos. Conta-se a fração das combinações de atributos que incluem F_k com taxa de acerto maior ou igual a um limiar. A relevância é calculada como a área embaixo da curva que se obtém variando-se o limiar do máximo valor possível para o menor valor. Emprega-se este método porque ele avalia a importância de um atributo específico considerando suas possíveis combinações com outros atributos. O método depende do algoritmo de classificação já que a ordenação de classificadores é baseada na taxa de acerto fornecida na classificação.

Um exemplo da aplicação do método AUC é ilustrado na Tabela 4.1, em que os classificadores estão organizados em ordem decrescente de precisão. Por exemplo, o classificador construído apenas com o atributo **B** obteve a maior taxa de acerto, o classificador construído com os atributos **A** e **B** obteve a segunda maior taxa de acerto e assim por diante. Os gráficos com a frequência acumulada para cada atributo são apresentados na Figura 4.6. Note que como o atributo **B** aparece mais frequentemente nas primeiras posições, sua área embaixo da curva é maior do que para os outros atributos.

4.5 Métodos de projeção

Na área de reconhecimento de padrões, métodos de projeção são essenciais para visualizar os dados multidimensionais. Além disso, técnicas de redução de dimensionalidade são úteis para amenizar problemas decorrentes da análise em altas dimensões, como a presença de dados esparsos que não fornecem significância estatística para reconhecimento de padrão. A seguir, dois métodos de redução de dimensionalidade são introduzidos.

Figura 4.6 – Exemplo de quantificação de relevância dos atributos **A** (esquerda), **B** (centro) e **C** (direita) usando o método AUC. A ordem de qualidade assumida pelos classificadores é dada na Tabela 4.1. Note que o atributo **B** cresce mais rapidamente pois aparece nos classificadores com maior qualidade. Por outro lado, o atributo **C** é considerado como o menos relevante já que aparece majoritariamente na cauda da Tabela 4.1.

4.5.1 Análise de componentes principais (PCA)

A análise de componentes principais (10, 65) (ou *Principal Component Analysis*) (PCA) é uma transformação matemática com uma mudança ortogonal para mapear um conjunto de observações de variáveis possivelmente correlacionadas em variáveis fracamente correlacionadas. A transformação é realizada para que a distribuição de pontos no novo espaço de fase gerado apresente máxima variância ao longo do componente principal (10, 65). O número de eixos novos (componentes principais) é sempre menor ou igual ao conjunto original de atributos. Na verdade, a quantidade final de eixos depende da precisão da projeção fornecida por este método. Suponha que o vetor de atributos seja representado por $\vec{\beta} = (\beta^{(1)}, \dots, \beta^{(\mathcal{M})})$. A matriz de covariância dos dados é:

$$\text{Cov} = \frac{(\vec{\beta} - \langle \vec{\beta} \rangle)(\vec{\beta} - \langle \vec{\beta} \rangle)^T}{n}, \quad (4.27)$$

onde n é o tamanho do banco de dados e $\langle \vec{\beta} \rangle$ é o vetor onde cada elemento é a média dos respectivos atributos. Uma vez que Cov 00 é uma matriz real de dimensões quadradas $\mathcal{M} \times \mathcal{M}$ e simétrica, ela possui um conjunto de autovalores χ_i positivos e decrescentes (i.e., $\chi_1 > \chi_2 \dots \chi_{\mathcal{M}} > 0$). Consequentemente, os autovetores $\vec{\nu}_i$ são ortogonais. Considere a matriz de transformação Υ :

$$\Upsilon = \begin{bmatrix} \leftarrow \vec{\nu}_1 \rightarrow \\ \leftarrow \vec{\nu}_2 \rightarrow \\ \leftarrow \vec{\nu}_3 \rightarrow \\ \vdots \\ \leftarrow \vec{\nu}_{\mathcal{M}} \rightarrow \end{bmatrix}. \quad (4.28)$$

Com esta matriz, é possível rotacionar o conjunto de dados original de acordo com a seguinte transformação:

$$\vec{\beta}_{\Upsilon} = \Upsilon \vec{\beta} \quad (4.29)$$

A redução de dimensionalidade é obtida selecionando apenas os primeiros q autovetores para fazer parte de Υ . Neste caso, a variabilidade dos dados que é mantida é dada por:

$$\vartheta_q = \sum_{i=1}^q \chi_i / \sum_{j=1}^{\mathcal{M}} \chi_j. \quad (4.30)$$

4.5.2 Análise de variáveis canônicas (CVA)

O método da análise de variáveis canônicas (10, 65) (*canonical variable analysis*) (CVA) é similar ao método de análise de componentes principais, já que também é um método de projeção e redução de dimensionalidade dos dados. Como extensão do PCA, este método considera a classe dos dados a fim de projetá-los de tal maneira que a separação das classes seja maximizada. Para descrever este método, considere a seguinte notação. A dispersão de cada classe é definida como

$$\Phi(c_i) = \sum_{\vec{\beta} \in c_i} \left(\vec{\beta} - \langle \vec{\beta} \rangle_{c_i} \right) \left(\vec{\beta} - \langle \vec{\beta} \rangle_{c_i} \right)^T \quad (4.31)$$

e a dispersão intra-classes é

$$\Phi_{(\text{intra})} = \sum_{c_i \in \mathcal{C}} \Phi(c_i). \quad (4.32)$$

A dispersão entre categorias distintas, denotada por $\Phi_{(\text{inter})}$, é definida matematicamente como:

$$\Phi_{(\text{inter})} = \sum_{c_i \in \mathcal{C}} \left(\langle \vec{\beta} \rangle_{c_i} - \langle \vec{\beta} \rangle \right) \left(\langle \vec{\beta} \rangle_{c_i} - \langle \vec{\beta} \rangle \right)^T. \quad (4.33)$$

O método minimiza a dispersão intra-classe e maximiza a dispersão inter-classes através da transformação:

$$\vec{\beta}_{\Upsilon} = \Upsilon \vec{\beta}, \quad (4.34)$$

onde Υ é a matriz diagonal contendo os autovetores da matriz $\Phi_{(\text{intra})}^{-1} \Phi_{(\text{inter})}$.

Capítulo 5

Análise das propriedades estatísticas de textos

Para compreender as propriedades de um sistema complexo muitas vezes é interessante estudar seu comportamento em comparação com um sistema gerado por um processo aleatório (70). Em termos de redes textuais, este estudo pode ser feito comparando a estrutura topológica das redes complexas com a estrutura de uma rede obtida de um texto aleatório equivalente, (i.e., um texto construído a partir do embaralhamento das palavras no sistema original). Este tipo de análise é fundamental para inferir, por exemplo, a presença/ausência de informação nas métricas que captam características textuais que vão além da frequência das palavras. Na prática, a identificação de medidas informativas pode ser útil na investigação da veracidade de manuscritos codificados uma vez que medidas informativas podem ser usadas para averiguar se o texto desconhecido foi escrito com padrões semelhantes a um texto aleatório sem significado.

Medidas extraídas de redes complexas modelando textos também podem ser estudadas em termos da sua dependência com fatores sintáticos e semânticos. Medidas de natureza predominantemente sintática supostamente devem variar quando o idioma varia, para um dado texto. Por outro lado, medidas de natureza predominantemente semântica supostamente possuem maior variabilidade para textos distintos escritos numa mesma língua em comparação com um texto fixo escrito em várias línguas. Definir a natureza das medidas pode ser fundamental também para a análise de textos codificados. Neste caso, medidas dependentes da sintaxe são úteis para identificar o idioma mais similar ao idioma do texto desconhecido, enquanto medidas de natureza semântica são úteis para capturar características relacionadas ao conteúdo do texto, como palavras-chave.

Neste capítulo apresentamos um arcabouço de análise de informatividade e dependência das métricas de redes complexas extraídas de textos. Também apresentamos uma metodologia para seleção de palavras-chave de manuscritos baseada na frequência e intermitência das palavras. As propriedades estatísticas das palavras são estudadas em uma análise

comparativa envolvendo o mesmo livro (o Novo Testamento da Bíblia) traduzido em 15 idiomas. Também foi usado um conjunto de 15 textos variados escritos em Inglês e Português a fim de verificar a variação das medidas com temáticas diferentes em uma mesma língua. Como aplicação do arcabouço, o famoso manuscrito Voynich (VMS) (71), cujo conteúdo é desconhecido, foi escolhido como estudo de caso. Por fim, este mesmo arcabouço é usado para testar a compatibilidade do VMS com línguas naturais e para identificar o tópico do texto, que pode ser útil para guiar o processo de decifração.

5.1 Metodologia de análise

As propriedades das medidas de texto empregadas nesta tese foram investigadas em duas dimensões de variabilidade. Medidas estatísticas de redes complexas podem variar devido (i) aos diferentes padrões sintáticos típicos de línguas diferentes e (ii) devido à semântica ou estilo do texto, para uma dada língua. Para estudar o comportamento das métricas com relação a (i), versões do Novo Testamento da Bíblia foram empregadas nos seguintes idiomas: Árabe, Basco, Inglês, Esperanto, Alemão, Grego, Hebraico, Húngaro, Coreano, Latim, Português, Russo, Suaíli, Vietnamita e Xhosa. O Novo Testamento foi escolhido para este tipo de análise pois é um dos poucos textos longos disponíveis em vários idiomas. Para estudar a variabilidade do tipo (ii) um conjunto de 15 textos em Português e 15 textos em Inglês foi utilizado. A lista de livros em Português e em Inglês é apresentada nas Tabelas 5.1 e 5.2, respectivamente. Neste caso, a escolha dos textos guiou-se pela tentativa de seleção de textos longos (para fornecer uma análise estatística significativa) tratando de assuntos diferentes para conferir uma maior variabilidade semântica dentro do gênero selecionado. Excepcionalmente na análise realizada no estudo de livros, o pré-processamento do texto descrito na Seção 3.4 (remoção de *stopwords* e lematização) não foi aplicado já que a abordagem de análise de métricas neste capítulo é independente do idioma. Por esse motivo esta metodologia pode ser aplicada até mesmo em textos cujo idioma é desconhecido, como é o caso do manuscrito Voynich.

5.1.1 Normalização das medidas

Uma vez que estamos interessados em medidas capazes de distinguir um texto com significado de suas versões aleatórias, cada uma das medidas X , $\langle X \rangle_2$ e $\gamma(X)$ foi normalizada pela média obtida em 10 textos aleatorizados (i.e., a ordem das palavras é trocada mas a frequência é mantida). Se $\mu(X_R)$ e $\sigma(X_R)$ são respectivamente a média e o desvio em 10 diferentes realizações de textos aleatorizados, a medida normalizada $\tilde{X}$ relacionada

Tabela 5.1 – Lista de livros em Inglês (ano de publicação, título e autor) para analisar a variabilidade de métricas textuais. A escolha dos textos guiou-se pela tentativa de seleção de textos longos tratando de assuntos diferentes a fim de conferir maior variabilidade semântica para o gênero estudado.

Ano	Título	Autor
1605	<i>The Advancement of Learning</i>	Francis Bacon
1612	<i>True Christianity</i>	Johann Arndt
1624	<i>Devotions upon Emergent Occasions</i>	John Donne
1651	<i>Leviathan</i>	Thomas Hobbes
1726	<i>Gulliver's Travels</i>	Jonathan Swift
1748	<i>Life's Progress through the Passions</i>	Eliza Haywood
1771	<i>The Expedition of Humphry Clinker</i>	Tobias Smollett
1792	<i>A Vindication of the Rights of Woman</i>	Mary Wollstonecraft
1811	<i>Sense and Sensibility</i>	Jane Austen
1826	<i>The Last of the Mohicans</i>	James Fenimore Cooper
1847	<i>Wuthering Heights</i>	Emily Brontë
1878	<i>The Return of the Native</i>	Thomas Hardy
1909	<i>The Lady of the Shroud</i>	Bram Stoker
1913	<i>The New Freedom</i>	Woodrow Wilson
1920	<i>The Age of Innocence</i>	Edith Wharton

à normalização de X é:

$$\tilde{X} = \frac{X}{\mu(X_R)}. \quad (5.1)$$

Dado que $\mu(X_R)$ apresenta um erro $\epsilon(\mu(X_R)) = \sigma(X_R)$ devido ao desvio observado nas realizações aleatórias, o erro $\epsilon(\tilde{X})$ observado na variável $\tilde{X}$ é dado por:

$$\epsilon(\tilde{X}) = \left| \frac{d\tilde{X}}{d\mu(X_R)} \right| \epsilon(\mu(X_R)) = \left| -\frac{X}{\mu^2(X_R)} \right| \sigma(X_R) = \frac{\sigma(X_R)}{\mu^2(X_R)} X = \frac{\sigma(X_R)}{\mu(X_R)} \tilde{X}. \quad (5.2)$$

A normalização efetuada na equação 5.1 é útil porque permite comparar cada medida como um modelo nulo. Dessa forma, uma medida fornece informação significativa somente se seu valor $\tilde{X}$ não é próximo de $\tilde{X} = 1$. Além disso, a influência do tamanho do vocabulário em outras medidas tende a ser minimizada.

5.1.2 Variabilidade entre línguas e textos

As medidas usadas nesta tese variam de texto para texto devido às propriedades sintáticas da linguagem. Em uma dada linguagem, existe também uma variação óbvia entre textos devido aos fatores estilísticos e semânticos. Dessa forma, em uma primeira aproximação podemos assumir que variações entre textos de uma medida $\tilde{X}$ ocorre em duas dimensões. Considere que $\tilde{X}_{t,l}$ denota o valor de $\tilde{X}$ para o texto t escrito na língua l . Se

Tabela 5.2 – Lista de livros em Português (ano de publicação, título e autor) para analisar a variabilidade de métricas textuais. A escolha dos textos guiou-se pela tentativa de seleção de textos longos tratando de assuntos diferentes a fim de conferir maior variabilidade semântica para o gênero estudado.

Ano	Título	Autor
1845	Moço Loiro	Joaquim Manuel de Macedo
1848	Os Dois Amores	Joaquim Manuel de Macedo
1862	Amor de Perdição	Camilo Castelo Branco
1863	Pupilas do Senhor Reitor	Júlio Dinis
1875	A Escrava Isaura	Bernardo Guimarães
1876	Helena	Machado de Assis
1879	O Sacrifício	Franklin Távora
1884	Casa de Pensão	Aluísio Azevedo
1887	A Relíquia	Eça de Queirós
1888	Os Maias	Eça de Queirós
1891	O Missionário	Inglês de Souza
1899	Dom Casmurro	Machado de Assis
1902	Os Sertões	Euclides da Cunha
1920	Cemitério dos Vivos	Lima Barreto
1948	Clara dos Anjos	Lima Barreto

tivéssemos acesso à matriz completa $\tilde{X}_{t,l}$, isto é, se todos os possíveis textos em cada língua possível pudessem ser analisados, poderíamos simplesmente comparar um novo texto t com a variação total da medida $\tilde{X}_{t,l}$ para, por exemplo, verificar para cada língua $l = \lambda$ se um texto dado texto $t = \tau$ é compatível com aquela língua. Na prática, podemos completar apenas algumas linhas e colunas e portanto alguns testes estatísticos são necessários a fim de caracterizar a variação de medidas específicas. Para textos distintos, $P(\tilde{X}_{t,l=\lambda})$ representa a distribuição da medida $\tilde{X}$ entre textos diferentes em uma dada língua $l = \lambda$ e $P(\tilde{X}_{t=\tau,l})$ a distribuição de $\tilde{X}$ para um texto fixo $t = \tau$ escrito em várias línguas. Dessa forma, $\mu(P)$ and $\sigma(P)$ representam a expectância e a variação da distribuição P . Como exemplo, a Figura 5.1 ilustra a distribuição de $\tilde{X} = b$ (número de bigramas duplicados) para três conjuntos de textos que utilizamos em nossa análise: 15 livros em Português (Tabela 5.2), 15 livros em Inglês (Tabela 5.1) e 15 versões do Novo Testamento escrito em 15 diferentes línguas. Consideramos também a média $\tilde{X}_m$ e o desvio padrão $\Delta\tilde{X}$ calculado sobre diferentes livros (por exemplo, cada um dos três conjuntos de 15 livros) e a correlação $\text{Corr}(\tilde{X}, M)$ entre $\tilde{X}$ e o tamanho do vocabulário M do livro. A Tabela 5.3 ilustra os valores de $\tilde{X}_m$, $\Delta\tilde{X}$ e $\text{Corr}(\tilde{X}, M)$ de todas as medidas em cada um dos três conjuntos de livros. A fim de melhor entender a dependência destas medidas com a linguagem (sintaxe) e o texto (semântica), a seguir realizamos alguns testes estatísticos para identificar as medidas que são mais apropriadas para abordar problemas específicos.

Tabela 5.3 – Valores normalizados de métricas para o conjunto de textos do Novo Testamento ($\tau = \text{nt}$) e romances do Português ($\lambda = \text{pt}$) e Inglês ($\lambda = \text{en}$). A correlação das medidas com o tamanho do vocabulário M é mostrada na última coluna.

$\tilde{X}$	$\tilde{X}_m \pm \Delta\tilde{X}$			$\text{Corr}(\tilde{X}, M)$
	$\tau = \text{nt}$	$\lambda = \text{en}$	$\lambda = \text{pt}$	
Vocabulário M	$5\,809 \pm 2\,665$	$4\,720 \pm 922$	$6\,921 \pm 1\,126$	+1,00
Expoente de Zipf γ_N	$1,99 \pm 0,11$	$1,93 \pm 0,06$	$2,01 \pm 0,09$	+0,86
Assortatividade r	$0,91 \pm 0,10$	$1,10 \pm 0,06$	$1,15 \pm 0,04$	+0,07
Diâmetro d	$1,44 \pm 0,58$	$1,32 \pm 0,38$	$1,07 \pm 0,14$	+0,08
Caminhos mínimos $\langle L \rangle$	$1,04 \pm 0,05$	$0,99 \pm 0,02$	$0,97 \pm 0,01$	+0,20
Caminhos mínimos $\langle L \rangle_2$	$1,08 \pm 0,04$	$1,04 \pm 0,02$	$1,03 \pm 0,01$	+0,34
Aglomerção $\langle C \rangle$	$0,83 \pm 0,13$	$0,97 \pm 0,04$	$0,97 \pm 0,03$	-0,34
Aglomerção $\langle C \rangle_2$	$0,66 \pm 0,13$	$0,65 \pm 0,08$	$0,63 \pm 0,07$	-0,58
Intermitência $\langle I \rangle$	$1,30 \pm 0,07$	$1,29 \pm 0,14$	$1,27 \pm 0,06$	-0,43
Intermitência $\langle I \rangle_2$	$1,32 \pm 0,05$	$1,32 \pm 0,14$	$1,26 \pm 0,09$	-0,26
Bigramas b	$0,18 \pm 0,15$	$0,05 \pm 0,04$	$0,10 \pm 0,05$	+0,27
Grau $\langle k \rangle$	$0,71 \pm 0,06$	$0,82 \pm 0,03$	$0,87 \pm 0,02$	+0,53
Grau $\langle k \rangle_2$	$0,71 \pm 0,07$	$0,89 \pm 0,05$	$1,00 \pm 0,04$	+0,26
Exp. da seletividade γ_S	$0,43 \pm 0,14$	$0,51 \pm 0,06$	$0,47 \pm 0,07$	-0,49
Seletividade $\langle S \rangle$	$1,32 \pm 0,18$	$1,13 \pm 0,03$	$1,07 \pm 0,02$	-0,51
Seletividade $\langle S \rangle_2$	$2,09 \pm 0,84$	$1,47 \pm 0,08$	$1,33 \pm 0,10$	-0,39
Motivo de rede m_A	$0,09 \pm 0,04$	$0,12 \pm 0,04$	$0,17 \pm 0,04$	+0,02
Motivo de rede m_B	$1,11 \pm 0,37$	$1,54 \pm 0,11$	$1,72 \pm 0,07$	-0,09
Motivo de rede m_C	$0,83 \pm 0,21$	$1,19 \pm 0,10$	$1,28 \pm 0,05$	+0,04
Motivo de rede m_D	$0,22 \pm 0,09$	$0,27 \pm 0,11$	$0,37 \pm 0,06$	+0,24
Motivo de rede m_E	$0,76 \pm 0,18$	$1,27 \pm 0,16$	$1,03 \pm 0,06$	-0,23
Motivo de rede m_F	$0,24 \pm 0,07$	$0,37 \pm 0,05$	$0,39 \pm 0,06$	-0,20
Motivo de rede m_G	$0,36 \pm 0,14$	$0,47 \pm 0,09$	$0,56 \pm 0,05$	+0,14
Motivo de rede m_H	$0,71 \pm 0,24$	$1,25 \pm 0,11$	$1,16 \pm 0,11$	-0,17
Motivo de rede m_I	$0,20 \pm 0,07$	$0,32 \pm 0,05$	$0,36 \pm 0,05$	-0,14
Motivo de rede m_J	$0,45 \pm 0,17$	$0,57 \pm 0,12$	$0,73 \pm 0,05$	+0,11
Motivo de rede m_K	$0,59 \pm 0,25$	$1,22 \pm 0,16$	$1,02 \pm 0,08$	-0,24
Motivo de rede m_L	$0,03 \pm 0,02$	$0,04 \pm 0,02$	$0,06 \pm 0,02$	+0,10
Motivo de rede m_M	$0,26 \pm 0,10$	$0,39 \pm 0,06$	$0,46 \pm 0,08$	-0,14

Figura 5.1 – Distribuição de $\tilde{X} = b$ para o Novo Testamento (círculos em preto), Inglês (círculos vermelhos) e Português (círculos azuis). A média $\tilde{X}_m$ para os três conjuntos de textos está representada por linhas pontilhadas. Note que nos três conjuntos de texto $\tilde{X} < 1$ e portanto a métrica considerada possui valores diferentes daqueles em textos aleatórios.

5.2 Distinguindo livros de sequências aleatórias

O primeiro objetivo aqui é identificar as medidas capazes de distinguir textos naturais de textos aleatorizados. Estas medidas serão denominadas como medidas informativas. Por exemplo, para $\tilde{X} = b$ na Figura 5.1, todos os valores são muito menores que 1 em todos os três conjuntos de textos, indicando que sua medida assume valores menores em textos naturais do que em textos aleatórios. A fim de quantificar a distância de um conjunto de valores $\{\tilde{X}\}$ para $\tilde{X} = 1$ definimos a quantidade $\Pi(\tilde{X} = 1, \{\tilde{X}\})$ como a proporção de elementos no conjunto $\{\tilde{X}\}$ para o qual $\tilde{X} = 1$ está incluído no intervalo $\tilde{X} \pm \epsilon(\tilde{X})$, onde $\epsilon(\tilde{X})$ aparece das flutuações devido à aleatoriedade presente no processo de embaralhamento conforme definido na equação 5.2. Isto leva à proposição da condição ζ_1 :

- **Condição ζ_1 :** $\tilde{X}$ é considerado informativo se $\Pi(\tilde{X} = 1, \{\tilde{X}\}) \rightarrow 0$ quando $|\{X\}| \rightarrow \infty$, onde $\{\tilde{X}\}$ representa um conjunto de valores de $\tilde{X}$ obtido para diferentes línguas ou textos, e $|\{\tilde{X}\}|$ é o número de elementos neste conjunto.

Os resultados obtidos a partir da aplicação de ζ_1 (com $\Pi(\tilde{X} = 1, \{\tilde{X}\})$) para todos os três conjuntos de textos no nosso banco de dados estão ilustrados na Tabela 5.4. As medidas que satisfazem ζ_1 estão indicadas com um círculo (●) na Tabela 5.4. Diversas medidas de rede (d , $\langle L \rangle$, $\langle C \rangle$, e $\langle k \rangle$) e os motivos m_C , m_E e m_K não satisfazem completamente ζ_1 . Consequentemente elas não podem ser usadas para distinguir um manuscrito da sua versão aleatorizada. Este fato é de certa forma surpreendente já que algumas das medidas anteriores foram úteis para captar sutilezas em textos, por exemplo na tarefa de

reconhecimento de autoria (ver Seção 6). Nesta última aplicação, no entanto, as redes representando textos eram desprovidas de *stopwords* e além disso as palavras remanescentes eram lematizadas (ver Seção 3). A média sobre as 50 palavras mais frequentes parece ser fundamental para satisfazer a condição ζ_1 para a medida de coeficiente de aglomeração e para os caminhos mínimos (note que $\langle C \rangle_2$ and $\langle L \rangle_2$ são informativos enquanto $\langle C \rangle$ e $\langle L \rangle$ não são). Isto significa que a informatividade destas quantidades está concentrada nas palavras mais frequentes. Por outro lado, para a medida de grau, um efeito oposto ocorre, isto é, $\langle k \rangle$ é informativo e $\langle k \rangle_2$ não é. A informatividade da medida de intermitência ($\langle I \rangle$ and $\langle I \rangle_2$) pode ser explicada pelo fato de que por construção a intermitência de palavras distribuídas aleatoriamente no texto possuem $I_i = 1$ (ver Seção 3.8). Uma vez que em textos escritos em linguagem natural muitas palavras tendem a se aglomerar em determinadas regiões (49) $\langle I \rangle > 1$ e $\langle I \rangle_2 > 1$. A seletividade S também é fortemente afetada pelo processo de embaralhamento. Palavras de textos aleatorizados tendem a ser menos seletivas, gerando um aumento em γ_S (isto é, palavras fortemente seletivas ocorrem esporadicamente) e uma diminuição em $\langle S \rangle$ e $\langle S \rangle_2$. A medida de seletividade está relacionada com o efeito de consistência de palavras^a (ver Ref. (72)), que foi verificado ser bastante comum principalmente nas palavras mais frequentes do Inglês. O número de bigramas b também é informativo pois $b \ll 1$ nos três conjuntos de textos, sugerindo que em línguas naturais o aparecimento da mesma palavra seguida no texto é improvável (quando comparado com textos aleatórios). Com relação aos motivos informativos, m_A , m_D , m_F , m_G , m_I , m_J , m_L e m_M raramente ocorrem em línguas naturais ($\langle X \rangle < 1$), enquanto o motivo m_B foi a única medida assumindo valores acima e abaixo de 1. A emergência deste motivo, portanto, parece depender da sintaxe, sendo bastante raro para Xhosa, Vietnamita, Suaíli, Coreano, Hebraico e Árabe.

5.3 Dependência com o estilo e com a língua

Estamos interessados agora em investigar quais medidas textuais são mais dependentes da linguagem do que do estilo do livro, e vice-versa. Medidas dependendo predominantemente da sintaxe são esperadas possuir maior variabilidade entre línguas diferentes do que entre textos. Por outro lado, medidas dependendo principalmente da estória (semântica) sendo contada são esperadas possuir maior variabilidade entre textos na mesma língua^b, isto é $t = \tau$. A variabilidade das medidas foi calculada com o coeficiente de variação $v = \Delta\tilde{X}/\tilde{X}_m$, onde $\Delta\tilde{X}$ e $\tilde{X}_m$ representam respectivamente o desvio padrão e a média cal-

^a Palavras consistentes são palavras com alta tendência de co-ocorrer sempre com as mesmas palavras vizinhas. Por este motivo, tais palavras são geralmente bastante seletivas.

^b Esta abordagem poderia ser estendida para considerar diferentes gêneros, uma vez que características distintas poderiam aparecer em romances, letras de músicas, enciclopédias, textos científicos, etc.

Tabela 5.4 – Verificação de que medidas satisfazem as condições ζ_1 , ζ_2 , ζ_3 e ζ_4 . $\text{Corr}(\tilde{X}, M)$ representa o coeficiente de correlação de Pearson entre $\tilde{X}$ e o vocabulário M . Assumimos que ζ_1 , ζ_2 , ζ_3 and ζ_4 são satisfeitas respectivamente quando $\Pi = 0, 00 \%$, $v_{t=\text{nt}, l} > v_{t=\text{nt}, l} \cap \iota(v_{t, l=\lambda}) \leq 0, 05$, $\iota(v_{t, l=\lambda}) \cup \iota(v_{t, l=\lambda})$ e $c(X_{t=\text{nt}, l=\lambda}, P(X_{t, l=\lambda})) > 0, 05$, onde nt representa o conjunto de textos do Novo Testamento. Medidas satisfazendo as condições para os três conjuntos de textos estão marcados com um círculo (●).

$\tilde{X}$	$\Pi(X = 1, \{X\})$			$v_{t=\text{nt}, l} / v_{t, l=\lambda}$		$c(X, P(X))$		$\text{Corr}(\tilde{X}, M)$	ζ_1	ζ_2	ζ_3	ζ_4
	$\tau = \text{nt}$	$\lambda = \text{en}$	$\lambda = \text{pt}$	$\lambda = \text{en}$	$\lambda = \text{pt}$	$\lambda = \text{en}$	$\lambda = \text{pt}$					
Vocabulário M	–	–	–	3,12	2,82	0,00	0,00	+1,00	–	●	●	
Exponente de Zipf γ_N	–	–	–	1,71	1,25	0,00	0,00	+0,86	–	●	●	
Assortatividade r	0,00 %	0,00 %	0,00 %	2,18	3,41	0,07	0,14	+0,07	●	●	●	●
Diâmetro d	12,50 %	37,50 %	43,75 %	1,41	3,16	0,00	0,00	+0,08		●		
Caminhos mínimos $\langle L \rangle$	12,50 %	0,00 %	0,00 %	2,07	7,57	0,76	0,68	+0,20		●	●	●
Caminhos mínimos $\langle L \rangle_2$	0,00 %	0,00 %	0,00 %	2,23	2,91	0,80	0,51	+0,34	●	●	●	●
Aglomerção $\langle C \rangle$	0,00 %	18,75 %	25,00 %	3,31	4,74	0,65	0,62	-0,34		●	●	●
Aglomerção $\langle C \rangle_2$	0,00 %	0,00 %	0,00 %	1,52	1,71	0,91	0,80	-0,58	●	●		●
Intermitência $\langle I \rangle_2$	0,00 %	0,00 %	0,00 %	0,47	1,03	0,59	0,45	-0,43	●		●	●
Intermitência $\langle I \rangle_2$	0,00 %	0,00 %	0,00 %	0,36	0,75	0,77	0,95	-0,26	●		●	●
Bigramas b	0,00 %	0,00 %	0,00 %	1,01	11,4	0,95	0,32	+0,27	●	●		●
Grau $\langle k \rangle$	0,00 %	0,00 %	0,00 %	1,44	3,99	0,00	0,01	+0,53	●	●	●	
Grau $\langle k \rangle_2$	0,00 %	0,00 %	12,50 %	1,93	2,81	0,01	0,01	+0,26		●	●	
Exp. da seletividade γ_S	0,00 %	0,00 %	0,00 %	2,53	2,26	0,88	0,69	-0,49	●	●	●	●
Seletividade $\langle S \rangle$	0,00 %	0,00 %	0,00 %	5,06	8,30	0,05	0,25	-0,51	●	●	●	
Seletividade $\langle S \rangle_2$	0,00 %	0,00 %	0,00 %	7,18	5,60	0,48	0,62	-0,39	●	●	●	●
Motivo de rede m_A	0,00 %	0,00 %	0,00 %	1,31	1,85	0,00	0,00	+0,02	●	●		
Motivo de rede m_B	0,00 %	0,00 %	0,00 %	3,75	7,67	0,00	0,00	-0,09	●	●	●	
Motivo de rede m_C	18,75 %	0,00 %	0,00 %	2,30	6,04	0,00	0,00	+0,04		●	●	
Motivo de rede m_D	0,00 %	0,00 %	0,00 %	0,97	2,45	0,00	0,00	+0,24	●	●		
Motivo de rede m_E	12,50 %	6,25 %	18,75 %	1,66	0,72	0,00	0,00	-0,23		●		
Motivo de rede m_F	0,00 %	0,00 %	0,00 %	1,87	1,80	0,00	0,00	-0,20	●	●		
Motivo de rede m_G	0,00 %	0,00 %	0,00 %	1,82	4,43	0,00	0,00	+0,14	●			
Motivo de rede m_H	0,00 %	0,00 %	0,00 %	2,67	3,66	0,00	0,00	-0,17	●	●	●	
Motivo de rede m_I	0,00 %	0,00 %	0,00 %	1,68	2,48	0,00	0,00	-0,14	●	●		
Motivo de rede m_J	0,00 %	0,00 %	0,00 %	1,76	5,19	0,00	0,00	+0,11	●	●		
Motivo de rede m_K	0,00 %	12,50 %	18,75 %	2,55	5,29	0,00	0,00	-0,24		●	●	
Motivo de rede m_L	0,00 %	0,00 %	0,00 %	1,53	1,85	0,04	0,35	+0,10	●	●		
Motivo de rede m_M	0,00 %	0,00 %	0,00 %	2,11	2,16	0,00	0,00	-0,14	●	●	●	

culada para os livros no conjunto $\{\tilde{X}\}$. Deste modo, assumimos que $\tilde{X}$ é mais dependente da língua do que do estilo se a condição ζ_2 é satisfeita:

- **Condição ζ_2 :** $\tilde{X}$ é mais dependente da língua (ou da sintaxe) do que do estilo (ou da semântica) se $v_{t=\tau,l} > v_{t,l=\lambda}$.

As medidas que não satisfazem a condição ζ_2 satisfazem a desigualdade $v_{t,l=\lambda} > v_{t=\tau,l}$ e portanto são mais dependentes do estilo/semântica do que da língua/sintaxe. A fim de quantificar se a diferença $v_{t=\tau,l} - v_{t,l=\lambda}$ é significativa, consideramos os intervalos de confiança de $v_{t=\tau,l}$ e $v_{t,l=\lambda}$. Seja $\iota(v)$ o intervalo de confiança para v calculado a partir da distribuição-t não central (73), então ζ_2 é válida se existe uma pequena intersecção dos intervalos de confiança. Em outras palavras:

- **Condição ζ_3 :** A inequação $v_{t=\tau,l} > v_{t,l=\lambda}$ (ou $v_{t,l=\lambda} > v_{t=\tau,l}$) é válida somente se $\iota(v_{t=\tau,l}) \cap \iota(v_{t,l=\lambda}) \rightarrow 0$ quando $|\{X\}| \rightarrow \infty$.

Os intervalos de confiança foram assumidos possuir um baixo grau de intersecção se

$$\frac{\iota(v_{t=\tau,l}) \cap \iota(v_{t,l=\lambda})}{\iota(v_{t=\tau,l}) \cup \iota(v_{t,l=\lambda})} \leq 0,05.$$

No cálculo dos intervalos de confiança, consideramos um nível de significância $\alpha = 0,95$.

Os resultados para as medidas satisfazendo as condições ζ_2 e ζ_3 estão ilustrados na Tabela 5.4. As medidas satisfazendo as condições ζ_2 e ζ_3 servem para examinar a dependência na sintaxe ou no estilo/semântica. O tamanho do vocabulário M , e as medidas de rede r , $\langle L \rangle$, $\langle L \rangle_2$, $\langle C \rangle$, $\langle k \rangle$ e $\langle k \rangle_2$ são mais dependentes da sintaxe do que da semântica. As medidas derivadas do conceito de seletividade (γ_S , $\langle S \rangle$ e $\langle S \rangle_2$) também são fortemente dependentes da linguagem. Com relação aos motivos, cinco deles satisfazem completamente ζ_2 e ζ_3 : m_B , m_C , m_H , m_K and m_M . Interessantemente, $\langle I \rangle$ e $\langle I \rangle_2$ são as únicas medidas com baixos valores de $v_{t=\tau,l}/v_{t,l=\lambda}$. Reciprocamente, a única medida que significativamente violou ζ_2 (e satisfaz ζ_3) foi $\langle I \rangle_2$, sugerindo que a intermitência média das palavras mais frequentes é mais dependente da semântica do que da sintaxe da língua.

5.4 Representatividade das medidas

A implementação prática do arcabouço geral de análise de medidas foi realizado a partir da quantificação da variação observada entre línguas usando apenas um livro (o Novo Testamento). Este livro foi escolhido devido à escassez de livros suficientemente longos disponíveis em várias línguas. Para garantir a robustez do arcabouço de análise é necessário determinar se as flutuações entre diferentes línguas são representativas das

flutuações observadas em variados livros. Por este motivo, investigamos as medidas $\tilde{X}$ cujos valores obtidos para um dado livro em uma dada língua λ ($\tilde{X}_{t=\tau, l=\lambda}$) são compatíveis com outros livros na mesma língua ($X_{t, l=\lambda}$). Para isto, definimos a compatibilidade de uma observação $\tilde{X}_{t=\tau, l=\lambda}$ como $c(\tilde{X}_{t=\tau, l=\lambda}, P(X_{t, l=\lambda}))$. A distribuição P foi calculada através da interpolação de Parzen (74) usando uma função gaussiana $\mathcal{G}$ como kernel:

$$\mathcal{G}(z) = \frac{1}{\sigma_{\mathcal{G}}\sqrt{2\pi}} \exp\left(-\frac{z - \mu_{\mathcal{G}}}{2\sigma_{\mathcal{G}}^2}\right). \quad (5.3)$$

Mais precisamente, a distribuição P foi construída adicionando distribuições $\mathcal{G}$ ao redor de cada $\tilde{X}$ observado em diferentes textos em uma dada língua fixa λ . Matematicamente, $c(\tilde{X}^* = \tilde{X}_{t=\tau, l=\lambda}, P)$ é calculado como

$$c(\tilde{X}^*, P) = 2 \times \int_{-\infty}^{\tilde{X}^*} P(X) dX, \quad (5.4)$$

caso $\tilde{X}^*$ se encontre antes da mediana da distribuição P . Caso contrário, $c(\tilde{X}^*, P)$ é calculado como

$$c(\tilde{X}^*, P) = 2 \times \int_{\tilde{X}^*}^{+\infty} P(X) dX. \quad (5.5)$$

Para fins práticos, consideramos que $\tilde{X}_{t=\text{nt}, l=\lambda}$ é compatível com outros livros escritos na mesma língua λ se ζ_4 é satisfeita:

- **Condição ζ_4 :** $\tilde{X}_{t=\text{nt}, l=\lambda}$ é uma medida representativa da língua λ se a compatibilidade entre $\tilde{X}_{t=\text{nt}, l=\lambda}$ e $P(\tilde{X}_{t, l=\lambda})$ for maior que 5%.

Para exemplificar este conceito, considere a Figura 5.2. Um conjunto de observações $\{\tilde{X}\}$ em textos escritos na língua $l = \lambda$ é representado por círculos azuis. A respectiva distribuição $P(\tilde{X}_{t, l=\lambda})$ obtida da estimação de Parzen é ilustrada em vermelho. O valor fictício $\tilde{X}_{t=\text{nt}, l=\lambda}$ obtido para a medida no Novo Testamento é ilustrado pelas linhas tracejadas **A** e **B**. Enquanto no primeiro caso a medida $\tilde{X}_{t=\text{nt}, l=\lambda}$ é compatível com uma medida em língua natural, no segundo caso a medida não pode ser considerada compatível pois é um *outlier* caracterizado por $c < 0,05$.

A representatividade das medidas calculadas para o Novo Testamento foi verificada usando a distribuição $P(\tilde{X})$ obtida do conjunto de livros escritos em Português e Inglês. O desvio padrão aplicado no método de Parzen foi o pior desvio entre Inglês e Português, isto é, $\sigma_{\mathcal{G}} = \min\{\sigma_{\text{pt}}, \sigma_{\text{en}}\}$. As medidas satisfazendo ζ_4 para o Inglês e para o Português estão apontadas na última coluna da Tabela 5.4. Com relação às medidas de rede, apenas $\langle L \rangle$, $\langle L \rangle_2$, $\langle C \rangle$ e $\langle C \rangle_2$ são representativas, sugerindo que elas são fracamente dependentes da variação de estilo. Além disso, $\langle I \rangle$, $\langle I \rangle_2$, B , γ_S , $\langle S \rangle_2$ e m_L mostraram ser medidas representativas.

Figura 5.2 – Exemplo de verificação da representatividade das medidas obtida para o Novo Testamento (linhas tracejadas) com medidas obtidas de línguas naturais (círculos azuis). Note que, diferentemente do caso **A** (ver equação 5.4), **B** é considerado um *outlier* porque $c(\tilde{X}_{t=nt,l=\lambda}, P(\tilde{X}_{t,l=\lambda})) < 0,05$ (ver equação 5.5)

5.5 Detecção de palavras-chave

Um problema fundamental nas ciências da informação é a detecção de palavras importantes uma vez que elas oferecem dicas sobre o conteúdo do texto. A identificação de palavras-chave também pode ser útil para guiar o processo de decodificação de manuscritos codificados, uma vez que os especialistas em criptografia podem focar sua atenção nas palavras mais relevantes. Técnicas tradicionais como a *term frequency-inverse document frequency* (tf-idf) são baseadas na análise de frequência (39). Basicamente, esta técnica associa alta relevância a uma palavra se ela é bastante frequente no documento em análise mas pouco frequente em outros documentos da coleção. A principal desvantagem associada a esta abordagem é a necessidade de um conjunto representativo de documentos no mesmo idioma. Obviamente, esta restrição torna impossível aplicar esta abordagem para identificar palavras de um documento fora do contexto da coleção. Uma outra possibilidade seria usar métodos baseados em entropia (50). No entanto, a aplicação destes métodos em casos é limitado uma vez que eles exigem que o manuscrito seja organizado em partições naturais, como capítulos e seções.

Para superar este problema, derivamos uma estratégia de detecção de palavras-chave que é baseada no fato de que palavras relevantes apresentam altos valores de intermitência dentro de um dado texto (49). Portanto, esta característica pode exercer o papel tradicionalmente exercido pelo termo denominado *inverse document frequency* (idf). Desta forma, a análise tf-idf pode ser redefinada aqui atribuindo-se a relevância Ω_i para a palavra i de

Tabela 5.5 – Lista de palavras chaves obtidas para o Novo Testamento (Inglês, Português e Alemão) usando a equação 5.6. Note que as palavras chaves encontradas para o Novo Testamento condizem com o contexto do manuscrito.

Língua	Palavras-chave
Português	nasceu, Pilatos, céus, bem-aventurados, Isabel, anjo, menino, vinha, sumo, sepulcro, joio, Maria, portanto, Herodes, talentos
Inglês	<i>begat, Pilates, talents, loaves, Herod, tares, vineyard shall, boat, demons, five, pay, sabbath, hear, whosoever</i>
Alemão	<i>zeugete, zentner, himmelreich, Pilatus, schwert, Maria, Elisabeth Etliches, unkraut, euch, weden, heuchler, tempel</i>

acordo com Ω :

$$\Omega_i = (I_i - 1)\sqrt{\log N_i}, \quad (5.6)$$

onde I_i é definida na equação 3.54 e N_i representa o número absoluto de ocorrências da palavra i . Outras combinações destes dois fatores podem ser utilizadas em aplicações específicas (p.e., para livros com tamanhos distintos um termo proporcional à frequência normalizada poderia ser utilizada no lugar de N_i). Note que com a introdução o fator I_i , palavras com $I \simeq 1$ assumem baixos valores de Ω mesmo quando elas são bastante frequentes. Embora existam outros métodos para detecção de palavras-chave baseados na análise da distribuição irregular das palavras (75) tais métodos são específicos para textos curtos. A fim de validar Ω , a equação 5.6 foi aplicada ao Novo Testamento em Português, Inglês e Alemão. A distribuição das palavras-chave para estes três idiomas com relação aos termos de frequência e intermitência está ilustrada na Figura 5.3. Uma inspeção da Tabela 5.18 revela que palavras representativas foram obtidas nos três idiomas. Dentre estas palavras, podemos citar personagens importantes como *Pilatos, Herodes, Isabel e Maria*; além de conceitos importantes do contexto bíblico como *nasceu, begat* (gerou), *céus, himmelreich* (paraíso), *heuchler* (hipócrita), *demons* (demônios) e *sabbath* (sábado). Note que as palavras-chave obtidas para os três idiomas não são as mesmas, embora o conteúdo analisado seja o mesmo. Isto sugere que a detecção de palavras-chave parece ser dependente do tradutor. De fato, substituições de palavras por sinônimos poderiam facilmente transformar uma palavra-chave em palavra comum. Finalmente, a lista de palavras-chave obtida nos três idiomas está ilustrada na Figura 5.3, em termos do fator de frequência e do fator de intermitência.

5.6 Estudo de caso: o manuscrito Voynich (VMS)

Nas seções anteriores introduzimos um arcabouço para identificar a dependência de diferentes medidas com a língua e com a estória de diferentes livros. Nesta seção

Figura 5.3 – Palavras-chave para o Novo testamento em (a) Português, (b) Inglês, e (c) Alemão e para (d) o manuscrito Voynich. N_i representa o número de ocorrência da palavra i no texto e I_i é a medida de intermitência definida na equação 3.54. As palavras-chave foram obtidas da equação 5.6 e estão marcadas com asterisco (*). As outras palavras estão indicadas por círculos (o). Note que as palavras-chave são caracterizadas por altos valores de I_i e N_i . Nas três línguas analisadas a palavra-chave com maior escore Ω (que corresponde à palavra *begat* em Inglês) pode ser explicada pela sua concentração (alta intermitência I) na descrição da genealogia de Jesus em duas passagens do Novo Testamento (Mateus 1:1-17 e Lucas 3:23-38).

investigamos como as medidas identificadas como relevantes podem fornecer informação para análise de textos. O manuscrito Voynich (VMS), adquirido pelo negociante de livros Wilfrid Voynich no início do século XX, é um livro de 240 páginas que remonta ao século XV. Seu aspecto misterioso tem cativado a atenção do público por séculos. De fato, o VMS tem sido estudado por criptógrafos profissionais, representando um desafio a estudiosos (58, 71). Algumas páginas originais do manuscrito estão ilustradas na Figura 5.4. Atualmente, é classificado entre as seis mais importantes cifras ainda não resolvidas (71). As várias hipóteses sobre o VMS podem ser sumarizadas em três categorias:

- Uma sequência de palavras sem significado.
- Um texto com significado escrito originalmente em uma língua existente que foi codificada (possivelmente com uma cifra mono-alfabética).
- Um texto com significado escrito em uma língua desconhecida (possivelmente construída).

Embora seja impraticável investigar sistematicamente todas estas hipóteses, realizamos algumas análises estatísticas a fim de verificar a viabilidade destes cenários. Para abordar o primeiro ponto, analisamos textos aleatorizados. Para considerar o segundo, consideramos 15 línguas distintas, incluindo o Esperanto que, por ser uma língua artificial, permite considerar também o último ponto. Neste estudo de caso não consideramos o efeito de criptografia poli-alfabética no texto porque toda a análise estatística seria influenciada pelas propriedades de codificação e portanto toda a informação a respeito do “idioma do VMS” seria perdida.

As propriedades estatísticas do VMS foram analisadas com o objetivo de responder as questões da Tabela 5.6 (uma visualização da rede obtida para o manuscrito Voynich é fornecida na Figura 5.5). Para cada uma destas perguntas, verificamos as medidas capazes de gerar resultados estatisticamente significativos. Para verificar se um dado texto é compatível com sua versão aleatória, os valores de $\tilde{X}$ calculados em textos escritos em línguas naturais devem estar sempre distantes de $\tilde{X} = 1$, e portanto apenas medidas informativas são capazes de responder a questão Q_1 . Para testar se um texto é consistente com alguma língua natural (questão Q_2), os textos empregados como base de comparação (isto é, o Novo Testamento) devem ser representativos da língua. Dessa forma, a condição ζ_4 precisa ser satisfeita na escolha de medidas apropriadas para responder a questão Q_2 . As condições ζ_2 e ζ_3 precisam ser satisfeitas por medidas apropriadas para responder Q_3 uma vez que a variância de estilo dentro de uma língua deve ser pequena, caso seja desejável determinar a língua mais similar. Caso contrário, um texto *outlier* em termos de estilo poderia ser equivocadamente considerado como pertencente à uma outra língua. Um raciocínio análogo pode ser aplicado na seleção de medidas úteis para identificar o estilo mais próximo. Finalmente, note que as respostas para as questões Q_3 e Q_4 necessitam de

Figura 5.4 – Visualização de algumas páginas do manuscrito Voynich, uma das mais importantes cifras ainda não resolvidas. As ilustrações detalhadas sugerem que este manuscrito trata de temas como astrologia, rituais de fertilidade e plantas. As páginas desta figura foram reproduzidas do website archive.org/stream/TheVoynichManuscript/Voynich_Manuscript.

uma comparação com livros de referência (por exemplo, o Novo Testamento no caso da Q_3). Por este motivo, medidas apropriadas devem satisfazer a condição ζ_4 para garantir que as medidas calculadas para os livros de referência sejam representativas da língua ou estilo.

Tabela 5.6 – Condições que precisam ser satisfeitas pelas medidas para responder cada uma das questões de interesse em textos. A condição ζ_1 é útil para selecionar as medidas informativas uma vez que esta condição assegura que textos aleatorizados podem ser distinguidos de textos escritos em linguagem natural. Para as questões Q_2 e Q_3 é necessário que a medida do Novo Testamento seja representativa do idioma, já que o Novo Testamento é usado como texto base. No caso da Q_4 , é necessário que os textos utilizados como base sejam representativos do estilo. Para a questão Q_3 , é desejável também que não haja intersecção entre as medidas de livros pertencentes a idiomas diferentes. Por este motivo, ζ_2 e ζ_3 precisam ser satisfeitas. Finalmente, para encontrar o estilo mais próximo na questão Q_4 , é desejável que a medida dependa fortemente do estilo.

Questões		ζ_1	ζ_2	ζ_3	ζ_4
Q_1	O texto é compatível com a versão aleatória?	•			
Q_2	O texto é compatível com línguas naturais?				•
Q_3	Qual é a língua mais similar?		•	•	•
Q_4	Qual é o estilo mais similar?			•	•

Figura 5.5 – Rede obtida para o manuscrito Voynich. Os vértices mais internos são aqueles mais frequentes (em vermelho e verde) e por este motivo fazem parte do núcleo da rede.

5.6.1 O manuscrito Voynich é compatível com textos aleatórios?

Antes da verificação da compatibilidade do manuscrito Voynich com textos aleatórios, verificamos se Q_1 pode ser corretamente respondida em um conjunto de livros escritos em Português ($l = \lambda_{ts} = \text{pt}$) e Inglês ($l = \lambda_{ts} = \text{en}$), denominados como banco de dados de teste (ver Tabelas 5.7 e 5.8). Um dado texto foi considerado como não aleatório se o intervalo $\tilde{X} \pm \epsilon(\tilde{X})$ não inclui $\tilde{X} = 1$. Para quantificar a distância de um texto para sua versão aleatorizada, definimos a distância D :

$$D = \frac{|\tilde{X} - 1|}{\epsilon(\tilde{X})}, \quad (5.7)$$

que quantifica quantos ϵ 's o valor de $\tilde{X}$ é distante de $\tilde{X} = 1$. Como esperado, os valores de $\tilde{X}$ calculados no banco de dados de teste para $\lambda_{ts} = \text{pt}$ e $\lambda_{ts} = \text{en}$ indicam que todos os textos não são compatíveis com suas versões aleatórias porque $D > 1$ (ver Tabelas 5.9, 5.10, 5.11 e 5.12), o que significa que o intervalo $X \pm \epsilon(\tilde{X})$ não inclui $\tilde{X} = 1$. Uma vez que a metodologia classificou apropriadamente os textos no banco de dados de teste como incompatíveis com as versões aleatórias, estamos agora em posição de aplicá-la ao VMS.

Tabela 5.7 – Lista de livros em Inglês empregados para validar a metodologia de verificação de compatibilidade com textos aleatorizados e textos escritos em língua natural. A sigla utilizada para identificar o livro aparece após o seu nome.

Título	Autor
<i>David Copperfield</i> (DCO)	Charles Dickens
<i>Dracula</i> (DRA)	Bram Stoker
<i>Evelina</i> (EVE)	Fanny Burney
<i>Great Expectations</i> (GEP)	Charles Dickens
<i>History of Tom Jones, a Foundling</i> (HTJ)	Henry Fielding
<i>Moby Dick</i> (MDC)	Herman Melville
<i>Persuasion</i> (PER)	Jane Austen
<i>Pride and Prejudice</i> (PPR)	Jane Austen
<i>Robinson Crusoe</i> (RCR)	Daniel Defoe
<i>Ulysses</i> (ULY)	James Joyce

Tabela 5.8 – Lista de livros em Português empregados para validar a metodologia de verificação de compatibilidade com textos aleatorizados e textos escritos em língua natural. A sigla utilizada para identificar o livro aparece após o seu nome.

Título	Autor
A Moreninha (MOR)	Joaquim Manuel de Macedo
Ao Correr da Pena (CPE)	José de Alencar
Memórias Póstumas de Brás Cubas (BRC)	Machado de Assis
Contrastes e confrontos (CON)	Euclides da Cunha
Esaú e Jacó (EJC)	Machado de Assis
Grande Sertão: Veredas (GSV)	João Guimarães Rosa
Memórias de um Sargento de Milícias (MSM)	Manuel Antônio de Almeida
O Guarani (GUA)	José de Alencar
Triste Fim de Policarpo Quaresma (POL)	Lima Barreto
O Primo Basílio (PBA)	Eça de Queirós

Tabela 5.9 – Distância D (equação 5.7) entre textos escritos em Inglês e suas versões aleatorizadas (parte 1). Note que uma vez que $D > 1$ para todos os textos e para todas as medidas consideradas, podemos pressupor que os textos diferem significativamente de suas versões aleatorizadas.

Livro	$\langle L \rangle_2$	$\langle C \rangle_2$	$\langle I \rangle$	$\langle I \rangle_2$	B	$\langle k \rangle$	γ_S
DCO	22,1	81,7	34,8	21,1	790,5	195,9	55,6
DRA	25,1	62,0	26,6	23,2	3234,7	185,7	45,0
EVE	31,6	90,6	35,9	13,0	946,5	243,2	57,5
GEP	31,2	41,6	22,8	41,9	1273,4	234,6	117,3
HTJ	31,9	90,3	17,2	14,4	36402,3	168,0	61,7
MDC	24,2	34,2	18,7	19,9	660,2	96,8	48,6
PER	34,4	87,5	24,5	26,1	7138,6	157,1	40,2
PPR	43,6	52,4	13,1	19,4	5604,7	238,8	24,8
RCR	50,4	80,3	21,9	20,1	88327,9	355,9	38,5
ULY	30,06	24,3	33,5	31,1	197,5	112,3	76,8

Tabela 5.10 – Distância D (equação 5.7) entre textos escritos em Inglês e suas versões aleatorizadas (parte 2). Note que uma vez que $D > 1$ para todos os textos e para todas as medidas consideradas, podemos pressupor que os textos diferem significativamente de sua versões aleatorizadas.

Livro	m_G	m_F	m_J	m_D	m_I	m_M	m_A	m_L
DCO	95,5	158,3	59,2	107,7	186,1	59,6	549,1	1030,8
DRA	79,5	71,2	66,2	86,9	146,2	86,6	483,2	1237,3
EVE	117,2	92,2	82,5	105,9	247,5	73,5	546,1	815,3
GEP	147,9	87,2	81,0	69,6	242,0	135,3	569,0	877,2
HTJ	132,2	83,9	57,4	93,9	147,3	58,6	561,8	483,4
MDC	125,9	91,9	78,3	75,9	138,0	63,8	540,9	929,3
PER	77,8	80,4	42,7	85,0	205,9	71,9	347,8	907,3
PPR	235,4	136,4	83,7	112,2	284,2	105,5	543,9	1376,7
RCR	166,6	103,4	81,0	136,6	251,2	51,7	794,9	1786,0
ULY	56,6	137,0	73,0	102,8	159,9	81,4	385,6	998,4

Tabela 5.11 – Distância D (equação 5.7) entre textos escritos em Português e suas versões aleatorizadas (parte 1). Note que uma vez que $D > 1$ para todos os textos e para todas as medidas consideradas, podemos pressupor que os textos diferem significativamente de sua versões aleatorizadas.

Livro	$\langle L \rangle_2$	$\langle C \rangle_2$	$\langle I \rangle$	$\langle I \rangle_2$	B	$\langle k \rangle$	γ_S
MOR	48,2	39,9	29,6	35,4	59,2	93,7	26,9
CPE	43,1	41,1	32,5	14,7	969,1	109,5	21,8
BRC	35,6	38,8	31,3	19,7	137,1	84,9	19,4
CON	47,8	55,5	14,0	13,4	563,7	98,4	14,3
EJC	41,7	49,3	22,2	35,3	79,4	99,1	12,8
GSV	24,3	14,4	30,7	20,7	46,1	75,5	19,4
MSM	35,7	33,2	19,2	29,5	272,6	116,9	41,9
GUA	46,4	74,5	27,7	26,5	255,6	90,6	27,8
POL	37,4	51,8	19,3	14,0	164,7	115,1	26,1
PBA	45,8	31,5	20,3	18,4	150,1	91,2	30,4

Tabela 5.12 – Distância D (equação 5.7) entre textos escritos em Português e suas versões aleatorizadas (parte 2). Note que uma vez que $D > 1$ para todos os textos e para todas as medidas consideradas, podemos pressupor que os textos diferem significativamente de sua versões aleatorizadas.

Livro	m_G	m_F	m_J	m_D	m_I	m_M	m_A	m_L
MOR	73,9	72,6	33,1	76,3	141,5	40,9	329,7	430,5
CPE	67,4	103,6	26,1	54,5	143,8	60,7	422,6	784,9
BRC	50,0	96,4	27,0	55,8	263,6	49,3	262,4	349,0
CON	37,7	100,0	18,3	23,3	181,2	97,0	410,3	980,7
EJC	53,8	81,3	18,6	47,5	149,0	48,3	191,3	286,9
GSV	55,3	55,6	17,9	30,2	115,8	30,7	201,1	415,2
MSM	74,7	93,5	41,7	62,1	124,9	47,2	584,6	871,7
GUA	58,1	172,6	28,9	67,6	244,0	87,7	350,6	1111,8
POL	44,3	82,1	32,3	55,2	112,9	83,7	239,2	398,3
PBA	50,4	99,3	45,1	60,3	130,3	90,3	302,4	400,2

Tabela 5.13 – Valores de $\tilde{X}$ para o manuscrito Voynich considerando apenas medidas informativas, isto é, medidas que satisfazem a condição ζ_1 . Com exceção de $\langle C \rangle_2$, todas as medidas indicam que o manuscrito Voynich não é compatível com textos de natureza aleatória.

$\tilde{X}$	$\langle L \rangle$	$\langle C \rangle_2$	$\langle I \rangle$	$\langle I \rangle_2$	B	$\langle k \rangle$	γ_S
$\tilde{X}_{\text{VMS}} - \epsilon(X_{\text{VMS}})$	1,069	0,981	1,423	1,875	2,333	0,948	0,617
$\tilde{X}_{\text{VMS}}$	1,071	0,999	1,433	1,890	2,637	0,949	0,692
$\tilde{X}_{\text{VMS}} + \epsilon(X_{\text{VMS}})$	1,072	1,017	1,443	1,904	2,940	0,950	0,768
D_{VMS}	47	0	44	61	5	51	23

Tabela 5.14 – Valores de $\tilde{X}$ para o manuscrito Voynich considerando apenas os motivos informativos, isto é, os motivos que satisfazem a condição ζ_1 . Note que todos os motivos sugerem que o manuscrito Voynich não é compatível com textos de natureza aleatória.

$\tilde{X}$	m_G	m_F	m_J	m_D	m_I	m_M	m_A	m_L
$\tilde{X}_{\text{VMS}} - \epsilon(X_{\text{VMS}})$	0,782	0,738	0,784	0,908	0,724	0,783	0,728	0,549
$\tilde{X}_{\text{VMS}}$	0,796	0,751	0,798	0,940	0,733	0,801	0,739	0,582
$\tilde{X}_{\text{VMS}} + \epsilon(X_{\text{VMS}})$	0,809	0,765	0,813	0,971	0,741	0,819	0,751	0,616
D_{VMS}	15	18	14	2	32	11	23	12

Os valores de $\tilde{X}$ para o VMS, denotados como $\tilde{X}_{\text{VMS}}$ nas tabelas 5.13 e 5.14, indicam que o VMS não é compatível com textos aleatórios, uma vez que o intervalo $\tilde{X}_{\text{VMS}} \pm \epsilon(\tilde{X}_{\text{VMS}})$ não inclui $\tilde{X} = 1$. Com exceção do coeficiente de aglomeração das palavras mais frequentes ($\langle C \rangle_2$), todas as medidas não incluem $\tilde{X} = 1$ no intervalo $\tilde{X}_{\text{VMS}} \pm \epsilon(\tilde{X}_{\text{VMS}})$, o que sugere que a ordem das palavras no VMS não é estabelecida aleatoriamente. As medidas que mais diferenciam o manuscrito Voynich dos textos aleatórios foi determinada quantitativamente usando a medida D_{VMS} , a partir da equação 5.7. As Tabelas 5.13 e 5.14 mostram que os maiores valores de D_{VMS} ocorrem para a medida de intermitência ($\langle I \rangle$ and $\langle I \rangle_2$) e para as medidas de rede ($\langle k \rangle$ and $\langle L \rangle_2$). Uma vez que as medidas de rede são fortemente afetadas por fatores sintáticos, os resultados obtidos sugerem que o VMS não é compatível com textos aleatórios sem significado na esfera sintática.

5.6.2 O manuscrito Voynich é compatível com textos escritos em línguas naturais?

A compatibilidade do manuscrito Voynich com textos escritos em línguas naturais foi verificado a partir da comparação das medidas apropriadas computadas para o VMS e as medidas do Novo Testamento escrito em 15 línguas. Analogamente à análise de compatibilidade com textos aleatórios realizada na Seção 5.6.1, validamos nossa estratégia no banco de dados de teste (Tabelas 5.7 e 5.8) da seguinte maneira. A compatibilidade com textos escritos em linguagem natural foi computada a partir das equações 5.4 e 5.5, onde P foi obtido do conjunto de textos do Novo Testamento ($t = \tau = \text{nt}$). O desvio

Tabela 5.15 – Compatibilidade dos textos em Inglês (banco de dados de teste) com textos escritos em linguagem natural (Novo Testamento). Dado que os textos são considerados incompatíveis com línguas naturais quando $c < 0,05$, conclui-se que, como esperado, todos os textos do banco de dados são compatíveis com línguas naturais.

Livro	r	$\langle L \rangle$	$\langle L \rangle_2$	$\langle C \rangle$	$\langle C \rangle_2$	$\langle I \rangle$	$\langle I \rangle_2$	B	$\langle S \rangle_2$	γ_S
DCO	0,20	0,91	0,74	0,69	0,96	0,91	0,60	0,94	0,55	0,74
DRA	0,19	0,99	0,81	0,31	0,93	0,49	0,96	0,60	0,29	0,55
EVE	0,12	0,98	0,45	0,69	0,83	0,98	0,87	0,98	0,68	0,63
GEP	0,33	0,97	0,77	0,79	0,77	0,87	0,39	0,86	0,23	0,89
HTJ	0,13	0,83	0,33	0,36	0,65	0,66	0,99	0,28	0,27	0,68
MDC	0,18	0,96	0,99	0,60	0,70	0,73	0,94	0,95	0,11	0,68
PER	0,25	0,85	0,61	0,43	0,99	0,78	0,94	0,53	0,47	0,48
PPR	0,15	0,91	0,42	0,60	0,96	0,39	0,74	0,46	0,65	0,48
RCR	0,19	0,95	0,55	0,38	0,84	0,41	0,85	0,23	0,47	0,39
ULY	0,21	0,58	0,72	0,73	0,54	0,81	0,43	0,65	0,15	0,98

Tabela 5.16 – Compatibilidade dos textos em Português (banco de dados de teste) com textos escritos em linguagem natural (Novo Testamento). Dado que os textos são considerados incompatíveis com línguas naturais quando $c < 0,05$, conclui-se que, como esperado, todos os textos do banco de dados são compatíveis com línguas naturais.

Livro	r	$\langle L \rangle$	$\langle L \rangle_2$	$\langle C \rangle$	$\langle C \rangle_2$	$\langle I \rangle$	$\langle I \rangle_2$	B	$\langle S \rangle_2$	γ_S
MOR	0,09	0,69	0,64	0,29	0,98	0,94	0,69	0,77	0,13	0,73
CPE	0,08	0,85	0,50	0,28	0,84	0,55	0,24	0,28	0,19	0,54
BRC	0,09	0,47	0,80	0,08	1,00	0,79	0,38	0,96	0,17	0,63
CON	0,07	0,58	0,27	0,37	0,19	0,48	0,22	0,35	0,39	0,52
EJC	0,07	0,72	0,79	0,24	0,57	0,75	0,69	0,91	0,08	0,41
GSV	0,15	0,43	0,79	0,07	0,57	0,43	0,36	0,68	0,24	0,58
MSM	0,10	0,83	0,60	0,60	0,91	0,86	1,00	0,77	0,70	0,84
GUA	0,07	0,80	0,31	0,62	0,51	0,94	0,50	0,70	0,80	0,87
POL	0,12	0,82	0,63	0,18	0,74	0,57	0,35	0,63	0,23	0,55
PBA	0,10	0,79	0,38	0,38	0,88	0,57	0,90	0,91	0,35	0,98

padrão σ_G de cada Gaussiana representando um livro no banco de dados de teste deve ser proporcional à variação de $\tilde{X}$ entre textos distintos em um idioma fixo. Por este motivo, o desvio selecionado foi o pior entre o observado entre o Português e Inglês. Os valores nas Tabelas 5.15 e 5.16 mostram que, como esperado, todos os livros dos dois banco de dados de teste são compatíveis com línguas naturais. Portanto, temos boas indicações que a estratégia proposta é capaz de inferir corretamente se um dado texto é compatível com línguas naturais.

A distância do manuscrito Voynich às línguas naturais foi estimada a partir do cálculo da compatibilidade $c(\tilde{X}_{\text{VMS}}, P(\tilde{X}_{t=\text{nt},l}))$ (ver equações 5.4 e 5.5). Neste caso, P foi construído a partir da adição de Gaussianas centradas ao redor de cada $\tilde{X}$ observado no Novo Testamento em diferentes línguas λ . A distribuição P para três medidas está ilustrada na Figura 13.6. Os valores de $c(\tilde{X}_{\text{VMS}}, P(\tilde{X}_{t=\text{nt},l}))$ na Tabela 5.17 confirmam que o VMS é compatível com

Tabela 5.17 – Compatibilidade do VMS com línguas naturais. Com exceção de $\langle I \rangle_2$ e B , as medidas calculadas para o VMS são consistentes com as medidas esperadas para textos escritos em línguas naturais.

$\tilde{X}_{\text{VMS}}$	r	$\langle L \rangle$	$\langle L \rangle_2$	$\langle C \rangle$	$\langle C \rangle_2$	$\langle I \rangle$	$\langle I \rangle_2$	B	$\langle s \rangle_2$	γ_S
c	0,14	0,62	0,99	0,96	0,05	0,39	0,00	0,00	0,09	0,12

Figura 5.6 – Distribuição das medidas para o Novo Testamento e comparação com o valor obtido para o manuscrito Voynich (linha pontilhada). As medidas consideradas são (a) $\tilde{X} = \langle I \rangle_2$ (intermitência das palavras mais frequentes); (b) $\tilde{X} = r$ (assortatividade) e (c) $\tilde{X} = \langle L \rangle$ (comprimento médio dos caminhos mínimos). Enquanto em (a) o VMS não é compatível com línguas naturais, em (b) e (c) a compatibilidade foi confirmada uma vez que $c(\tilde{X}_{\text{VMS}}, P) > 0,05$.

línguas naturais para a maioria das medidas adequadas para responder Q_2 (ver Tabela 5.6). As exceções são o número de bigramas repetidos b e a intermitência média das palavras mais frequentes $\langle I \rangle_2$. Um alto valor de b é uma característica particular do VMS pois esta quantidade é muito maior que o esperado ao acaso, um padrão que é raro em línguas naturais. Similarmente, o valor de $\langle I \rangle_2$ calculado para o VMS é maior do que os valores tipicamente observados em línguas naturais (conforme ilustra a Figura 13.6(a)), embora a intermitência absoluta das palavras mais frequentes do VMS não esteja muito distante da mesma quantidade observada em línguas naturais. Uma vez que a intermitência está relacionada com a distribuição em larga escala de palavras (chaves) em textos, especulamos que o valor de $\langle I \rangle_2$ para o VMS é inconsistente com línguas naturais porque o VMS é um compêndio de tópicos distintos.

5.6.3 Qual língua é mais similar ao manuscrito Voynich?

Nesta seção abordamos a escolha do idioma mais próximo a um texto de maneira genérica. No entanto, mostraremos que devido à limitação do banco de dados empregado

(o Novo Testamento), não podemos obter uma previsão confiável a respeito do idioma mais próximo de um manuscrito. Dado um texto τ , a língua mais similar a este texto é escolhida de acordo com o seguinte procedimento. Inicialmente calcula-se a distância euclidiana (usando a normalização z (76) das medidas apropriadas para responder a questão Q_3 na Tabela 5.6) entre o livro τ e as versões do Novo Testamento. Seja $\mathcal{R}_{\lambda,\tau}$ a posição obtida pela língua λ no texto τ . Isto é, λ representa a $\mathcal{R}_{\lambda,\tau}$ -ésima língua mais próxima ao texto τ . Dado um conjunto de textos $\mathcal{T}$ escritos na mesma língua, este procedimento gera uma lista $\mathcal{R}_{\lambda,\tau}$ para cada $\tau \in \mathcal{T}$. Neste caso, é interessante combinar os diferentes valores de $\mathcal{R}_{\lambda,\tau}$ considerando o produto das posições normalizadas:

$$\delta_\lambda(\mathcal{T}) = \prod_{\tau \in \mathcal{T}} \frac{\mathcal{R}_{\lambda,\tau}}{|\mathcal{T}|}, \quad (5.8)$$

onde $|\mathcal{T}|$ representa a quantidade de textos presentes no banco de dados $\mathcal{T}$. Esta escolha é motivada pelo fato de que $\mathcal{R}_{\lambda,\tau}/|\mathcal{T}|$ corresponde à probabilidade de se obter ao acaso uma classificação tão boa quanto $\mathcal{R}_{\lambda,\tau}$ em uma inferência de *ranks* aleatória. Desse modo, $\delta_\lambda(\mathcal{T})$ na equação 5.8 corresponde à probabilidade de se obter em uma classificação aleatória uma *rank* melhor ou igual ao obtido pelo nosso método. A partir da classificação das línguas de acordo com $\delta_\lambda(\mathcal{T})$, obtemos uma lista das línguas mais prováveis para os textos em $\mathcal{T}$.

Em nossos experimentos de validação com $|\mathcal{T}| = 15$ textos conhecidos para cada uma das línguas (Tabelas 5.7 e 5.8) verificamos que as medidas apropriadas para responder Q_3 geraram resultados para os livros em Português e Inglês que nem sempre coincidiram com a língua correta. No caso do banco de dados de teste em Português, o Português foi classificado como a segunda melhor língua (após o Grego), enquanto que no banco de dados de teste em Inglês as línguas mais similares foram o Grego e o Russo. Neste caso, o Inglês foi classificado apenas na sexta posição. Embora a língua mais similar não foi compatível com a língua real dos livros, o valor de δ_λ encontrado foi significativamente melhor do que um sistema de classificação aleatório ($p\text{-value} = 1,0 \times 10^{-7}$ e $4,3 \times 10^{-5}$, respectivamente no banco de dados de testes de livros em Português e Inglês.).

A razão pela qual o procedimento anterior é incapaz de prever corretamente a língua de um texto está diretamente relacionada com o uso de apenas um exemplo (uma versão do Novo Testamento) para cada língua, enquanto que em métodos de classificação robustos vários exemplos são empregados para cada uma das classes do problema. Dessa forma, a escolha do idioma mais similar ao idioma do manuscrito Voynich exigirá a introdução de vários livros representando cada idioma, o que é um desafio devido à inexistência de uma vasta quantidade de textos traduzidos em vários idiomas.

Tabela 5.18 – Lista de palavras-chave para o manuscrito Voynich usando a equação 5.6.

Livro	Palavras-chave
Voynich	<i>cthy, qokeedy, shedy, qokain, chor, lkaiin, qol, lchedy, sho, qokaiin, olkeedy, qokal, gotain, dchor, otedy</i>

5.6.4 Palavras-chave do manuscrito Voynich

Nas seções 5.6.1 e 5.6.2 verificamos que o manuscrito Voynich não é compatível com textos aleatórios, sendo compatível com textos escritos em língua natural. Por este motivo, acreditamos que o VMS é passível de ser decodificado e portanto, a detecção das palavras mais relevantes pode ser útil no processo de decodificação. Devido à presença de um único documento escrito no “idioma” Voynich, a aplicação da estratégia *tf-idf* se torna impossível. Da mesma forma, como não é possível conhecer as partições naturais do VMS, torna-se impossível aplicar as estratégias baseadas em entropia (50). Como alternativa aos métodos tradicionais, aplicamos a metodologia de detecção de palavras-chave descrita na Seção 5.5. A lista de palavras obtidas, apresentada na Tabela 5.18, é potencialmente útil para decodificadores uma vez que estes profissionais podem focar seu estudo nas palavras mais relevantes.

5.7 Conclusões

Neste capítulo foram analisadas as propriedades de medidas de texto em uma série de 15 idiomas de famílias distintas. Medidas foram estudadas em termos da sua capacidade de distinguir textos escritos em linguagem natural de textos aleatorizados, e da sua dependência com fatores sintáticos ou semânticos. Com a definição de um arcabouço de análise das medidas, foi possível verificar que a maioria das medidas é informativa em redes de co-ocorrência sem pré-processamento, indicando que estas capturam informação além da frequência das palavras. Aplicando estudos comparativos no mesmo texto (Novo Testamento) nas 15 línguas estudadas e em 15 livros escritos em Português e Inglês foi possível identificar se as métricas dependem mais da sintaxe do idioma ou da semântica do texto. A maioria das medidas de rede depende predominantemente da sintaxe do texto, enquanto que as medidas de intermitência se mostraram dependentes da semântica. Note que a análise de propriedades das métricas foi baseada inteiramente nas propriedades estatísticas das palavras, não exigindo nenhum conhecimento específico sobre o significado das palavras ou alfabeto de codificação dos textos. Por este motivo, a metodologia de estudo das propriedades das medidas estatísticas de texto pode ser usada na prática, por exemplo para estudar características de manuscritos com codificação desconhecida, assim

como foi feito para o manuscrito Voynich.

Capítulo 6

Análise de estilo de textos

A Estilometria analisa o estilo linguístico através da recuperação de atributos descrevendo padrões específicos de escrita. Estudos de estilometria baseiam-se na discriminação de padrões de documentos em uma mesma língua, gênero ou tópico capazes de distinguir características mais sutis como padrões de gênero de autor. Também tem sido investigada a identificação de transições no estilo com o passar de tempo. Hoje, existe interesse em aplicações forenses e em características textuais na Web, como a detecção de padrões que identifiquem características de raça, gênero ou qualidade dos textos (77–79). Os primeiros estudos de estilometria remontam ao início do século XIX com a publicação do trabalho de Mendenhall (80) para reconhecimento de autoria a partir de medidas estatísticas. Outras pesquisas importantes foram realizadas por Yule (81) e por Mosteller e Wallace (51), também relacionados ao reconhecimento de autoria.

As técnicas tradicionais para análise estilística baseiam-se na sua grande maioria na frequência de entidades (p.e., letras, sílabas, palavras etc). Abordagens mais complexas incluem uso de informação sintática ou de informação obtida a partir de técnicas de compressão de dados. Uma representação comum para fins de estudos estilísticos é a modelagem de textos como vetores, onde cada elemento representa a frequência de palavras (ou n -gramas) específicos. Este tipo é comumente denominada de análise *bag of words* (39). Grande parte dos trabalhos de estilo também tem se baseado na eliminação de palavras de conteúdo, já que a aparição destas palavras específicas não depende do estilo mas sim do assunto tratado pelo texto. Embora representando uma minoria das abordagens, conceitos físicos como entropia para estudo da riqueza de vocabulário ou recorrência de padrões de aparição de palavras no texto também já foram usados para caracterizar o estilo de textos (82).

Neste capítulo propomos uma abordagem complementar à abordagem tradicional. Modelamos textos como redes complexas conforme explicitado no Capítulo 2 e estudamos os padrões inferidos através da análise das medidas de redes complexas extraídas dos textos. Neste caso, ao contrário das abordagens tradicionais onde o relacionamento não trivial

entre palavras é muitas vezes desprezado, consideramos associações mais complexas de palavras. Além disso, diferentemente da análise tradicional, eliminamos palavras funcionais a fim de verificar que o relacionamento entre palavras de conteúdo também pode ser útil na análise estilométrica, como sugerido por trabalhos anteriores (83). Mais especificamente, estudamos duas aplicações com análise de estilo em textos. Na Seção 6.1, abordamos o problema de reconhecimento de autoria e na Seção 6.2 apresentamos uma análise de evolução estilística a partir de métricas de redes complexas.

6.1 Reconhecimento de autoria

No campo de reconhecimento de autoria, o objetivo é identificar o autor de documentos cujas identidades são desconhecidas (27). Métodos tradicionais quantitativos realizam análises estatísticas por exemplo no tamanho das palavras (uma perspectiva histórica é apresentada na Ref. (84)). Uma importante contribuição para a área foi introduzida por Mosteller e Wallace (51), que mostraram que a frequência de palavras funcionais (tais como “algum”, “de”, “um” e “sobre”) é útil para caracterizar o estilo de autores. Palavras frequentes também são responsáveis pelo sucesso da abordagem proposta por físicos em que a distância dos *ranks* de frequência de palavras é usada para computar similaridade entre textos (85–87). Outros atributos empregados recentemente são: comprimento de palavras e sentenças, frequência de sinais de pontuação e contrações; e frequência de grafemas, colocações e palavras. Nesta seção, avaliamos como as métricas de rede e séries temporais dependem do estilo do autor.

A investigação da dependência da topologia da rede e outras medidas com a autoria foi estudada em um conjunto de texto de oito autores. Cinco livros foram usados para caracterizar cada um dos autores, conforme mostra a Tabela 6.1. Cada livro foi caracterizado pelas várias medidas apresentadas nas Seções 2.4, 3.6 e 3.8, resumidas segundo o procedimento da Seção 3.9. A média modificada $\langle X \rangle_2$ foi calculada com a versão ponderada com logaritmo (ver equação 13.6 no Capítulo 3), pois essa versão permitiu alcançar maiores taxas de acerto de reconhecimento de autoria. Então, a autoria foi associada usando algoritmos de reconhecimento de padrões. Em todos estes algoritmos, a saída final é a associação de um autor específico para o livro testado. A eficiência é quantificada simplesmente como a fração de associação de autorias realizada corretamente.

Tabela 6.1 – Conjunto de livros para os experimentos de reconhecimento de autoria. Cada um dos 8 autores foi caracterizado por um conjunto de 5 livros.

Título	Autor
<i>Uncle Bernac</i>	Arthur Conan Doyle
<i>The Tragedy of the Korosko</i>	Arthur Conan Doyle
<i>The Valley of Fear</i>	Arthur Conan Doyle
<i>The War in South Africa</i>	Arthur Conan Doyle
<i>Through the Magic Door</i>	Arthur Conan Doyle
<i>Dracula's Guest</i>	Bram Stoker
<i>The Jewel Of Seven Stars</i>	Bram Stoker
<i>The Lady of the Shroud</i>	Bram Stoker
<i>Lair of the White Worm</i>	Bram Stoker
<i>The Man</i>	Bram Stoker
<i>Coral Reefs</i>	Charles Darwin
<i>The Expression of the Emotions</i>	Charles Darwin
<i>Geological Observations on South America</i>	Charles Darwin
<i>The Different Forms of Flowers</i>	Charles Darwin
<i>Volcanic Islands</i>	Charles Darwin
<i>American Notes</i>	Charles Dickens
<i>A Tale of Two Cities</i>	Charles Dickens
<i>Hard Times</i>	Charles Dickens
<i>Great Expectations</i>	Charles Dickens
<i>Barnaby Rudge</i>	Charles Dickens
<i>A Changed Man</i>	Thomas Hardy
<i>A Pair of Blue Eyes</i>	Thomas Hardy
<i>Far from the Madding Crowd</i>	Thomas Hardy
<i>The Hand of Ethelberta</i>	Thomas Hardy
<i>Jude the Obscure</i>	Thomas Hardy
<i>My Man Jeeves</i>	Pelham Grenville Wodehouse
<i>Tales of St. Austin's</i>	Pelham Grenville Wodehouse
<i>The Adventures of Sally</i>	Pelham Grenville Wodehouse
<i>The Clicking of Cuthbert</i>	Pelham Grenville Wodehouse
<i>The Man with Two Left Feet</i>	Pelham Grenville Wodehouse
<i>The Works of Allan Poe (5 volumes)</i>	Edgar Alan Poe
<i>Beasts and Super Beasts</i>	Hector Hugh Munro
<i>The Chronicles of Clovis</i>	Hector Hugh Munro
<i>The Toys of Peace</i>	Hector Hugh Munro
<i>The Unbearable Bassington</i>	Hector Hugh Munro
<i>When William Came</i>	Hector Hugh Munro

Tabela 6.2 – Taxa de acerto com os três melhores classificadores usando todos atributos e a melhor combinação destes atributos.

Classificadores	C4.5	κ -NN	<i>Naive Bayes</i>
Todos atributos	50,0%	47,5%	42,5%
Melhor caso	62,5%	65,0%	62,5%

Tabela 6.3 – *p-values* referentes às taxas de acerto apresentadas na Tabela 6.2. As probabilidades correspondem à chance de se obter ao acaso uma taxa de acerto maior ou igual em uma tentativa.

Classificadores	C4.5	κ -NN	<i>Naive Bayes</i>
Todos atributos	1×10^{-8}	6×10^{-8}	2×10^{-6}
Melhor caso	5×10^{-9}	4×10^{-10}	5×10^{-9}

6.1.1 Resultados da classificação

Os resultados sumarizados na Tabela 6.2 indicam que as taxas de acerto variam entre 42,5 % e 50,0 % quando todos os atributos são utilizados. Os resultados são estatisticamente significativos (ver Tabela 6.3), o que confirma que os atributos estudados capturam sutilezas de escrita que são específicas para cada autor. A fim de analisar mais em detalhe o desempenho dos diferentes métodos, consideramos os casos em que apenas alguns atributos foram incluídos nos algoritmos. Foram testadas 2^{15} combinações de 15 atributos obtendo-se uma taxa de acerto máxima igual a 65,0 %.

6.1.2 Relevância de atributos

Para entender a dependência entre estilo de autores e as métricas estudadas, identificamos a relevância de cada atributo segundo o método AUC. Neste método a relevância dos atributos é avaliada considerando suas possíveis inter-relações. Os resultados para os três melhores algoritmos de reconhecimento de padrões estão ilustrados nas colunas 2-4 da Tabela 13.4. Os três atributos que aparecem como mais importantes são $\langle N \rangle$, γ_I e $\langle N \rangle_2$. A classificação da relevância das métricas parece ser pouco dependente do algoritmo de reconhecimento de padrões. De fato, o coeficiente de correlação de *ranks* de Spearman^a confirma a consistência observada. Esta é uma forte indicação de que a análise vai além de resultados específicos de algoritmos e de fato captura a influência dos atributos.

Também é importante comparar os resultados de relevância considerando cada atributo de forma isolada. Este tipo de análise foi realizado verificando a taxa de acerto dos

^a As correlações observadas foram 0,29 (*p-value* = 0,145), 0,49 (*p-value* = 0,032) e 0,67 (*p-value* = 0,003) respectivamente para os pares C4.5/ κ -NN, C4.5/Bayes e κ -NN/Bayes. Neste caso, os *p-values* são calculados adotando como hipótese nula que os *ranks* são independentes.

Tabela 6.4 – *Rank* de atributos baseado na taxa de acerto dos classificadores. 1 significa primeira posição, 2 significa segunda posição e assim por diante. Os resultados para cada classificador (C4.5, κ -NN e *Bayes*) estão ilustrados para cada um dos métodos de quantificação de relevância: método AUC (análise conjunta, colunas 2-4) e método KL (análise isolada, coluna 5). A análise isolada também foi realizada verificando a taxa de acerto de cada algoritmo, quando apenas um dos atributos é utilizado (colunas 6-8). A última coluna mostra o coeficiente de correlação de Pearson entre cada atributo e o tamanho do vocabulário M (número de palavras distintas) calculado em todos os 40 livros. Os atributos na tabela estão ordenados em ordem decrescente da média geométrica dos *ranks* obtidos nos 3 métodos (esta ordenação corresponde também aos *ranks* da probabilidades de cada atributo obter ao acaso um *rank* igual ou melhor ao obtido nos 3 esquemas de ordenação).

	Análise Conjunta			Análise Isolada				Correlação com M
	C4.5	κ -NN	<i>Bayes</i>	KL	C4.5	κ -NN	<i>Bayes</i>	
$\langle N \rangle_2$	6	1	1	3	2	5	1	-0,90
γ_I	2	2	2	10	12	9	10	-0,08
$\langle N \rangle$	1	6	3	2	1	2	3	-0,96
$\langle L \rangle$	7	4	6	9	5	3	8	0,85
$\langle B \rangle$	5	8	5	1	3	1	2	0,98
$\langle I \rangle_2$	10	3	10	15	15	12	12	-0,34
$\langle L \rangle_2$	8	7	8	8	5	7	5	0,85
$\langle C \rangle$	12	11	4	6	5	5	5	-0,87
γ_L	4	13	11	13	10	9	13	-0,13
γ_B	3	14	14	11	8	9	9	-0,07
$\langle B \rangle_2$	9	9	9	7	9	14	5	0,88
$\langle C \rangle_2$	11	10	7	5	4	3	4	-0,87
$\langle I \rangle$	13	5	12	12	13	15	10	-0,29
γ_N	15	12	13	4	10	8	13	0,81
γ_C	14	15	15	14	14	12	15	0,07

classificadores quando apenas o atributo é utilizado na classificação. A relevância isolada também foi analisada com o método KL. Os resultados estão ilustrados nas colunas 5-8 da Tabela 13.4. Note que alguns atributos com alta relevância no método que considera a interação de atributos não são discriminativos quando analisados isoladamente (p.e., a assimetria da intermitência γ_I). Por outro lado, atributos que são bem classificados na análise de atributos isolada nem sempre aparecem entre os mais bem classificados na análise de relevância que considera a interação entre atributos. A influência das métricas estudadas no estilo de autores pode ser resumida da seguinte maneira:

- **Frequência:** esta foi a quantidade mais eficiente no reconhecimento de autoria, em que $\langle N \rangle$ e $\langle N \rangle_2$ aparecem entre os três atributos mais importantes. Isto significa que o número distinto de palavras no texto também é um fator importante no reconhecimento de autoria, já que

$$\langle N \rangle = \frac{\text{comprimento do livro}}{M} \quad (6.1)$$

e o comprimento do livro foi mantido constante. A média modificada $\langle N \rangle_2 = \langle \log N \rangle$ também captura este aspecto, incluindo a proporção de palavras frequentes e pouco frequentes. Por outro lado, o fraco desempenho de γ_N ($= \alpha$ em $p(N) \sim N^{-\alpha}$) possivelmente é consequência da universalidade da lei de Zipf (para um tamanho de livro fixo (9)).

- *Betweenness:* o *betweenness* se mostrou útil devido à sua correlação com o tamanho do vocabulário M (veja última coluna da Tabela 13.4). Em termos de rede, esta relação corresponde a uma relação linear entre $\langle B \rangle$ e o tamanho da rede (M é o número de vértices). $\langle B \rangle_2$ e γ_B apresentaram baixa relevância para a tarefa. Este resultado sugere que o número de palavras com alto *betweenness* não é um atributo relevante para distinguir o estilo de autores.
- **Caminhos mínimos:** este foi o atributo de rede que apresentou melhor desempenho. $\langle L \rangle$ quantifica a distância típica das palavras aos *hubs* da rede (i.e., as palavras mais frequentes). O bom desempenho aponta para uma dependência no estilo dos autores. Com relação às quantidades $\langle L \rangle_2$ e γ_L , o desempenho foi inferior ao observado para $\langle L \rangle$. Isto indica que a dependência de estilo em L é mais forte para valores típicos do que para valores de palavras frequentes.
- **Coeficiente de aglomeração:** o fraco desempenho de todas as quantidades relacionadas ao coeficiente de aglomeração sugere que autores possuem pouca liberdade para variar esta medida em redes de co-ocorrência. A última posição no *rank* de γ_C na Tabela 13.4 sugere que a fração das palavras usadas em contextos específicos (palavras com alto coeficiente de aglomeração) não depende dos diferentes estilos

de autores. As duas médias $\langle C \rangle$ e $\langle C \rangle_2$ assumem valores similares e são relevantes apenas quando usadas isoladamente, possivelmente devido à sua correlação com o tamanho do vocabulário M .

- **Intermitência:** após a medida de frequência, a intermitência foi a quantidade mais relevante na Tabela 13.4. Em especial, a assimetria da distribuição γ_I desempenha papel fundamental na discriminação de autoria. Isto quer dizer que a fração de palavras utilizadas como palavras-chave representa um fator importante para reconhecer autoria. Por outro lado, a média simples $\langle I \rangle$ foi pouco relevante. Dentre todas as quantidades analisadas, aquela que apresentou o menor valor absoluto de correlação com a frequência foram as medidas de intermitência, o que explica porque γ_I apresentou fraco desempenho quando avaliado isoladamente. Finalmente, $\langle I \rangle_2$ apresentou desempenho melhor que $\langle I \rangle$, sugerindo que a intermitência das palavras mais frequentes é mais relevante que a intermitência de todas as palavras do texto.

6.1.3 Comparação com métodos tradicionais

O objetivo dos experimentos realizados para reconhecimento de autoria foi avaliar a dependência e a importância dos diferentes fatores para a tarefa de reconhecimento de autoria. Mesmo assim, é interessante comparar a eficiência dos resultados obtidos com outros métodos tradicionais da literatura. Uzuner e Katz (88) utilizaram um conjunto de textos similar ao da Tabela 6.1, construído com oito autores. Eles usaram cinco conjuntos de atributos, incluindo estatísticas simples e análises estatísticas mais sofisticadas (veja por exemplo a Tabela 3 da Ref. (88)). Nosso melhor resultado (taxa de acerto igual a 65 %) é comparável com o segundo melhor caso (acerto de 62 %) citado na Ref. (88) quando “elementos sintáticos de expressão” são utilizados. O melhor resultado encontrado por Uzuner e Katz (88) (acerto de 87 %) foi obtido usando palavras funcionais. Em outro estudo, Grieve (89) obteve taxas de acerto variando entre 33 % e 87 % para o caso de 5 autores e taxas variando entre 18 % e 80 % para 10 autores (Tabela 9 da Ref. (89)). Neste caso, um conjunto de 34 atributos foi utilizado. Os resultados desta tese são melhores que a mediana dos resultados citados anteriormente, obtidos com um conjunto distinto de atributos. Novamente, os melhores resultados relatados por Grieve (89) foram obtidos com atributos baseados na frequência das palavras funcionais. Tais resultados estão de acordo com os primeiros da área relatados por Mosteller e Wallace (51), que usaram a frequência das palavras funcionais para identificar autoria.

A fim de confirmar a importância das palavras funcionais, uma série de esquemas de predição foi implementada usando a frequência N_i de palavras frequentes (são na maioria palavras funcionais). Diferentemente da abordagem utilizada até agora, baseada na média

e nas propriedades de escala, empregamos a frequência de palavras específicas diretamente como entrada para os algoritmos de classificação. Neste caso, usamos apenas o algoritmo κ -NN porque os outros algoritmos não forneceram resultados satisfatórios com muitos atributos. A partir da lista de 70 *stopwords* da Tabela 2.5 da Ref. (51), obtivemos uma taxa de acerto comparável ao melhor resultado de nossa análise (62,5 %). Seguindo o estudo descrito na Ref. (89), duas outras listas de palavras foram também usadas: todas as 1978 palavras que aparecem em pelo menos um dos livros de cada autor, levando a uma taxa de acerto de 90 %; e com todas as 209 palavras que aparecem pelo menos uma vez em todos os livros, levando a uma taxa de acerto de 82,5 %. Estes resultados confirmam que o uso apenas de atributos de redes e intermitência não representa a melhor metodologia. Apesar disso, a combinação de atributos pode ser útil para aperfeiçoar a discriminação.

A fim de ilustrar como os atributos propostos neste projeto podem ser complementares aos métodos tradicionais, alguns experimentos foram propostos usando a frequência e a intermitência do conjunto de palavras compostos pelas cinco palavras mais frequentes de cada livro. A taxa de acerto no reconhecimento de autoria quando consideramos apenas a frequência foi de 72,5 %. Quando apenas a intermitência foi usada, a taxa obtida foi de 37,5 %. Embora esta última taxa de acerto não seja impressionante, ela ainda é estatisticamente significativa ($p = 2,2 \times 10^{-4}$) e mostra que os valores de intermitência de palavras específicas entre autores distintos são diferentes. Quando ambos os atributos são combinados, a taxa de acerto aumenta para 80 %. Isto sugere que a combinação adequada de atributos tradicionais e atributos baseados na caracterização de textos como sistemas complexos e dinâmicos pode ser útil para melhorar o desempenho de reconhecimento de autoria.

6.2 Estilos literários de livros

Várias descobertas relacionadas à linguagem e aos valores culturais têm sido resultado da aplicação de métodos estatísticos para analisar grandes quantidades de textos. Exemplos recentes desta análise são o estudo dos padrões estatísticos em milhões de livros (90) e o estudo de mensagens do *twitter* para análise de humor (91) ou para análise de tendência de uso da linguagem entre países distintos (92). Em vários destes exemplos, conhecimento é inferido a partir da análise do conteúdo semântico dos textos. Nesta seção, estudamos a evolução do estilo da linguagem através de uma abordagem baseada na organização estrutural das redes modelando textos. Em especial, mostramos que padrões estilísticos de métricas de redes são úteis não apenas para identificar autoria (como verificado na Seção 6.1), mas também para caracterizar estilos de um ponto de vista mais amplo. Em outras palavras, verificamos se livros escritos na mesma época apresentam padrões de escrita

similares.

6.2.1 Agrupamento de livros por época

Para analisar a variação de estilo em livros de autores variados empregamos 58 livros, cuja data de publicação varia entre 1590 a 1922. As Tabelas 6.5 e 6.6 fornecem os detalhes dos livros. Tendo limitado o tamanho dos livros aos seus primeiros 21900 *tokens* a fim de minimizar a dependência das medidas com o comprimento do texto, cada um dos livros foi representado pela modelagem de co-ocorrência de palavras descrita na Seção 3.5.1. Como na análise de reconhecimento de autoria, cada livro foi caracterizado pelas medidas apresentadas nas Seções 2.4, 3.6 e 3.8, resumidas segundo o procedimento da Seção 3.9. A média modificada $\langle X \rangle_2$ foi calculada apenas nas palavras mais frequentes (ver equação 13.6 no Capítulo 3) pois essa versão permitiu maior distinguibilidade de livros por épocas.

Para verificar como os estilos se diferenciam, uma divisão inicial arbitrária dos livros em 6 intervalos de 50 anos de acordo com a data de publicação foi obtida, levando aos grupos ilustrados na Figura 6.1. Houve pouca distinção, especialmente quando o desvio padrão ilustrado em forma de elipses (93) dentro da figura foi considerado. A separação só era razoável para períodos distantes no tempo, já que suas elipses não se interseccionam. Esta dificuldade em se distinguir movimentos literários provavelmente ocorre porque não existe motivo para que grupos separados apareçam com divisões arbitrárias de 50 anos. A distinguibilidade verificada com o método de análise de componentes principais (PCA) também mostrou ser fraca.

A possibilidade de distinção para uma divisão não arbitrária foi verificada com um procedimento automático para encontrar a melhor partição do banco de dados através de uma estratégia de otimização. Escolheu-se a melhor partição sob a condição de que livros com data de publicação consecutiva deveriam pertencer ao mesmo grupo ou pertencer às bordas de grupos consecutivos. Em outras palavras, após a ordenação do conjunto de livros de acordo com a data de publicação, variamos a quantidade de grupos e a posição dos seus delimitadores, quantificando para cada caso a qualidade dos grupos de acordo com os índices de qualidade SWC e DN. Uma boa separabilidade entre grupos foi obtida quando o número de grupos variou entre 3 e 7, conforme apontado pelos dois índices de qualidade. A melhor partição, que é estatisticamente significativa (ver Figura 6.3), foi obtida com 6 grupos, levando ao agrupamento ilustrado na Figura 6.2. Surpreendentemente, os 6 períodos inferidos a partir desta análise coincidem com os estilos literários tradicionais, conforme listado na Tabela 6.7.

Outras características importantes também podem ser inferidas a partir da análise da Figura 6.2. Grupos referentes a períodos literários consecutivos se posicionam normalmente

Tabela 6.5 – Livros usados nos experimentos de análise de variação de estilo com a época de publicação. O ano de publicação está indicado após o nome do livro.

Livro	Autor
<i>The Faerie Queene</i> (1590)	Edmund Spenser
<i>Don Quixote</i> (1605)	Miguel de Cervantes
<i>The Advancement of Learning</i> (1605)	Francis Bacon
<i>True Christianity</i> (1612)	Johann Arndt
<i>Devotions upon Emergent Occasions</i> (1624)	John Donne
<i>Leviathan</i> (1651)	Thomas Hobbes
<i>The Love Letters of Dorothy Osborne</i> (1653)	Dorothy Osborne
<i>Experiments and Considerations Touching Colours</i> (1664)	Robert Boyle
<i>The Pilgrim's Progress</i> (1678)	John Bunyan
<i>Love-Letters Between a Nobleman and His Sister</i> (1684)	Aphra Behn
<i>The Spectator</i> (1711)	Joseph Addison
<i>Robinson Crusoe</i> (1719)	Daniel Defoe
<i>Gulliver's Travels</i> (1726)	Jonathan Swift
<i>Selected Sermons of Jonathan Edwards</i> (1731-1750)	Jonathan Edwards
<i>The Letters of Horace Walpole</i> (1735-1748)	Horace Walpole
<i>Pamela, or Virtue Rewarded</i> (1740)	Samuel Richardson
<i>The Fortunate Foundlings</i> (1744)	Eliza Haywood
<i>Clarissa, or, the History of a Young Lady</i> (1748)	Samuel Richardson
<i>Life's Progress through the Passions</i> (1748)	Eliza Haywood
<i>The History of Tom Jones, a Foundling</i> (1749)	Henry Fielding
<i>The Vicar of Wakefield</i> (1761)	Oliver Goldsmith
<i>The Life and Opinions of Tristram Shandy</i> (1767)	Laurence Sterne
<i>Thoughts on the Present Discontents</i> (1770)	Edmund Burke
<i>The Expedition of Humphry Clinker</i> (1771)	Tobias Smollett
<i>Evelina</i>	Frances Burney
<i>Cecilia, Memoirs of an Heiress</i> (1782)	Frances Burney
<i>A Vindication of the Rights of Woman</i> (1791)	Mary Wollstonecraft
<i>Autobiography of Benjamim Franklin</i> (1793)	Benjamim Franklin
<i>The Mysteries of Udolpho</i> (1794)	Ann Radcliffe

próximos na projeção, indicando mudanças suaves no estilo literário ao longo do tempo. A mesma conclusão pode ser inferida a partir da análise hierárquica na Figura 6.4, realizada com o método Ward (94). A exceção a essa tendência foi a mudança referente ao período 1794 – 1818 → 1826 – 1906, que pode ter sido uma consequência de uma mudança drástica desencadeada pela Revolução Francesa de 1789. Com relação à variância entre grupos, a menor ocorreu no período 1590 – 1653, enquanto que a maior ocorreu no período 1906 – 1922. Estes resultados são intuitivos já que uma menor variação de estilo é esperada em períodos antigos, enquanto que nos períodos mais recentes uma menor uniformidade pode ter sido resultado da coexistência de vários estilos de escrita.

Tabela 6.6 – Livros usados nos experimentos de análise de variação de estilo com a época de publicação (continuação). O ano de publicação está indicado após o nome do livro.

Livro	Autor
<i>Sense and Sensibility</i> (1811)	Jane Austen
<i>Pride and Prejudice</i> (1813)	Jane Austen
<i>Emma</i> (1815)	Jane Austen
<i>Frankenstein</i> (1818)	Mary Shelley
<i>Persuasion</i> (1818)	Jane Austen
<i>The Last of the Mohicans</i> (1826)	James Fenimore Cooper
<i>The Voyage of the Beagle</i> (1839)	Charles Darwin
<i>The Black Tulip</i> (1844)	Alexandre Dumas
<i>Wuthering Heights</i> (1847)	Emily Bronte
<i>David Copperfield</i> (1850)	Charles Dickens
<i>Great Expectations</i> (1861)	Charles Dickens
<i>The Headless Horseman</i> (1865)	Mayne Reid
<i>The Innocents Abroad</i> (1869)	Mark Twain
<i>Daniel Deronda</i> (1876)	George Eliot
<i>Two on a Tower</i> (1882)	Thomas Hardy
<i>Life's Little Ironies</i> (1894)	Thomas Hardy
<i>Dracula</i> (1897)	Bram Stoker
<i>Jane Eyre</i> (1897)	Charlotte Bronte
<i>The Invisible Man</i> (1897)	Herbert George Wells
<i>The Jewel of Seven Stars</i> (1903) s	Bram Stoker
<i>The House of Mirth</i> (1905)	Edith Wharton
<i>The Jungle</i> (1906)	Upton Sinclair
<i>The Lady of the Shroud</i> (1909)	Bram Stoker
<i>The Lair of the White Worm</i> (1911)	Bram Stoker
<i>Pollyanna</i> (1913)	Eleanor H. Porter
<i>The New Freedom</i> (1913)	Woodrow Wilson
<i>Dear Enemy</i> (1915)	Jean Webster
<i>The Age of Innocence</i> (1920)	Edith Wharton
<i>Ulysses</i> (1922)	James Joyce

6.2.2 Relevância de atributos

Os fatores mais importantes para a separabilidade dos estilos literários foram determinados de duas formas distintas. Na primeira técnica, uma métrica foi considerada relevante quando os valores referentes a grupos distintos se mostraram significativamente diferentes. A lista de medidas e o valor correspondente de p -value para a diferença das medidas para um dado par de grupos são dados na Tabela 6.8. A assimetria da distribuição do comprimento médio dos caminhos mínimos γ_L e o tamanho do vocabulário M apresentaram as maiores variações. Note que esses resultados são similares aos encontrados para o problema de reconhecimento de autoria, já que estas duas medidas também foram úteis para caracterizar o estilo pessoal de autores (ver Seção 6.1.2). No segundo tipo de

Figura 6.1 – *Scatter plot* (projção CVA) representando o estilo de cada livro de acordo com 6 estilos literários. Cada estilo é representado por um conjunto de 10 livros. O gráfico interno ilustra a dispersão dos estilos literários.

avaliação, um atributo foi considerado relevante quando se mostrou capaz de fornecer boa distinguibilidade entre grupos considerando as interdependências entre as métricas. Este tipo de avaliação foi realizado através do cálculo da importância de cada medida para o eixo de projeção. Os resultados nas Tabelas 6.9 e 6.10 apontam o coeficiente de aglomeração ($\langle C \rangle$ e $\langle C \rangle_2$) como o principal fator de distinguibilidade. Uma vez que existem evidências que o coeficiente de aglomeração quantifica o quanto palavras estão restritas a contextos específicos ou genéricos^b (27), estes resultados sugerem que a frequência de uso de palavras

^b Palavras de contexto específico são aquelas aparecendo em uma pequena quantidade de contextos. Por exemplo, a palavra “professor” usualmente induz conceitos relacionados ao ambiente de aprendizado. Por outro lado, palavras genéricas podem

Tabela 6.7 – Relação entre o melhor agrupamento baseado em fatores topológicos de redes complexas e os estilos tradicionais.

Agrupamento	Estilo	Nome	Referência
1590 - 1653	1558 - 1603	Período Elisabetano	(95)
1664 - 1761	1660 - 1798	Neoclassicismo/Renascimento	(96–98)
1767 - 1793	1660 - 1798	Neoclassicismo/Renascimento	(96, 99)
1794 - 1818	1764 - 1820	Literatura gótica	(96, 99)
1826 - 1906	1830 - 1900	Realismo	(96)
1826 - 1906	1865 - 1900	Naturalismo	(96, 100)
1906 - 1922	1890 - 1940	Modernismo	(96, 101)

Figura 6.2 – *Scatter plot* ilustrando o melhor agrupamento obtido para os estilos literários. A elipses de variância evidenciam que há maior distinguibilidade de períodos em relação a distinguibilidade obtida com duração fixa de períodos literários (Figura 6.1).

Figura 6.3 – Teste de significância realizado para (a) silhueta simplificada SWC e (b) índice Dunn DN. Os histogramas representam os valores de qualidade dos *clusters* considerando uma distribuição aleatória de pontos. As linhas pontilhadas correspondem ao valor do índice computado para o agrupamento da Figura 6.2. Uma vez que o valor da silhueta simplificada para o caso aleatório $SWC_{rand} = 0,187 \pm 0,036$ é sempre menor que o valor obtido para o agrupamento da Figura 6.2, o agrupamento inferido neste último caso é estatisticamente significativo. Esta mesma conclusão vale para o índice Dunn, já que $DN_{rand} = 0,059 < DN = 0,207$.

Tabela 6.8 – Lista de transições mais significativas. Quando avaliadas individualmente, as medidas mais importantes para distinguir *clusters* são o tamanho do vocabulário M e o terceiro momento da distribuição do comprimento médio dos caminhos mínimos γ_L .

Medida	Transição	p -value
Vocabulário M	1590 – 1653 $\rightarrow$ 1794 – 1818	0,048
Vocabulário M	1664 – 1761 $\rightarrow$ 1767 – 1793	0,051
Vocabulário M	1664 – 1761 $\rightarrow$ 1826 – 1906	0,001
Vocabulário M	1767 – 1793 $\rightarrow$ 1794 – 1818	0,011
Vocabulário M	1794 – 1818 $\rightarrow$ 1826 – 1906	$< 0,001$
Assortatividade r	1590 – 1653 $\rightarrow$ 1767 – 1793	0,008
Assortatividade r	1590 – 1653 $\rightarrow$ 1826 – 1906	0,044
Assortatividade r	1664 – 1761 $\rightarrow$ 1767 – 1793	0,041
Assortatividade r	1664 – 1761 $\rightarrow$ 1826 – 1906	0,006
Caminhos mínimos $\langle L \rangle$	1664 – 1761 $\rightarrow$ 1826 – 1906	0,049
Caminhos mínimos $\langle L \rangle_2$	1664 – 1761 $\rightarrow$ 1906 – 1922	0,050
Caminhos mínimos σ_L	1590 – 1653 $\rightarrow$ 1906 – 1922	0,031
Caminhos mínimos σ_L	1664 – 1761 $\rightarrow$ 1906 – 1922	0,022
Caminhos mínimos σ_L	1767 – 1793 $\rightarrow$ 1906 – 1922	0,023
Caminhos mínimos σ_L	1826 – 1906 $\rightarrow$ 1906 – 1922	$< 0,001$
Caminhos mínimos γ_L	1590 – 1653 $\rightarrow$ 1826 – 1906	0,028
Caminhos mínimos γ_L	1590 – 1653 $\rightarrow$ 1906 – 1922	$< 0,001$
Caminhos mínimos γ_L	1664 – 1761 $\rightarrow$ 1906 – 1922	$< 0,001$
Caminhos mínimos γ_L	1767 – 1793 $\rightarrow$ 1906 – 1922	0,001
Caminhos mínimos γ_L	1794 – 1818 $\rightarrow$ 1906 – 1922	0,019
Caminhos mínimos γ_L	1826 – 1906 $\rightarrow$ 1906 – 1922	$< 0,001$
Aglomeração $\langle C \rangle$	1664 – 1761 $\rightarrow$ 1767 – 1793	0,048
Aglomeração $\langle C \rangle$	1664 – 1761 $\rightarrow$ 1826 – 1906	0,051
Aglomeração $\langle C \rangle_2$	1664 – 1761 $\rightarrow$ 1767 – 1793	0,054
Aglomeração $\langle C \rangle_2$	1664 – 1761 $\rightarrow$ 1826 – 1906	0,055
Aglomeração σ_C	1664 – 1761 $\rightarrow$ 1767 – 1793	0,054
Aglomeração γ_C	1590 – 1653 $\rightarrow$ 1767 – 1793	0,045

Tabela 6.9 – Importância de cada medida para a análise de variáveis canônicas. As medidas de coeficiente de aglomeração e comprimento médio dos caminhos mínimos se mostraram as mais relevantes no eixo principal.

Quantidades (Primeiro eixo)	Relevância (Primeiro eixo)
$\langle C \rangle_2$	33,3 %
$\langle C \rangle$	31,6 %
σ_C	6,6 %
$\langle L \rangle$	6,4 %
r	5,1 %

calculado matematicamente como:

$$\vec{a}_i = \frac{1}{i} \sum_{j=1}^i \vec{v}_j. \quad (6.2)$$

Tabela 6.10 – Importância de cada medida para o segundo eixo da análise de variáveis canônicas. O coeficiente de aglomeração e o comprimento médio dos caminhos mínimos apareceram como as mais relevantes no eixo secundário.

Quantidade (Segundo eixo)	Relevância (Segundo eixo)
$\langle C \rangle$	34,5 %
$\langle C \rangle_2$	33,7 %
$\langle L \rangle_2$	9,5 %
$\langle L \rangle$	9,4 %
σ_C	3,4 %

Figura 6.5 – Dinâmica (a) do coeficiente de aglomeração médio $\langle C \rangle$; (b) do tamanho do vocabulário M ; e (c) do coeficiente de lei de potência da distribuição de frequência γ_N . Enquanto $\langle C \rangle$ e M oscilam, γ_N tende a diminuir com o tempo, evidenciando que palavras são usadas mais uniformemente nos últimos períodos.

Dado $\vec{a}_i$, o *estado oposto* ao estado atual i (a Figura 6.6a fornece uma interpretação geométrica) é dado por:

$$\vec{r}_i = \vec{v}_i + 2(\vec{a}_i - \vec{v}_i) = 2\vec{a}_i - \vec{v}_i. \quad (6.3)$$

Tendo calculado os valores $\vec{r}_i$ e $\vec{v}_i$, o *vetor oposição* $\vec{D}_i$ em relação a $\vec{v}_i$ (ver Figura 6.6a) é dado por:

$$\vec{D}_i = \vec{r}_i - \vec{v}_i. \quad (6.4)$$

Para dois elementos consecutivos na série i e j , o vetor $\vec{M}_{ij}$ representando a mudança de estilo é

$$\vec{M}_{ij} = \vec{v}_j - \vec{v}_i. \quad (6.5)$$

O vetor $\vec{M}_{ij}$ é importante porque o seu módulo $\|\vec{M}_{ij}\|$ quantifica a mudança de estilo em relação ao estado anterior $\vec{v}_i$. Com $\vec{M}_{ij}$, o *índice de oposição* op_{ij} é computado pelo

Figura 6.6 – Ilustração das quantidades empregadas para definir os conceitos de oposição (op), *skewness* (skw) e contra-dialética (cd).

componente de $\vec{M}_{ij}$ projetado em $\vec{D}_i$:

$$\text{op}_{ij} = \frac{\vec{M}_{ij} \cdot \vec{D}_i}{\|\vec{D}_i\|}. \quad (6.6)$$

Note que se o elemento atual da série tende a se opor ao anterior, então o componente de $\vec{M}_{ij}$ em $\vec{D}_i$ assumirá um alto valor.

O índice *skewness* skw_{ij} , ilustrado na Figura 6.6a, é definido como a distância entre $\vec{v}_j$ e a linha definida por $\vec{D}_i$. Este índice quantifica como o elemento atual da série se distancia do movimento de oposição. Ele é útil, por exemplo, para identificar oscilações triviais dentro da linha L_i , já que neste caso uma série de movimentos com $\text{skw} = 0$ seria observado.

A dialética entre três elementos consecutivos i , $j = i + 1$ e $k = j + 1 = i + 2$ é quantificada da seguinte maneira. Se $\vec{v}_k$ é o resultado da síntese de estilos representados por $\vec{v}_i$ e $\vec{v}_j$, então a distância d_{ik} entre $\vec{v}_k$ e a mediana ML_{ij} definida por $\vec{v}_i$ e $\vec{v}_j$ (ver Figura 6.6b) será pequena. O índice de contra-dialética^c cd_{ik} é

$$\text{cd}_{ik} = \frac{d_{ik}}{\|\vec{M}_{ij}\|}. \quad (6.7)$$

Detalhes adicionais relacionados à definição do conceito de oposição, *skewness* e contra-dialética são fornecidos na Ref. (102).

A aplicação dos índices definidos anteriormente levou aos seguintes resultados relacionados à evolução dos estilos literários. As mudanças de estilo entre dois grupos consecutivos parecem ter sido influenciadas por movimentos de oposição aos estilos anteriores, assim como geralmente ocorre com a temática^d. Isto pode ser observado a partir da análise do

^c Note que denominamos ρ_{ik} como índice de *contra*-dialética (e não dialética) pois ele é definido como uma distância. Dessa forma, há uma relação inversa entre ρ_{ik} e o conceito de dialética.

^d Por exemplo, enquanto o Romantismo é caracterizado pelo subjetivismo, individualismo e nacionalismo, o Realismo é caracterizado pelo objetivismo e valorização de personagens inspirados na realidade.

coeficiente de oposição op_{ij} na Tabela 6.11, que quantifica geometricamente o conceito de oposição. Com exceção do período $1826 - 1906 \rightarrow 1909 - 1922$, todos os valores de op_{ij} foram positivos. Além disso, pode-se notar que o movimento de oposição é mais intenso que o movimento *skewness* (skw_{ij}), que quantifica o quanto a transição do estilo atual se desviou do movimento de oposição. A evolução do estilo também foi estudada em termos da contra-dialética cd_{ijk} . Em outras palavras, estudamos quantitativamente o quanto um estilo pode ser considerado como consequência dos dois anteriores. Neste caso, nenhum padrão claro pôde ser identificado na Tabela 6.12. O período com menor valor de contra-dialética cd_{ijk} e portanto com maior dialética aconteceu no século XIX. Portanto, em termos de estilos de escrita, o realismo é o estilo literário que mais se aproxima de uma síntese de estilos anteriores.

Tabela 6.11 – Índices de oposição (op_{ij}) e *skewness* (skw_{ij}). Note que $op_{ij} > 0$ para a maioria das transições, o que confirma que as transições são realizadas em oposição ao movimento anterior. Além disso, o movimento de oposição é praticamente sempre mais intenso que o movimento de *skewness*.

Período	op_{ij}	skw_{ij}
1590 - 1653 $\rightarrow$ 1664 - 1761	1,00	0,00
1664 - 1761 $\rightarrow$ 1767 - 1793	0,39	0,08
1767 - 1793 $\rightarrow$ 1794 - 1818	0,35	0,18
1794 - 1818 $\rightarrow$ 1826 - 1906	1,09	0,07
1826 - 1906 $\rightarrow$ 1909 - 1922	-0,01	0,08

Tabela 6.12 – Índice de contra-dialética (cd_{ijk}), quantificando o quanto um período pode ser visto como resultado dos dois estilos anteriores. Por apresentar o menor valor de contra-dialética, o realismo (1826 - 1906) é o estilo literário que mais se aproxima de uma síntese de estilos anteriores.

Período	cd_{ijk}
1590 - 1653 $\rightarrow$ 1664 - 1761 $\rightarrow$ 1767 - 1793	0,76
1664 - 1761 $\rightarrow$ 1767 - 1793 $\rightarrow$ 1794 - 1818	1,49
1767 - 1793 $\rightarrow$ 1794 - 1818 $\rightarrow$ 1826 - 1906	0,39
1794 - 1818 $\rightarrow$ 1826 - 1906 $\rightarrow$ 1909 - 1922	0,69

6.3 Conclusões

Os estudos de análise de autoria e de evolução da escrita mostraram que o estilo de diferentes autores e estilos literários deixam marcas em medidas de textos baseadas em

redes de co-ocorrência e na intermitência das palavras. As classificações e agrupamentos estatisticamente significativos obtidos no reconhecimento de padrões de estilo em livros mostram de modo inequívoco que a dependência com estes atributos pode ser usada na prática. Com relação à importância dos diferentes atributos, o *rank* de atributos pode depender do banco de dados, dos atributos selecionados e dos algoritmos de classificação. Conforme enfatizado na literatura (89), diferentes algoritmos e atributos precisam ser testados em um dado corpus antes da implementação de qualquer algoritmo de reconhecimento de padrões. No entanto, a robustez de nossos resultados usando algoritmos baseados em paradigmas distintos sugere que a importância dos atributos estudados vai além de características específicas dos algoritmos. Em especial, três padrões merecem destaque na análise de estilo: (i) a assimetria da distribuição da intermitência das palavras γ_I (na análise de reconhecimento de autoria), que é baseada na distribuição de larga escala de palavras e quantifica a porcentagem de palavras-chave (alto valor de $I \equiv \sigma_T / \langle T \rangle$) usadas no livro; (ii) o coeficiente de aglomeração (na análise de estilos literários) indicando que a proporção de palavras de natureza específica tem variado; e (iii) a média do comprimento dos caminhos mínimos da rede de co-ocorrência $\langle L \rangle$ (ambas as análises), que é baseada na conectividade de curto alcance e detecta a distância típica de palavras a todas as outras. A natureza distinta destas quantidades sugere um papel complementar na captura das propriedades de curto e de longo alcance dos textos.

O foco da análise de estilo desta tese se concentrou na avaliação de diferentes atributos e não na *maximização* da eficiência dos algoritmos de atribuição de autoria. Isto fica evidente quando comparamos, por exemplo, o melhor resultado de atribuição de autoria (65%) com as propostas tradicionais da literatura (90%). Enquanto a fonte do sucesso das abordagens tradicionais concentra-se na observação de que palavras funcionais são essenciais para detecção do estilo de autores (51, 88, 89), nesta tese focamos a análise nas palavras de conteúdo. Desta forma, os resultados podem ser entendidos como complementares à análise tradicional baseada em palavras funcionais. Mais especificamente, nossos resultados sugerem que especialmente medidas de intermitência, aglomeração e caminhos mínimos podem aperfeiçoar as técnicas de análise de estilo em textos, se usadas em combinação com outros atributos para extrair padrões (89).

Capítulo 7

Desambiguação de palavras

O termo ambiguidade refere-se à habilidade de se expressar de modo que pelo menos duas interpretações sejam possíveis quando o contexto não é especificado. Este fenômeno ocorre em várias situações e em particular nas expressões da linguagem humana (fala e escrita). Embora seres humanos sejam capazes de resolver ambiguidades quase que imediatamente, o processo de desambiguação automática em textos ainda representa um desafio aos linguistas e outros pesquisadores ligados ao estudo da linguagem. Grande parte da dificuldade enfrentada na área ocorre devido à escassez de grandes bases de conhecimento (i.e. conjunto de textos anotados, dicionários, redes semânticas, etc) de aplicação genérica. Infelizmente, as bases de conhecimento atuais são na sua maioria bastante limitadas e restritas a contextos específicos quando comparadas à base de conhecimento humana.

O problema de desambiguação automática de palavras é fundamental para o sucesso de aplicações de PLN que necessitam de tratamento semântico dos dados. Este é o caso, por exemplo, da tradução automática, pois a correta tradução depende sobretudo do sentido adequado ao contexto identificado para a palavra alvo. Ao ampliar a análise tradicional de textos baseada em estatísticas léxico-sintáticas, a tarefa de desambiguação também pode ser potencialmente útil para o desenvolvimento da Web Semântica (103) e tradução automática (22–24, 104) como ferramenta de análise mais profunda em uma vasta quantidade de textos.

O desenvolvimento das pesquisas em desambiguação iniciou-se na década de 1940, sendo concebida como tarefa fundamental para a tradução automática. Apesar da euforia inicial, ao longo do tempo foi notado que a complexidade da tarefa de identificar contextos representava um dos principais entraves ao desenvolvimento de tradutores automáticos. No início, o problema foi tratado com o auxílio de bases de conhecimento e informações estatísticas simples relacionadas ao uso de palavras e sentidos. Infelizmente, esta abordagem se mostrou limitada devido às restrições de tamanho das bases de conhecimento. Mais recentemente, métodos estatísticos robustos foram introduzidos e recursos léxicos têm sido utilizados (105). Atualmente, as estratégias de desambiguação automática podem ser

agrupadas em dois paradigmas:

1. **Paradigma profundo:** baseia-se no uso de bases de conhecimento, no intuito de simular a base de conhecimento humana. Na prática, esta estratégia é inferior ao paradigma superficial já que bases de conhecimentos atuais são relativamente limitadas a contextos específicos (105). Outro obstáculo aparece devido à necessidade da criação manual de recursos de conhecimento, já que a manutenção destas bases representa um esforço caro e demorado (106), que precisa ser repetido a cada mudança específica de contexto de desambiguação (105).
2. **Paradigma superficial:** neste método, a informação a respeito de contexto é obtida através da análise estatística da vizinhança da palavra ambígua. Por exemplo, se a palavra “tecido” aparece próxima à palavra ambígua “manga”, então o sentido associado à palavra é aquele relacionado ao vestimento. Por outro lado, se a palavra “árvore” aparece próxima à palavra ambígua, então o sentido mais provável é aquele relacionado à fruta. Abordagens deste tipo geralmente fornecem melhores resultados na prática, quando comparados com a abordagem tradicional (105).

Neste capítulo, adotamos uma abordagem estrutural que não utiliza bases de conhecimento. Neste tipo de análise, analisamos os padrões de conexão na estrutura topológica da rede de co-ocorrência de palavras. Ao contrário dos métodos tradicionais, o rótulo das palavras representando o contexto é eliminado e a análise é concentrada nos padrões de recorrência da estrutura da rede.

7.1 Técnicas tradicionais de desambiguação

Algoritmos tradicionais de desambiguação são baseados em técnicas de identificação de contexto pelo uso de vários atributos contextuais. Estes atributos podem ser classificados como:

- **Atributos locais:** caracterizam o contexto local da palavra alvo. Tipicamente, a análise é realizada nas características (p.e., a classe gramatical) de um pequeno número de palavras ao redor da palavra ambígua.
- **Atributos globais:** um contexto maior da palavra ambígua é estudado. A análise de contexto pode envolver uma sentença, um parágrafo ou ainda estruturas maiores.
- **Atributos sintáticos:** o relacionamento sintático entre a palavra ambígua e as outras palavras da sentença é usado como identificador de contexto.

- **Atributos semânticos:** a informação semântica das palavras sem ambiguidade (ou com ambiguidade resolvida) é usada como atributo.

Atributos locais, globais, sintáticos ou semânticos têm sido usados em muitos algoritmos por cerca de 70 anos. Um dos mais simples e famosos é o algoritmo *Lesk* (107), baseado na ideia de que as palavras identificadoras de contexto são semanticamente relacionadas ao sentido adequado à palavra ambígua. Mais especificamente, para cada um dos sentidos possíveis o algoritmo seleciona aquele que compartilha o maior número de palavras de sua definição no dicionário com o contexto ao redor da palavra ambígua. Extensões deste método e a descrição de vários outros podem ser encontradas nas Refs. (105, 107).

7.1.1 Abordagens tradicionais com *bag-of-words*

A técnica baseada na caracterização *bag-of-words* será exemplificada nesta seção com o classificador *Naive Bayes*. A ideia desta técnica é analisar a recorrência de contextos. Cada palavra de conteúdo ao redor da palavra ambígua pode contribuir para caracterizar um contexto. No caso da decisão bayesiana, a seguinte regra é utilizada:

Regra de decisão ótima Bayesiana : decida pelo sentido θ' se $P(\theta'|c) > P(\theta_k|c)$ para todo $\theta_k \neq \theta'$, onde $P(\theta_k|c)$ representa a probabilidade do sentido θ_k ocorrer no contexto c .

Usualmente, o valor de $P(\theta_k|c)$ não é conhecido, mas ele pode ser computado com o teorema de Bayes:

$$P(\theta_k|c) = \frac{P(c|\theta_k)}{P(c)}P(\theta_k). \quad (7.1)$$

Neste caso $P(\theta_k)$ é chamada de probabilidade a priori do sentido θ_k , isto é, a chance de o sentido θ_k ocorrer independentemente de qualquer contexto. A evidência sobre o contexto aparece no termo $P(c|\theta_k)$. No problema de discriminação de sentidos, $P(c)$ pode ser simplesmente ignorado por ser constante para todos os sentidos. Desta forma, o sentido mais provável pode ser obtido da seguinte maneira:

$$\begin{aligned} \theta' &= \arg \max_{\theta_k} P(\theta_k|c) \\ &= \arg \max_{\theta_k} \frac{P(c|\theta_k)}{P(c)} P(\theta_k) \\ &= \arg \max_{\theta_k} P(c|\theta_k) P(\theta_k) \\ &= \arg \max_{\theta_k} [\log P(c|\theta_k) + \log P(\theta_k)]. \end{aligned} \quad (7.2)$$

Note que o logaritmo foi tomado no último passo para simplificar a computação. Usando a simplificação do classificador *Naive Bayes* de independência dos atributos e considerando o

esquema de que palavras vizinhas j representam o contexto, i.e. $c = \{j | j \in c\}$, obtém-se:

Hipótese do Naive Bayes : $P(c|\theta_k) = P(\{j | j \in c\}|\theta_k) = \prod_{j \in c} P(j|\theta_k)$.

Note que, com esta suposição, toda a organização das palavras dentro do contexto considerado é ignorada. Além disso, devido à independência dos atributos, a presença de uma palavra na vizinhança indepente da presença de outra. Com estas considerações, o sentido correto θ' é escolhido como:

$$\theta' = \arg \max_{\theta_k} [\log P(\theta_k) + \sum_{j \in c} \log P(j|\theta_k)]. \quad (7.3)$$

7.1.2 Abordagens tradicionais com grafos

Os conceitos e metodologias de grafos também têm sido usados em algoritmos de desambiguação automática de sentido de palavras. Por exemplo, o algoritmo *HyperLex* (109) é baseado na construção de redes de co-ocorrência em nível de parágrafo, cujas arestas são estabelecidas entre palavras sempre que elas co-ocorrem no mesmo parágrafo. O peso das conexões é dado por $w_{ij} = 1 - \max\{P(i|j), P(j|i)\}$, onde $P(i|j) = N_{ij}/N_j$, $P(j|i) = N_{ij}/N_i$ e N_{ij} representa o número de vezes em que as palavras i e j apareceram no mesmo parágrafo. Após a construção do grafo, um passo iterativo é aplicado ao algoritmo de forma que o vértice com maior grau seja selecionado como *hub* e todos os seus vizinhos sejam não elegíveis como *hub* nas próximas interações. Este processo é finalizado quando o maior grau remanescente é menor que um dado limiar. O conjunto de *hubs* identificados na fase anterior, que são presumidos representar os possíveis sentidos, são conectados à palavra ambígua com peso $w_{ij} = 0$. Como próximo passo, a árvore geradora mínima é identificada a fim de ser usada como estrutura de discriminação de sentido para instâncias específicas da palavra alvo. Note que desta forma as palavras mais relacionadas ao *hubs* são usadas como representantes daquele sentido. A Figura 7.1 ilustra a árvore geradora mínima obtida para um exemplo fictício. Seja $\mathcal{J} = \{j_1, j_2, \dots, j_i, \dots, j_n\}$ o contexto em que a palavra ambígua está inserida. Um vetor $\vec{\text{scr}}$ é associado a cada $j_i \in \mathcal{J}$, de forma que o k -ésimo componente $\vec{\text{scr}}_k$ representa a contribuição do k -ésimo *hub* h_k :

$$\text{scr}_k = \begin{cases} 1/(1 + d(h_k, j_i)) & \text{se o } \textit{hub } h_k \text{ é um ancestral de } j_i \text{ na árvore} \\ 0 & \text{caso contrário.} \end{cases} \quad (7.4)$$

onde $d(h_k, j_i)$ representa a distância entre o *hub* h_k e o vértice j_i . Em seguida, todos os escores scr_k associados com todas as palavras j_i são somados e o *hub* que recebe o maior scr é escolhido como o sentido mais apropriado para a palavra ambígua.

Figura 7.1 – Árvore geradora mínima para a palavra ambígua *bar* em um texto fictício. A palavra ambígua é conectada com os *hubs* que representam possíveis sentidos assumidos da palavra. Através do algoritmo de árvore geradora mínima, os *hubs* são conectados às palavras mais representativas de cada sentido. Exemplo adaptado da Ref. (105).

Embora o algoritmo *HyperLex* e outros algoritmos de desambiguação utilizem a estrutura de co-ocorrência em grafo, ele é fortemente baseado nas características locais e semânticas do texto. A abordagem de redes complexas, descrita na próxima seção, emprega a estrutura da rede de co-ocorrência, que é baseada principalmente em fatores sintáticos (ver Capítulo 5).

7.2 Relacionamento entre topologia de redes e sentidos das palavras

A maioria das técnicas tradicionais é baseada na semântica para identificar o contexto de palavras ambíguas. O contexto estrutural identificado por meio de medidas topológicas em redes complexas ainda não foi explorado para a desambiguação. Em especial, é interessante verificar se sentidos distintos de uma palavra podem ser caracterizados topologicamente, sem auxílio de informação semântica ou de rótulo de palavras vizinhas. Para analisar o relacionamento entre estrutura da rede de co-ocorrência e semântica, 10 palavras ambíguas do Inglês foram estudadas no conjunto de livros da Tabela 7.1. A lista de palavras e sentidos assumidos são apresentados na Tabela 7.2. Nesta análise, cada uma das ocorrências da palavra ambígua é considerada como um vértice distinto na rede de modo que os diferentes sentidos possam ser analisados.

Tabela 7.1 – Lista de livros com data de publicação e respectivos autores para análise das palavras ambíguas da Tabela 7.2.

Livro	Autor
<i>Pride and Prejudice</i> (1813)	Jane Austen
<i>American Notes</i> (1842)	Charles Dickens
<i>Coral Reefs</i> (1842)	Charles Darwin
<i>A Tale of Two Cities</i> (1859)	Charles Dickens
<i>The Moonstone</i> (1868)	Wilkie Collins
<i>The Expression of the Emotions</i> (1872)	Charles Darwin
<i>A Pair of Blue Eyes</i> (1873)	Thomas Hardy
<i>Jude the Obscure</i> (1895)	Thomas Hardy
<i>Dracula's Guest</i> (1897)	Bram Stoker
<i>Uncle Bernac</i> (1897)	Arthur Conan Doyle
<i>The Tragedy of the Korosko</i> (1898)	Arthur Conan Doyle
<i>The Return of Sherlock Holmes</i> (1903)	Arthur Conan Doyle
<i>Tales of St. Austin's</i> (1903)	Pelham Grenville Wodehouse
<i>The Chronicles of Clovis</i> (1911)	Hector Hugh Munro
<i>A Changed Man; and Other Tales</i> (1913)	Thomas Hardy
<i>Beasts and Super Beasts</i> (1914)	Hector Hugh Munro
<i>The Wisdom of Father Brown</i> (1914)	Gilbert Keith Chesterton
<i>My Man Jeeves</i> (1919)	Pelham Grenville Wodehouse

Cada uma das ocorrências das palavras ambíguas foi caracterizada com as medidas locais de rede apresentadas no Capítulo 3. Para aperfeiçoar a caracterização topológica local das palavras, utilizamos também a definição hierárquica das medidas locais (ver Seção 2.5) e medidas simples de vizinhança como média e desvio do grau ($\langle k_n \rangle$ e Δk_n , respectivamente) e média e desvio do *strength* ($\langle s_n \rangle$ e Δs_n , respectivamente). O esquema de classificação foi montado de forma que os atributos sejam os valores das medidas extraídas da rede e a classe a ser inferida seja um dos sentidos apresentados na Tabela 7.2. O processo de reconhecimento de padrões para desambiguação está ilustrado na Figura 7.2.

Tabela 7.2 – Lista das 10 palavras ambíguas usadas na desambiguação automática de sentido. Os diferentes sentidos estão separados por ponto-e-vírgula (;).

Palavra	Sentido
<i>bear</i>	urso; suportar; lembrar (“ <i>bear in mind</i> ”)
<i>jam</i>	geleia; apertar
<i>just</i>	leal, justo, honrado; neste momento; somente, unicamente, só
<i>march</i>	Março (terceiro mês); marcha, caminhada
<i>rock</i>	pedra; balançar, chacoalhar; nome (p.e., Chris Rock)
<i>ring</i>	toque de sino, campainha; anel, círculo, argola
<i>save</i>	salvar, resgatar; poupar, economizar
<i>present</i>	no presente momento, agora; mostrar, apresentar
<i>close</i>	perto, próximo; fechar, finalizar, terminar
<i>note</i>	perceber; bilhete, aviso

Figura 7.2 – Exemplo de desambiguação para uma dada palavra ambígua com frequência $N_i = 7$. Os sentidos são identificados com as cores azul (3 instâncias) e vermelho (4 instâncias). Para simplificar, consideramos no exemplo apenas as medidas topológicas de grau k e coeficiente de aglomeração C . Note que enquanto o sentido em azul é caracterizado por altos valores de aglomeração e baixo grau, o sentido em vermelho possui alto grau e baixa aglomeração.

Para avaliar a qualidade da tarefa, a eficiência da classificação foi verificada em termos de sua significância estatística tomando como modelo nulo um classificador que presume não existir nenhuma relação entre a estrutura da palavra ambígua na rede e seu sentido no texto.

A Tabela 7.3 mostra na segunda e terceira colunas, respectivamente, a taxa de acerto e o valor correspondente do p -value p_{CN} relativo a um classificador que sempre associa o sentido mais provável (i.e., o mais frequente no banco de dados) à palavra ambígua. Uma taxa de acerto significativa foi obtida para 9 das 10 palavras consideradas. Isto significa que sentidos podem ser discriminados com precisão significativa usando apenas a informação topológica da rede e desconsiderando a semântica (ou o rótulo) das palavras que identificam o contexto. Um exemplo de *scatter plot* ilustrando a discriminação para a palavra “ring” é fornecido na Figura 7.3, onde cada eixo representa uma combinação linear das medidas fornecida pela técnica da análise de variáveis canônicas. Este forte relacionamento entre topologia e sentidos ocorre possivelmente porque as medidas de redes de co-ocorrência geralmente estão relacionadas a fatores sintáticos (ver Capítulo 5), que por sua vez estão atados a aspectos semânticos (108) através de um relacionamento não trivial.

Embora o foco do estudo deste projeto seja a análise de relacionamento entre estrutura topológica local e sentido de palavras, é interessante comparar os resultados a partir de métricas de redes complexas com os resultados obtidos da identificação de contexto baseada na recorrência de contexto. Esta abordagem é denominada aqui como abordagem tradicional. Para realizar este tipo de classificação, classificadores foram induzidos usando como atributo a lista de palavras que aparecerem pelo menos uma vez ao redor da palavra ambígua. Assim, o tamanho da janela ω de identificação de contexto (ω representa o

Tabela 7.3 – Resultados obtidos caracterizando palavras ambíguas com medidas estruturais de redes complexas. A taxa de acerto, o p -value p_{CN} (o modelo nulo é um classificador que associa sempre o sentido majoritário ao sentido da instância) e os melhores classificadores estão ilustrados para todas as palavras. O menor conjunto de medidas com melhor resultado em cada caso é mostrado na última coluna. Com exceção da palavra “close”, todas as classificações se mostraram estatisticamente significativas ($p_{CN} < 5,0 \times 10^{-2}$). Note que mesmo usando muitas medidas para caracterizar o contexto estrutural, os melhores classificadores foram obtidos com apenas 5 no máximo.

Palavra	Acerto	p_{CN}	Classif.	Medidas
<i>save</i>	87,64 %	$1,6 \times 10^{-3}$	κ -NN	$\langle s_n \rangle$ e L
<i>note</i>	84,53 %	$4,7 \times 10^{-2}$	κ -NN	$C^{(3)}$, $s^{(3)}$ e Δs_n
<i>march</i>	86,95 %	$1,9 \times 10^{-3}$	κ -NN	$C^{(3)}$, $\langle k_n \rangle$ e B
<i>present</i>	71,14 %	$2,2 \times 10^{-2}$	κ -NN	$s^{(4)}$, $\langle k_n \rangle$ e Δs_n
<i>jam</i>	100,0 %	$6,0 \times 10^{-3}$	κ -NN	$C^{(4)}$, $s^{(4)}$ e Δk_n
<i>ring</i>	84,61 %	$2,8 \times 10^{-4}$	κ -NN	$C^{(4)}$, $k^{(4)}$, $s^{(4)}$, Δk_n e L
<i>just</i>	51,28 %	$1,6 \times 10^{-2}$	κ -NN	$C^{(2)}$ e $C^{(3)}$
<i>bear</i>	61,95 %	$8,0 \times 10^{-3}$	κ -NN	$C^{(4)}$, $k^{(3)}$ e $s^{(4)}$
<i>rock</i>	79,30 %	$9,7 \times 10^{-9}$	C4.5	$k^{(2)}$, $s^{(2)}$, $k^{(3)}$, $k^{(4)}$ e $s^{(4)}$
<i>close</i>	72,20 %	$8,0 \times 10^{-2}$	κ -NN	$C^{(2)}$, $s^{(2)}$, $\langle s_n \rangle$ e Δs_n

Figura 7.3 – Análise de variáveis canônicas para a palavra “ring”. Os sentidos considerados são (i) som de um sino e (ii) anel. Enquanto o sentido (i) é mais regular, sendo caracterizado pela baixa dispersão, o sentido (ii) tende a ser mais heterogêneo do ponto de vista de sua estrutura topológica.

número de palavras) representa um parâmetro do classificador. Foram considerados aqui os valores $\omega = \{5, 20, 50\}$. A Tabela 7.4 mostra os classificadores com as maiores taxas de acerto (com menor p -value). Para as primeiras 5 palavras a abordagem estrutural de redes apresenta melhor eficiência que o método tradicional. Isto significa que os padrões de recorrência de topologia local podem ser ainda mais relevantes que os padrões de recorrência semântica das palavras. Apesar destes resultados, o uso somente da abordagem estrutural não é aconselhável já que as medidas de redes complexas são estatisticamente confiáveis somente quando computadas em textos longos. Em princípio, as técnicas poderiam ser combinadas, já que a análise estrutural provavelmente fornece informação complementar à análise de rótulo das palavras do contexto. Com relação aos métodos, algumas observações merecem destaque. Dentre os algoritmos convencionais, a abordagem de redes é mais eficiente com o algoritmo κ -NN, enquanto que o algoritmo *Naive Bayes* é mais eficiente quando usado com a estratégia tradicional. Estas diferenças provavelmente ocorrem devido ao número distinto de atributos disponíveis em cada uma das palavras^a. Finalmente, em 6 das 10 palavras estudadas, a análise tradicional com 5 vizinhos apresentou maior taxa de acerto do que a mesma análise realizada com mais vizinhos.

Tabela 7.4 – Resultados da caracterização de palavras ambíguas usando a abordagem tradicional (segunda coluna) e a abordagem baseada em métricas de redes complexas (terceira coluna). O p -value (p_{TR} para a abordagem tradicional e p_{CN} para a abordagem com redes) foi calculado considerando como modelo nulo um classificador baseado no sentido majoritário, i.e., um classificador que sempre associa o sentido mais frequente. Os melhores classificadores ilustrados na quarta coluna referem-se à abordagem tradicional.

Palavra	p_{TR}	p_{CN}	Classif.
<i>save</i>	$6,2 \times 10^{-1}$	$1,6 \times 10^{-3}$	κ -NN
<i>note</i>	$3,8 \times 10^{-1}$	$4,7 \times 10^{-2}$	κ -NN
<i>march</i>	$1,4 \times 10^{-2}$	$1,9 \times 10^{-3}$	κ -NN
<i>present</i>	$7,6 \times 10^{-1}$	$2,2 \times 10^{-2}$	<i>Bayes</i>
<i>jam</i>	$4,0 \times 10^{-1}$	$6,0 \times 10^{-3}$	<i>Bayes</i>
<i>ring</i>	$7,3 \times 10^{-6}$	$2,8 \times 10^{-4}$	<i>Bayes</i>
<i>just</i>	$1,8 \times 10^{-4}$	$1,6 \times 10^{-2}$	<i>Bayes</i>
<i>bear</i>	$3,3 \times 10^{-5}$	$8,0 \times 10^{-3}$	C4.5
<i>rock</i>	$4,4 \times 10^{-16}$	$9,7 \times 10^{-9}$	<i>Bayes</i>
<i>close</i>	$2,3 \times 10^{-11}$	$8,0 \times 10^{-2}$	<i>Naive</i>

7.3 Análise de relevância dos atributos

A contribuição de cada medida para a discriminação de sentidos pode ser inicialmente examinada observando o subconjunto de medidas que gerou os melhores classificadores

^a Enquanto na abordagem estrutural o número de atributos é fixado pelo número de medidas, na análise tradicional o número de atributos varia de acordo com a frequência da palavra ambígua analisada.

Figura 7.4 – *Scatter plot* para a palavra “save”. Os sentidos utilizados foram (i) economizar e (ii) resgatar, salvar. Enquanto (i) é caracterizado por altos valores de L e baixos valores de $\langle s_n \rangle$; (ii) usualmente assume baixos valores de L .

(última coluna da Tabela 7.3). Note que há uma predominância de medidas hierárquicas, sugerindo que contextos mais profundos de análise na rede melhoram a caracterização do sentido das palavras. Embora várias medidas tenham sido utilizadas, o melhor desempenho para a tarefa foi obtido com 5 ou menos atributos. Para algumas palavras, foi possível alcançar a taxa de acerto ótima com apenas 2 atributos. Este é o caso da palavra “save”, cuja distribuição de sentidos para as medidas $\langle s_n \rangle$ e L é ilustrada na Figura 7.4. Quantitativamente, a relevância das medidas foi avaliada com o método KL; que avalia atributos isoladamente e com o método AUC, que considera a interação de atributos. O *ranking* obtido, ilustrado na Tabela 10.4, indica que a relevância das métricas varia de palavra para palavra. Além disso, uma métrica pode ser considerada relevante quando analisada isoladamente mas pouco relevante quando combinada com outras métricas. De acordo com o método KL, as métricas que apareceram mais vezes com alta relevância (entre as três melhores) foram $C^{(3)}$ e L . Segundo o método AUC, o coeficiente de aglomeração calculado em altos níveis ($C^{(4)}$) também é um atributo relevante, assim como o *strength* em alto nível ($s^{(3)}$ e $s^{(4)}$). Estes resultados sugerem que geralmente a relevância das métricas para a discriminação de sentido é dependente do *strength* dos vizinhos “virtuais” (i.e., vizinhos em níveis superiores). Diferentemente da análise isolada, na análise combinada o comprimento médio dos caminhos mínimos L não foi relevante.

Tabela 7.5 – *Rankings* de relevância de métricas obtidas com o método Kullback-Leibler (KL) e com o método AUC. De acordo com o método KL, as métricas que apareceram mais vezes com alta relevância (entre as três melhores) foram $C^{(3)}$ e L . Segundo o método AUC, o coeficiente de aglomeração calculado em altos níveis ($C^{(4)}$) também é atributo relevante, assim como o *strength* em alto nível ($s^{(3)}$ e $s^{(4)}$).

Método Palavra	Ranking KL			Ranking AUC				
	1	2	3	1	2	3	4	5
<i>bear</i>	$C^{(2)}$	$\langle s_n \rangle$	$\langle k_n \rangle$	B	$s^{(3)}$	$k^{(2)}$	$k^{(4)}$	Δk_n
<i>jam</i>	$C^{(4)}$	$C^{(3)}$	$\langle s_n \rangle$	$s^{(3)}$	$C^{(4)}$	$C^{(2)}$	$s^{(4)}$	$k^{(4)}$
<i>just</i>	$C^{(4)}$	$C^{(3)}$	–	$k^{(3)}$	$s^{(4)}$	$C^{(4)}$	$s^{(3)}$	$k^{(4)}$
<i>march</i>	–	–	–	B	$C^{(4)}$	$k^{(2)}$	$s^{(4)}$	$k^{(4)}$
<i>ring</i>	L	–	–	B	$C^{(4)}$	$k^{(4)}$	$s^{(3)}$	L
<i>present</i>	–	–	–	$\langle s_n \rangle$	$C^{(4)}$	$\langle k_n \rangle$	$s^{(4)}$	$k^{(3)}$
<i>close</i>	–	–	–	$C^{(3)}$	L	$s^{(4)}$	$\langle k_n \rangle$	$k^{(4)}$
<i>note</i>	$C^{(3)}$	$s^{(4)}$	$s^{(2)}$	$s^{(3)}$	B	Δk_n	Δs_n	$s^{(2)}$
<i>save</i>	L	–	–	Δs_n	$s^{(2)}$	L	$k^{(3)}$	$C^{(4)}$
<i>rock</i>	L	$s^{(2)}$	$k^{(2)}$	$\langle k_n \rangle$	$k^{(2)}$	$s^{(2)}$	$s^{(4)}$	$s^{(3)}$

7.4 Desambiguação baseada na classificação híbrida

Conforme relatado nos resultados da Seção 7.3, o uso de redes complexas pode fornecer informação de contexto estrutural útil para a desambiguação. Esta mesma ideia de empregar a estrutura em substituição à semântica pode ser aplicada no nível de reconhecimento de padrões. Após a extração das medidas referentes à estrutura das palavras ambíguas, os dados são apresentados em uma matriz atributo-valor (61, 64), que é usada como entrada para as técnicas de reconhecimento de padrão. Algoritmos tradicionais como κ -NN são baseados na semântica destes dados. Ou seja, se dois objetos estão próximos no espaço de atributos, então provavelmente pertencerão a um mesmo sentido. Mas os dados podem também ser estudados a partir de sua organização estrutural no espaço de atributos, onde sentidos de palavras são conectados de acordo com a similaridade das métricas extraídas das redes de co-ocorrência. Desta forma, duas palavras são supostas como tendo o mesmo sentido caso suas organizações estruturais no espaço de atributos sejam similares.

A análise de reconhecimento de padrões na estrutura organizacional das instâncias de palavras é feita com o algoritmo de reconhecimento de padrões de híbrido, descrito no Capítulo 4, Seção 4.1.4. Para exemplificar o funcionamento deste algoritmo, considere o exemplo da Figura 7.5. Duas classes de sentidos diferentes são evidentes: a classe laranja (quadrado) e a classe azul (círculo). Durante a fase de classificação, os parâmetros para construir a rede foram $\kappa = 3$ e $\epsilon = 0,02$ (ver Seção 4.1.4). No exemplo, a classe laranja apresenta um padrão bem definido: uma estrutura piramidal, enquanto que a classe azul não apresenta nenhum padrão claro de formação. Suponha que o objetivo seja classificar a classe verde (triângulo) usando apenas informação do conjunto de treinamento. As

Figura 7.5 – Exemplo de classificação com o algoritmo híbrido que combina as técnicas de baixo e alto nível. O comportamento das fronteiras de decisão varia conforme o valor de λ . As fronteiras de decisão ilustradas correspondem a (a) $\lambda = 0,00$; (b) $\lambda = 0,50$; and (c) $\lambda = 0,80$.

Figuras 7.5(a), 7.5(b) e 7.5(c) exibem as fronteiras de decisão das duas classes para $\lambda = 0$, $\lambda = 0,5$ e $\lambda = 0,8$, respectivamente. Quando $\lambda = 0$, apenas o classificador tradicional de baixo nível é utilizado no classificador híbrido. Neste caso, a instância de teste não é corretamente classificada. Quando $\lambda = 0,5$, o classificador de baixo nível e o baseado na estrutura dos dados são utilizados com igual peso, levando a fronteira de decisão a se aproximar da classe azul devido ao forte padrão da classe laranja. Isto acontece porque conforme o valor de λ aumenta, a classe mais estruturada tende a deter maior poder de decisão e por este motivo a área de decisão correspondente à outra classe é diminuída. Quando o valor de λ é aumentado ainda mais, os padrões estruturais da classe laranja se tornam tão fortes a ponto de a fronteira de decisão se aproximar bastante da classe azul.

A verificação de eficiência de discriminação de sentidos usando a abordagem topológica na caracterização das palavras e a abordagem topológica no espaço de atributos é mostrada na Tabela 7.6. Também é mostrado o valor ótimo do parâmetro λ , que fornece o peso atribuído ao fator topológico na classificação. Note que quando λ é selecionado de tal forma a maximizar a qualidade classificação, a eficiência da discriminação é aumentada em comparação com a eficiência dos classificadores de baixo nível (κ -NN, C4.5 e *Naive Bayes*). É fundamental observar que $\lambda \neq 0$ na maioria das palavras estudadas, sugerindo que padrões de organização estrutural emergem também no espaço de atributos. Quando os atributos baseados na recorrência de palavras vizinhas são aplicados (ver seção anterior), também há melhoria no desempenho de discriminação, como mostra a Tabela 7.7. Estes resultados sugerem que a organização estrutural das palavras não é útil somente quando examinada em termos de medidas de redes de adjacência, mas também quando consideradas no espaço de atributos.

A técnica de alto nível também foi aplicada em combinação com a caracterização de caminha do turista (38). Como no caso de caracterização com medidas tradicionais de redes complexas, dois paradigmas foram empregados na caracterização de sentidos: o tradicional baseado na recorrência de vizinhos (ver Tabela 7.8) e o baseado na estrutura das redes de adjacência (ver Tabela 7.9). Em ambos os casos, podemos ver que os melhores resultados ocorrem quando $\lambda > 0$, o que novamente implica que os padrões estruturais no espaço de atributos emergem tanto na caracterização topológica como na caracterização semântica. Em particular, a dependência mais forte (com menores p -values) entre sentidos

Tabela 7.6 – Abordagem estrutural para discriminar sentidos de palavras ambíguas. Instâncias de palavras ambíguas são caracterizadas com métricas topológicas locais de redes complexas e a discriminação foi feita com classificadores tradicionais de baixo nível (κ -NN, C4.5 e *Naive Bayes*) e com o classificador híbrido. Note que a taxa de acerto é otimizada quando padrões estruturais de organização no espaço de atributos são considerados na classificação, já que o parâmetro λ^* que otimiza a discriminação de sentidos é na maioria das vezes $\lambda^* \neq 0$.

Palavra	Baixo nível		Híbrido	
	Classif.	Acerto	Acerto	λ^*
<i>save</i>	κ -NN	87,6 %	90,2 %	0,10
<i>save</i>	C4.5	79,8 %	84,2 %	0,25
<i>save</i>	<i>Bayes</i>	83,1 %	86,5 %	0,15
<i>note</i>	κ -NN	84,5 %	85,3 %	0,10
<i>note</i>	C4.5	78,4 %	80,3 %	0,15
<i>note</i>	<i>Bayes</i>	78,4 %	80,3 %	0,15
<i>march</i>	κ -NN	87,0 %	88,4 %	0,10
<i>march</i>	C4.5	60,9 %	71,9 %	0,35
<i>march</i>	<i>Bayes</i>	73,9 %	76,2 %	0,20
<i>present</i>	κ -NN	71,1 %	71,1 %	0,00
<i>present</i>	C4.5	64,7 %	65,5 %	0,05
<i>present</i>	<i>Bayes</i>	73,9 %	73,9 %	0,00
<i>jam</i>	κ -NN	100,0 %	100,0 %	0,00
<i>jam</i>	C4.5	80,0 %	84,4 %	0,20
<i>jam</i>	<i>Bayes</i>	90,0 %	91,9 %	0,10
<i>ring</i>	κ -NN	84,6 %	89,0 %	0,20
<i>ring</i>	C4.5	71,4 %	74,6 %	0,20
<i>ring</i>	<i>Bayes</i>	69,2 %	72,4 %	0,20
<i>just</i>	κ -NN	51,3 %	57,8 %	0,35
<i>just</i>	C4.5	47,0 %	52,0 %	0,30
<i>just</i>	<i>Bayes</i>	46,2 %	50,3 %	0,30
<i>bear</i>	κ -NN	62,0 %	66,1 %	0,20
<i>bear</i>	C4.5	62,0 %	67,4 %	0,25
<i>bear</i>	<i>Bayes</i>	60,9 %	65,5 %	0,20
<i>rock</i>	κ -NN	78,4 %	84,5 %	0,20
<i>rock</i>	C4.5	79,3 %	86,0 %	0,30
<i>rock</i>	<i>Bayes</i>	70,3 %	74,9 %	0,15
<i>close</i>	κ -NN	72,2 %	80,6 %	0,25
<i>close</i>	C4.5	68,7 %	79,9 %	0,35
<i>close</i>	<i>Bayes</i>	68,7 %	77,7 %	0,25

Tabela 7.7 – Abordagem baseada na recorrência de palavras vizinhas para discriminar sentidos de palavras ambíguas. As instâncias das palavras ambíguas foram caracterizadas de acordo com a frequência das $\omega = 5$ palavras vizinhas. O acerto representa a porcentagem de instâncias corretamente classificadas. Note que a eficiência da abordagem híbrida nunca é inferior à abordagem de baixo nível.

Palavra	Baixo nível		Híbrido	
	Classif.	Acerto	Acerto	λ^*
<i>save</i>	κ -NN	79,5 %	84,7 %	0,25
<i>save</i>	C4.5	78,9 %	84,1 %	0,25
<i>save</i>	Bayes	76,6 %	81,9 %	0,25
<i>note</i>	κ -NN	82,6 %	86,4 %	0,20
<i>note</i>	C4.5	79,5 %	83,1 %	0,15
<i>note</i>	Bayes	78,3 %	82,1 %	0,20
<i>march</i>	κ -NN	82,8 %	89,7 %	0,30
<i>march</i>	C4.5	82,8 %	87,4 %	0,30
<i>march</i>	Bayes	62,5 %	72,9 %	0,35
<i>present</i>	κ -NN	62,9 %	69,4 %	0,20
<i>present</i>	C4.5	57,0 %	61,4 %	0,15
<i>present</i>	Bayes	60,2 %	65,3 %	0,20
<i>jam</i>	κ -NN	76,5 %	87,3 %	0,35
<i>jam</i>	C4.5	76,5 %	89,5 %	0,40
<i>jam</i>	Bayes	82,4 %	90,1 %	0,30
<i>ring</i>	κ -NN	80,7 %	81,9 %	0,05
<i>ring</i>	C4.5	84,4 %	84,4 %	0,00
<i>ring</i>	Bayes	83,5 %	83,5 %	0,05
<i>just</i>	κ -NN	47,4 %	56,3 %	0,25
<i>just</i>	C4.5	38,0 %	45,0 %	0,30
<i>just</i>	Bayes	52,6 %	57,2 %	0,25
<i>bear</i>	κ -NN	48,0 %	52,9 %	0,25
<i>bear</i>	C4.5	62,0 %	62,0 %	0,00
<i>bear</i>	Bayes	48,0 %	53,5 %	0,25
<i>rock</i>	κ -NN	69,7 %	75,7 %	0,30
<i>rock</i>	C4.5	68,5 %	73,1 %	0,20
<i>rock</i>	Bayes	77,5 %	77,5 %	0,00
<i>close</i>	κ -NN	64,8 %	69,0 %	0,20
<i>close</i>	C4.5	71,9 %	74,3 %	0,15
<i>close</i>	Bayes	75,3 %	78,8 %	0,15

Figura 7.6 – Análise de variabilidade do comprimento do transiente e ciclos para dois sentidos da palavra *close*. Cada cor representa um sentido distinto. Ambas as quantidades são úteis para discriminar sentidos pois sentidos distintos são caracterizados por diferentes padrões de complexidade.

e organização estrutural na rede de adjacência foi observada para as palavras *rock* e *close*. Na abordagem semântica, a maior dependência foi encontrada observada para as palavras *rock* e *ring*.

É possível explicar a razão pela qual a caracterização com a caminhada do turista apresenta desempenho igual ou superior à caracterização de medidas topológicas tradicionais de redes complexas. A caminhada do turista naturalmente captura a informação topológica local até global simplesmente variando o comprimento da memória do turista. Conforme incrementamos a memória do turista, ela gradualmente captura mais informação global da rede, já que o comprimento do ciclo e o transiente tendem a variar mais do que quando pequenos valores de μ são selecionados. Diferentemente, cada medida tradicional de rede complexa captura de uma maneira estática uma característica especificada da rede. Por exemplo, o grau captura estritamente organização local de cada vértice; o coeficiente de aglomeração captura relações entre vizinhos, e a assortatividade captura uma informação global de cada componente da rede. Portanto, a idéia de capturar gradualmente as características da rede com o ciclo c e com a memória μ também parece ser útil e dependente dos sentidos. Interessantemente, sentidos diferentes são caracterizados por diferentes tempos necessários para alcançar o estado estacionário, i.e. o estado onde não existe variação em resposta ao aumento de μ (ver exemplos nas Figuras 7.6 e 7.7). Estes resultados sugerem que a complexidade de cada componente representando um sentido pode ser capturada com a introdução de comprimentos de memória mais longos, os quais são úteis para aperfeiçoar a identificação de sentidos de palavras polissêmicas.

Em suma, quando o sentido das palavras é analisado do ponto de vista de redes complexas, padrões (de organização estrutural no espaço de atributos) ocultos aos classificadores tradicionais são identificados, fazendo com que a classificação estrutural e até mesmo a

Tabela 7.8 – Abordagem semântica para discriminação de sentido de palavras ambíguas. Os sentidos foram caracterizados de acordo com a frequência de aparecimento das 5 palavras vizinhas ambíguas e a discriminação de sentidos foi realizada com classificadores de baixo (κ -NN, C4.5 e Bayes) e alto nível. $\Delta\%$ representa o ganho obtido a técnica híbrida em uma avaliação realizada com a técnica de validação cruzada. O p -value corresponde à probabilidade de a taxa de acerto obtida com a técnica híbrida ser obtida com um classificador aleatório. Note que em todos os casos a técnica de alto nível é melhor ou igual que a técnica tradicional de baixo nível.

Palavra	Técnica	$\Delta\%$	p -value	λ^*
<i>save</i>	κ -NN	3.8%	1.8×10^{-1}	0.20
<i>save</i>	C4.5	6.9%	7.3×10^{-2}	0.30
<i>save</i>	Bayes	6.7%	1.8×10^{-1}	0.35
<i>note</i>	κ -NN	5.3%	7.6×10^{-3}	0.25
<i>note</i>	C4.5	5.5%	5.3×10^{-2}	0.25
<i>note</i>	Bayes	6.7%	5.3×10^{-2}	0.30
<i>march</i>	κ -NN	7.7%	3.3×10^{-4}	0.35
<i>march</i>	C4.5	7.7%	3.3×10^{-4}	0.30
<i>march</i>	Bayes	8.9%	1.5×10^{-1}	0.30
<i>present</i>	κ -NN	5.9%	1.0×10^{-2}	0.20
<i>present</i>	C4.5	3.1%	8.9×10^{-1}	0.10
<i>present</i>	Bayes	4.2%	5.3×10^{-1}	0.15
<i>jam</i>	κ -NN	13.5%	4.6×10^{-2}	0.40
<i>jam</i>	C4.5	13%	4.6×10^{-2}	0.45
<i>jam</i>	Bayes	0.0%	4.6×10^{-2}	0.00
<i>ring</i>	κ -NN	5.7%	3.4×10^{-5}	0.25
<i>ring</i>	C4.5	5.4%	7.0×10^{-7}	0.20
<i>ring</i>	Bayes	3.5%	3.4×10^{-5}	0.15
<i>just</i>	κ -NN	11.7%	6.8×10^{-5}	0.50
<i>just</i>	C4.5	10.5%	5.5×10^{-2}	0.40
<i>just</i>	Bayes	10.7%	1.0×10^{-6}	0.45
<i>bear</i>	κ -NN	5.7%	2.3×10^{-1}	0.20
<i>bear</i>	C4.5	3.4%	1.1×10^{-3}	0.15
<i>bear</i>	Bayes	6.9%	1.3×10^{-1}	0.30
<i>rock</i>	κ -NN	0.0%	3.6×10^{-4}	0.00
<i>rock</i>	C4.5	1.5%	1.7×10^{-4}	0.10
<i>rock</i>	Bayes	0.0%	1.0×10^{-7}	0.00
<i>close</i>	κ -NN	5.7%	2.1×10^{-1}	0.25
<i>close</i>	C4.5	4.2%	1.7×10^{-3}	0.15
<i>close</i>	Bayes	4.7%	3.7×10^{-6}	0.15

Tabela 7.9 – Abordagem topológica para discriminação de sentido de palavras ambíguas. Os sentidos foram caracterizados com medidas topológicas de redes complexas e a discriminação de sentidos foi realizada com classificadores de baixo (κ -NN, C4.5 e Bayes) e alto nível. $\Delta\%$ representa o ganho obtido a técnica híbrida em uma avaliação realizada com a técnica de validação cruzada. O p -value corresponde à probabilidade de a taxa de acerto obtida com a técnica híbrida ser obtida com um classificador aleatório. Note que em todos os casos a técnica de alto nível é melhor ou igual que a técnica tradicional de baixo nível.

Palavra	Técnica	$\Delta\%$	p -value	λ^*
<i>save</i>	κ -NN	3.3%	1.6×10^{-3}	0.15
<i>save</i>	C4.5	6.2%	4.1×10^{-2}	0.25
<i>save</i>	Bayes	2.9%	4.1×10^{-2}	0.10
<i>note</i>	κ -NN	1.6%	1.1×10^{-1}	0.10
<i>note</i>	C4.5	3.0%	5.6×10^{-1}	0.10
<i>note</i>	Bayes	3.0%	5.6×10^{-1}	0.10
<i>march</i>	κ -NN	2.3%	3.3×10^{-4}	0.15
<i>march</i>	C4.5	12.5%	6.8×10^{-2}	0.40
<i>march</i>	Bayes	6.0%	2.6×10^{-2}	0.25
<i>present</i>	κ -NN	2.5%	2.7×10^{-3}	0.30
<i>present</i>	C4.5	3.1%	1.7×10^{-1}	0.25
<i>present</i>	Bayes	3.9%	2.8×10^{-5}	0.40
<i>jam</i>	κ -NN	0.0%	6.0×10^{-3}	0.00
<i>jam</i>	C4.5	5.6%	4.6×10^{-2}	0.20
<i>jam</i>	Bayes	2.8%	4.6×10^{-2}	0.10
<i>ring</i>	κ -NN	7.8%	3.3×10^{-8}	0.35
<i>ring</i>	C4.5	6.9%	2.5×10^{-1}	0.30
<i>ring</i>	Bayes	6.5%	6.9×10^{-2}	0.25
<i>just</i>	κ -NN	8.7%	3.1×10^{-5}	0.45
<i>just</i>	C4.5	8.8%	1.0×10^{-3}	0.35
<i>just</i>	Bayes	7.1%	5.9×10^{-3}	0.30
<i>bear</i>	κ -NN	5.2%	2.6×10^{-4}	0.25
<i>bear</i>	C4.5	6.9%	1.1×10^{-4}	0.30
<i>bear</i>	Bayes	6.1%	2.6×10^{-4}	0.25
<i>rock</i>	κ -NN	3.7%	1.9×10^{-10}	0.15
<i>rock</i>	C4.5	5.2%	2.3×10^{-12}	0.15
<i>rock</i>	Bayes	3.0%	1.5×10^{-5}	0.10
<i>close</i>	κ -NN	9.6%	6.0×10^{-8}	0.35
<i>close</i>	C4.5	11.4%	1.9×10^{-6}	0.40
<i>close</i>	Bayes	10.7%	6.8×10^{-6}	0.35

Figura 7.7 – Análise de variabilidade do comprimento do transiente e ciclos para dois sentidos da palavra *present*. Cada cor representa um sentido distinto. Ambas as quantidades são úteis para discriminar sentidos pois sentidos distintos são caracterizados por diferentes padrões de complexidade.

classificação baseada na recorrência de palavras vizinhas sejam aperfeiçoadas.

7.5 Conclusões

Neste capítulo avaliamos a aplicabilidade do modelo de redes complexas para a tarefa de desambiguação em textos. Caracterizando a estrutura local dos vértices representando palavras ambíguas, observou-se que os diferentes sentidos afetam a organização estrutural das redes de co-ocorrência. Ressalte-se que a discriminação para algumas palavras usando a organização estrutural foi melhor que aquela alcançada com a análise de recorrência de palavras vizinhas. Com relação à relevância das métricas, a melhor medida varia de palavra para palavra. De maneira geral, as métricas hierárquicas se mostraram úteis para a tarefa.

Além de aplicar os conceitos e metodologias de redes complexas para modelar o relacionamento entre palavras por relações de adjacência no texto, consideramos também o relacionamento entre palavras por relações de similaridade estrutural no espaço de atributos. Para modelar a relação entre sentido neste novo espaço, usamos um classificador híbrido combinando classificadores tradicionais de baixo nível com aqueles baseados em padrões estruturais. Os resultados evidenciaram que a inclusão do termo topológico foi responsável por aperfeiçoar a capacidade de discriminação. Este resultado sugere que as palavras com o mesmo sentido apresentam os mesmos padrões organizacionais não apenas no nível textual mas também no espaço de atributos. Isto reforça a premissa de que estas redes não são totalmente desorganizadas, já que um dado sentido tende a formar padrões

que o descrevem unicamente.

Capítulo 8

Consistência de palavras

Desde os primórdios da humanidade, a habilidade de comunicação tem desempenhado um papel fundamental para a preservação da vida e manutenção das relações sociais. A escrita, uma das principais manifestações da comunicação, cuja invenção remonta a 3200 a.C., também se estabeleceu como uma das principais habilidades desenvolvidas pela humanidade. Entre as principais vantagens da língua escrita em relação à falada estão a capacidade de portabilidade e permanência, que garantem que os pensamentos, idéias, histórias e fatos sejam preservadas através dos séculos. Apesar desta ubiquidade, a habilidade de se escrever um texto conciso, coerente e claro ainda não pode ser considerada uma tarefa trivial ou comum. Mesmo após a aquisição da linguagem, a construção de um texto preciso e bem concatenado exige, além de um pensamento organizado, a habilidade de uso recursos lingüísticos expressivos aliada à interpretação analítica da realidade.

Além das dificuldades impostas pelos mecanismos da gramática na língua escrita (110), existe ainda um fator relacionado ao nível semântico dos detalhes necessários para recriar a idéia original do autor. Como não é possível especificar todos os detalhes em um texto finito, o autor deve sempre ajustar o nível desejado de detalhamento. Assim, quando poucos detalhes são fornecidos o leitor deve preencher os espaços em branco usando seu próprio conhecimento e experiência sobre o mundo. Por outro lado, em um texto excessivamente pormenorizado, pouco espaço é deixado para inferências, o que torna a leitura mais objetiva. Esta dicotomia entre objetividade e subjetividade tem sido explorada de diversas maneiras em diferentes gêneros de escrita. Enquanto textos científicos, jornais, revistas e relatórios tendem a usar uma abordagem mais objetiva, textos literários e artísticos preferem uma apresentação mais subjetiva. Em ambos os casos, o grau de objetividade (ou subjetividade) varia de acordo com o tamanho do texto (textos longos tendem a ser mais detalhados), e o número de palavras descritivas (um texto com muitos adjetivos, por exemplo, tende a ser muito detalhado). O contexto induzido pelas palavras parece ser igualmente importante no nível de detalhamento (111, 112), uma vez que as palavras que induzem contextos restritos tendem a limitar a capacidade de extrapolar idéias, tornando o texto mais objetivo.

Tomando a relação entre indução e objetividade como motivação (113), neste capítulo abordamos o problema de quantificar/classificar a consistência inerente das palavras, isto é, o grau de preservação de suas vizinhança (contexto), no intuito de entender as razões pelas quais uma palavra é usada de uma forma mais ou menos consistente. Usamos o termo consistência porque as palavras cujo contexto é preservado em suas inúmeras ocorrências tenderá a ser utilizada consistentemente por diferentes autores, inferindo sempre um mesmo contexto. Usando os conceitos e metodologias de redes complexas para analisar a relação entre os conceitos, desenvolvemos uma série de índices para medir a consistência. Estes índices são baseados na ideia de que, se uma palavra induz um conjunto limitado de contextos, então a vizinhança dessa palavra será mantida mesmo em textos escritos por diferentes autores. Esta hipótese é de fato razoável, uma vez que é sabido que as palavras sintaticamente relacionados também tendem a ser semanticamente relacionadas. Neste capítulo estudamos ainda a distribuição da medida proposta e sua correlação com outros fatores linguísticos das palavras, tais como frequência na língua e nível de ambiguidade. A aplicabilidade dos índices de consistência em tarefas reais será ilustrada no problema de reconhecimento de autoria.

8.1 Rede dos autores

Os contextos distintos na qual cada palavra é utilizada foram investigados com um banco de dados obtidos no projeto Gutenberg. Este banco de dados é apresentado na Tabela 8.1. Embora o número de livros utilizados para cada autor seja diferente, o tamanho do *corpus* para cada autor foi mantido constante (180.500 *tokens*). Dessa forma, a diferença do tamanho do *corpora* pouco influencia a análise de consistência das palavras.

Após a criação de uma rede para cada livro de cada autor, as redes de livros de um mesmo autor foram mescladas para a obtenção da rede do autor, que reflete as associações de palavras geradas por aquele autor. Em outras palavras, se um dado vértice aparece em uma das redes de livros do autor, então ela aparecerá na rede do autor correspondente. Analogamente, dois vértices são conectados na rede de autor se ambas aparecem conectadas em pelo menos uma rede de livro do autor. O processo de derivação das redes de autores é ilustrado na Figura 8.1.

8.2 Índices de consistência

Nesta seção, descrevemos os índices propostos para medir a consistência $\mathbb{C}$ de uso das palavras. Como o cálculo da consistência necessita que as palavras ocorram em todas as

Tabela 8.1 – Banco de dados empregado nos experimentos de análise de consistência de palavras.

Livro	Autor
<i>Uncle Bernac - A Memory of the Empire</i>	Arthur Conan Doyle
<i>The Tragedy of the Korosko</i>	Arthur Conan Doyle
<i>The Valley of Fear</i>	Arthur Conan Doyle
<i>The War in South Africa</i>	Arthur Conan Doyle
<i>The White Company</i>	Arthur Conan Doyle
<i>Through the Magic Door</i>	Arthur Conan Doyle
<i>The Adventures of Sherlock Holmes</i>	Arthur Conan Doyle
<i>Dracula's Guest</i>	Bram Stoker
<i>The Jewel Of Seven Stars</i>	Bram Stoker
<i>The Lady of the Shroud</i>	Bram Stoker
<i>Lair of the White Worm</i>	Bram Stoker
<i>The Man</i>	Bram Stoker
<i>Coral Reefs</i>	Charles Darwin
<i>On the Origin of Species</i>	Charles Darwin
<i>The Voyage of the Beagle</i>	Charles Darwin
<i>The Different Forms of Flowers on Plants</i>	Charles Darwin
<i>American Notes</i>	Charles Dickens
<i>A Tale of Two Cities</i>	Charles Dickens
<i>Hard Times</i>	Charles Dickens
<i>The Old Curiosity Shop</i>	Charles Dickens
<i>A Changed Man; and Other Tales</i>	Thomas Hardy
<i>Desperate Remedies</i>	Thomas Hardy
<i>Far from the Madding Crowd</i>	Thomas Hardy
<i>The Hand of Ethelberta</i>	Thomas Hardy
<i>My Man Jeeves</i>	P. G. Wodehouse
<i>Tales of St. Austin's</i>	P. G. Wodehouse
<i>The Adventures of Sally</i>	P. G. Wodehouse
<i>The Clicking of Cuthbert</i>	P. G. Wodehouse
<i>The Gem Collector</i>	P. G. Wodehouse
<i>The Man with Two Left Feet</i>	P. G. Wodehouse
<i>The Pothunters</i>	P. G. Wodehouse
<i>The Swoop!</i>	P. G. Wodehouse
<i>The White Feather</i>	P. G. Wodehouse
<i>Jacob's Room</i>	Virginia Woolf
<i>Monday or Tuesday</i>	Virginia Woolf
<i>Night and Day</i>	Virginia Woolf
<i>The Voyage Out</i>	Virginia Woolf
<i>Lyrical Ballads</i>	William Wordsworth
<i>Poetical Works</i>	William Wordsworth

Figura 8.1 – Exemplo de derivação de uma rede específica de um autor a partir da rede de livros. Note que se um vizinho aparece em pelo menos uma das redes de livros, então ele aparecerá também na rede do autor.

redes de autores, limitamos nossa análise para as 2.880 palavras que ocorreram em pelo menos um livro de cada autor.

8.2.1 Histograma de co-ocorrência de vizinhos

Para derivar este índice para uma dada palavra w , considere o vetor $\vec{v}(w)$, que armazena a frequência dos $k_n(w)$ vizinhos distintos de w considerando todas as redes de autores. Dessa forma, cada elemento $v_i(w)$ de $\vec{v}(w)$ varia de 1 até 8, pois $n = 8$ autores foram empregados na análise. A consistência baseada em $\vec{v}(w)$ aumenta com a soma das componentes do vetor, já pois se o elemento $v_i(w)$ assume altos valores então a associação correspondente é repetida por diversos autores. A consistência de histogramas é calculada como

$$\mathbb{C}^{(hist)}(w) = \frac{s(w) - k_n(w)}{(n-1)k_n(w)} = \frac{s(w) - k_n(w)}{7k_n(w)}, \quad (8.1)$$

onde s é dado por

$$s(w) = \sum_{i=1}^{k_n(w)} v_i(w). \quad (8.2)$$

Note que, enquanto o numerador da equação 8.1 representa a soma de $v_i(w)$ obtido mínima soma possível (que ocorre quando cada um dos $k_n(w)$ vizinhos ocorrem em somente uma das redes de autores), o denominador representa a variação da soma (o maior valor da soma ocorre quando cada vizinho aparece em n redes de autores). Por este motivo, $\mathbb{C}^{(hist)}$ varia entre zero e um.

8.2.2 Similaridade cosseno

A consistência pode ser calculada para as palavras considerando duas redes de autores $A^{(i)} = \{a_{xy}^{(i)}\}$ e $A^{(j)} = \{a_{xy}^{(j)}\}$, contendo o conjunto de vértices denotados por $V^{(i)}$ e $V^{(j)}$, respectivamente. Inicialmente, $V^{(i)}$ e $V^{(j)}$ são expandidos (mantendo os *links* originais) de forma que o novo conjunto de vértices seja $V^{(ij)} = V^{(i)} \cup V^{(j)}$. Seja w a palavra cuja consistência deve ser calculada. O número de vizinhos de w aparecendo em ambas as redes $A^{(i)}$ e $A^{(j)}$ é

$$\mathcal{N}^{(ij)}(w) = \sum_k a_{wk}^{(i)} a_{wk}^{(j)}. \quad (8.3)$$

Apesar de capturar o número de vizinhos em comum, é difícil interpretar se o valor $\mathcal{N}^{(ij)}$ é alto ou baixo, pois esta medida não é normalizada. A versão normalizada foi definida através da divisão da equação 8.3 pela média geométrica dos graus da palavra w em ambas as redes. Desta forma, se $k_w^{(i)}$ e $k_w^{(j)}$ representam os graus da palavra w nas redes $A^{(i)}$ e

$A^{(j)}$, então a medida normalizada pode ser definida como

$$\tilde{\mathcal{N}}^{(ij)}(w) = \sum_x a_{wx}^{(i)} a_{wx}^{(j)} [k_w^{(i)} k_w^{(j)}]^{-\frac{1}{2}}. \quad (8.4)$$

Para considerar todos os pares de autores na quantificação da consistência, definimos esta medida como a média de $\tilde{\mathcal{N}}^{(ij)}(w)$ sobre os $n(n-1)/2$ pares de redes distintas de autores:

$$\mathbb{C}^{(cos)}(w) = \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j=i+1}^n \tilde{\mathcal{N}}^{(ij)}(w). \quad (8.5)$$

É válido mencionar que $\mathbb{C}^{(cos)}$ varia entre zero e um, de maneira análoga a $\tilde{\mathcal{N}}^{(ij)}$. Na verdade, $\mathbb{C}^{(cos)}$ pode ser interpretado como o cosseno do ângulo formado entre os vetores cujos elementos indicam a presença ou ausência de uma dada palavra como vizinho. Dessa forma, se esses vetores são similares, o ângulo entre eles será pequeno e o cosseno assumirá valores altos, indicando uma alta consistência.

8.2.3 Correlação de Pearson

Uma segunda possibilidade normalizar a quantidade $\mathcal{N}^{(ij)}(w)$ na equação 8.3 é através da comparação com uma vizinhança escolhida aleatoriamente. O número esperado de vizinhos comuns para a palavra w nas redes $A^{(i)}$ e $A^{(j)}$ é dado por $k_w^{(i)}$ e $k_w^{(j)}$. Se a palavra w na rede $A^{(i)}$ escolhe aleatoriamente um dos seus $k_w^{(i)}$ vizinhos, então a probabilidade de se escolher um vértice que também é vizinho de w em $A^{(j)}$ é $M^{-1}k_w^{(j)}$. Aplicando o mesmo raciocínio para os $k_w^{(i)}$ vizinhos de w em $A^{(i)}$, o número esperado de vizinhos comuns é $M^{-1}k_w^{(i)}k_w^{(j)}$. Portanto, se $\mathcal{N}^{(ij)}(w) > M^{-1}k_w^{(i)}k_w^{(j)}$, a consistência é maior do que o valor esperado. Desta forma, a consistência pode ser definida como a diferença entre o número real de vizinhos compartilhados e o número esperado aleatoriamente:

$$\begin{aligned} \mathbb{C}^{(r)}(w) &= \mathcal{N}^{(ij)} - M^{-1}k_w^{(i)}k_w^{(j)} = \sum_x a_{wx}^{(i)} a_{wx}^{(j)} - M^{-1}k_w^{(i)}k_w^{(j)} \\ &= \sum_x a_{wx}^{(i)} a_{wx}^{(j)} - M \langle k_w^{(i)} \rangle \langle k_w^{(j)} \rangle \\ &= \sum_x \left(a_{wx}^{(i)} a_{wx}^{(j)} - \langle k_w^{(i)} \rangle \langle k_w^{(j)} \rangle \right) \\ &= \sum_x \left(a_{wx}^{(i)} - \langle k_w^{(i)} \rangle \right) \left(a_{wx}^{(j)} - \langle k_w^{(j)} \rangle \right), \end{aligned} \quad (8.6)$$

onde $\langle k_w \rangle$ representa o grau normalizado pelo número de vértices da rede:

$$\langle k_w \rangle = M^{-1} \sum_x a_{wx}. \quad (8.7)$$

A equação 8.6 pode ser interpretada como uma medida de covariância, i.e., uma correlação não normalizada. Para transformar $\mathbb{C}^{(r)}$ em uma medida normalizada, $\mathbb{C}^{(r)}$ é dividido pelos correspondentes desvios padrões dos vetores $a_{wx}^{(i)}$ e $a_{wx}^{(j)}$, $x \in \{1, 2, \dots, M\}$. Com esta normalização, $\mathbb{C}^{(r)}$ se transforma no coeficiente de correlação de Pearson:

$$\tilde{\mathbb{C}}^{(r)}(w) = \sum_x \left(a_{wx}^{(i)} - \langle k_w^{(i)} \rangle \right) \left(a_{wx}^{(j)} - \langle k_w^{(j)} \rangle \right) \left[\sum_x \left(a_{wx}^{(i)} - \langle k_w^{(i)} \rangle \right)^2 \sum_x \left(a_{wx}^{(j)} - \langle k_w^{(j)} \rangle \right)^2 \right]^{-\frac{1}{2}}. \quad (8.8)$$

Uma vez que a equação 8.8 representa uma correlação, $\tilde{\mathbb{C}}^{(r)}(w) \in [-1, 1]$. Para manter a consistência com os outros índices nos limites de variação, a seguinte transformação linear foi aplicada em $\tilde{\mathbb{C}}^{(r)}(w)$:

$$\tilde{\mathbb{C}}^{(r)}(w) \mapsto \frac{\tilde{\mathbb{C}}^{(r)}(w) + 1}{2}. \quad (8.9)$$

Desta forma, se $\tilde{\mathbb{C}}^{(r)}(w) > 0,5$ então a quantidade de vizinhos compartilhados é maior que o esperado aleatoriamente.

8.2.4 Índice Leicht-Holme-Newman

No cálculo da consistência baseada na correlação de Pearson, o índice $\tilde{\mathbb{C}}^{(r)}(w)$ foi derivado da diferença entre o número de vizinhos comuns encontrados e o número de vizinhos compartilhados caso as ligações fossem estabelecidas aleatoriamente. Na definição do índice Leicht-Holme-Newman, que é inspirado na medida de similaridade Leicht-Holme-Newman, a nova medida de consistência é obtida a partir da razão das duas quantidades mencionadas anteriormente:

$$\begin{aligned} \mathbb{C}^{(LHN)}(w) &= \mathcal{N}^{(ij)}(w) \left[k_w^{(i)} k_w^{(j)} \right]^{-1} \\ &= M \sum_x a_{wx}^{(i)} a_{wx}^{(j)} \left[\sum_x a_{wx}^{(i)} \sum_x a_{wx}^{(j)} \right]^{-1}. \end{aligned} \quad (8.10)$$

O limiar de interesse é agora $\mathbb{C}^{(LHN)}(w) = 1$, já que para $\mathbb{C}^{(LHN)}(w) > 1$ a consistência assume um valor acima do do valor esperado aleatoriamente.

8.2.5 Índice Sørensen

Diferentemente dos outros índices descritos anteriormente, este índice é baseado na dissimilaridade do conjunto de vizinhos de uma dada palavra em duas redes. Os dois

vetores $A_w^{(i)}$ e $A_w^{(j)}$ que armazenam os vizinhos da palavra w foram usados novamente. A dissimilaridade dos vizinhos foi calculada neste caso com a distância euclidiana normalizada:

$$\begin{aligned}\mathbb{C}^{(euc)}(w) &= \sum_x \left(a_{wx}^{(i)} - a_{wx}^{(j)} \right)^2 \left(k_w^{(i)} + k_w^{(j)} \right)^{-1} \\ &= \sum_x \left(a_{wx}^{(i)} + a_{wx}^{(j)} - 2a_{wk}^{(i)}a_{wk}^{(j)} \right) \left(k_w^{(i)} + k_w^{(j)} \right)^{-1} \\ &= 1 - 2\mathcal{N}^{(ij)}(w) \left(k_w^{(i)} + k_w^{(j)} \right)^{-1}.\end{aligned}\quad (8.11)$$

Uma vez que a distância entre os dois vetores foi dividida pela máxima distância possível $(k_w^{(i)} + k_w^{(j)})$, $\mathbb{C}^{(euc)}(w) \in [0, 1]$. Para tornar esta medida consistente com as outras medidas definidas até aqui, convertamos $\mathbb{C}^{(euc)}(w)$ em uma medida de similaridade:

$$\mathbb{C}^{(s)} = 1 - \mathbb{C}^{(euc)} = 2\mathcal{N}^{(ij)}(w) \left(k_w^{(i)} + k_w^{(j)} \right)^{-1}.\quad (8.12)$$

8.2.6 Frequência de vizinhos compartilhados

Ao contrário dos índices definidos até aqui, este índice considera a frequência com que um dado vizinhos foi empregado por cada autor. Esta medida é justificada pelo fato de que vizinhanças podem ocorrer com as mais variadas frequências. Por exemplo, se uma dada associação de palavras ocorre 100 vezes para um autor e apenas uma vez para outro autor (considerando textos do mesmo tamanho), o índice de consistência seria máximo de acordo com os índices de consistência definidos até agora. Entretanto, esta combinação de palavras claramente não é consistente, já que a frequência de vizinhos é bastante diferente.

A disparidade de frequências foi considerada da seguinte maneira. Suponha que a palavra w possua $k_n(w)$ vizinhos distintos no *corpus* de n autores. Seja v_k um dos vizinhos de w . Se v_k aparece associado a w $f_k^{(i)}$ vezes em $A^{(i)}$ e $f_k^{(j)}$ vezes em $A^{(j)}$, então a consistência de w com relação à associação $w \leftrightarrow v_k$ é

$$\mathbb{C}_{ij}^{(f)}(w \leftrightarrow v_k) = 1 - |f_k^{(i)} - f_k^{(j)}| (f_k^{(i)} + f_k^{(j)})^{-1}.\quad (8.13)$$

Para considerar todos os vizinhos, $\mathbb{C}_{ij}^{(f)}$ é calculada como uma média sobre $\mathbb{C}_{ij}^{(f)}(w \leftrightarrow v_k)$, dado que $f_k^{(i)} + f_k^{(j)} > 0$. Assumindo que esta condição ocorre τ vezes, a consistência baseada na frequência de vizinhos compartilhados é calculada como:

$$\mathbb{C}_{ij}^{(f)}(w) = \tau^{-1} \sum_k \mathbb{C}_{ij}^{(f)}(w \leftrightarrow v_k).\quad (8.14)$$

Tabela 8.2 – Parâmetros obtidos pelo ajuste da curva de probabilidade de consistência com curvas log-normais. A adequabilidade do ajuste foi confirmada pelos valores $R^2 \sim 1$ e teste chi-square $\chi^2 \ll 1$.

Índice	x_c	σ	$\mathcal{A} (\times 10^3)$	R^2	$\chi^2 (\times 10^5)$
Frequência	$0,02416 \pm 0,0011$	$0,97641 \pm 0,03091$	$5,42 \pm 0,14$	0,991	1,53
Histograma	$0,02389 \pm 0,0018$	$0,89291 \pm 0,04733$	$3,42 \pm 0,19$	0,972	2,68
Sørensen	$0,05076 \pm 0,0020$	$1,04088 \pm 0,02737$	$10,55 \pm 0,22$	0,994	0,825
Pearson	$0,04134 \pm 0,0037$	$1,11074 \pm 0,04754$	$6,83 \pm 0,36$	0,986	0,149
Cosseno	$0,06411 \pm 0,0052$	$1,08812 \pm 0,04677$	$12,32 \pm 0,58$	0,989	1,52
LNH	$3,97112 \pm 0,0536$	$0,65551 \pm 0,01073$	$904,14 \pm 12,81$	0,994	1,17

8.3 Distribuição e inter correlação de índices de consistência

Os índices de consistência definidos na Seção 8.2 foram utilizados primeiramente para analisar a distribuição da consistência no conjunto de 2.880 palavras analisadas. As distribuições exibiram um pico e duas caudas assimétricas para todos os índices analisados, como ilustra a Figura 8.2. Esta distribuição mostra que apenas algumas palavras assumem valores extremos de consistência. Em especial, palavras altamente consistente são bastante raras. Formalmente, esta descrição pode ser confirmada pela distribuição log-normal obedecida por estes índices. Em todos os casos, os dados puderam ser explicados pela função log-normal

$$f(x; x_c, \sigma, \mathcal{A}) = \frac{\mathcal{A}}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{(\ln x/x_c)^2}{2\sigma^2}\right), \quad (8.15)$$

onde x_c , σ e $\mathcal{A}$ correspondem aos parâmetros livres da distribuição. A Tabela 8.2 sumariza os parâmetros obtidos para cada caso, e confirma a adequabilidade do ajuste com o valor de Pearson igual a $R^2 \sim 1$ e teste chi-square $\chi^2 \ll 1$.

Distribuições log-normais são amplamente encontradas em contextos não-lingüísticos (114–119). Na lingüística, seu aparecimento é menos frequente. Esta distribuição têm sido observada no comprimento das sentenças e na duração de palavras faladas em conversas por telefone (120, 121). Log-normais geralmente são geradas por processos seguindo leis proporcionalidade (122–124). Por esta razão, a consistência pode ser visualizada como um resultado de um processo de crescimento governado pela constante α_k , que também segue uma distribuição log-normal. Para verificar por que esta afirmação é verdadeira, suponha que uma determinada palavra é inicialmente utilizada por apenas 2 autores com consistência $\mathbb{C}_0$. Para cada novo autor usando esta palavra, a consistência atual aumenta

Figura 8.2 – Distribuição da consistência para 2.880 palavras.

ou diminui de acordo com o fator α_k :

$$\mathbb{C}_k = \alpha_k \mathbb{C}_0. \quad (8.16)$$

Após $n + 2$ autores usarem esta palavra, a consistência final será dada por

$$\mathbb{C}_n = \mathbb{C}_0 \prod_{i=1}^n \alpha_i. \quad (8.17)$$

Na verdade, a consistência de uma palavra é quantificada como uma porcentagem da consistência atual a cada novo uso, sendo esta porcentagem independente da consistência atualmente observada. Uma vez que assumimos que o parâmetro α_k é log-normalmente distribuído, então $\mathbb{C}_k$ seguirá também uma distribuição log-normal, pois o produto de distribuições log-normais também gera uma distribuição log-normal. A suposição de α_k seguir uma distribuição log-normal pode ser desconsiderada se considerarmos que a palavra está presente no vocabulário de vários autores. Sabendo-se que $\ln \mathbb{C}_n$ é dado por

$$\ln \mathbb{C}_n = \sum_{i=1}^n \ln \alpha_i + \ln \mathbb{C}_0, \quad (8.18)$$

e que $\sum_{i=1}^n \alpha_i$ segue uma distribuição normal (pelo teorema do limite central), é possível dizer que $\ln \mathbb{C}_n$ é normalmente distribuído e portanto $\mathbb{C}_n$ segue uma distribuição normal.

Um comportamento similar para os índices de consistência na Figura 8.2 pode indicar que os índices de consistência são correlacionados. De fato, o coeficiente de correlação de Pearson entre os pares de índices foram todos próximos de $\text{Corr} = 1$, com exceção do índice LNH. A correlação variou de $\text{Corr} = 0,902$ entre $\mathbb{C}^{(r)}$ e $\mathbb{C}^{(hist)}$ até $\text{Corr} = 0,996$ entre os índices $\mathbb{C}^{(cos)}$ e $\mathbb{C}^{(s)}$. Portanto, cinco dos índices são equivalentes e podem ser

Tabela 8.3 – Coeficientes de correlação de Pearson e Spearman entre o índice LNH e os outros índices de consistência. Em ambos as medidas de correlação, os valores obtidos são baixos.

Índice	Pearson	Spearman
Histogram	0,044	0,166
Pearson	0,318	0,347
Cosseno	0,301	0,315
Sørensen	0,293	0,295

usados da mesma forma. Com relação ao índice LHN, ele é fracamente correlacionado com os outros índices, como confirma a segunda coluna da Tabela 8.3. Mesmo quando a correlação de *ranks* é usada, baixos valores de correlação foram observados, conforme ilustra a terceira coluna da Tabela 8.3. Portanto, para quantificar a consistência de textos em estudos futuros, basta considerar as medidas baseadas na diferença e na razão entre o número de vizinhos compartilhados reais e o esperado aleatoriamente. Por uma questão de completude, optamos por analisar todos os índices nas próximas seções.

8.4 Correlação entre consistência e atributos linguísticos

Para entender como a consistência medida de acordo com os índices propostos está relacionado com outros fatores lingüísticos, examinamos a relação entre a consistência e o número de vizinhos distintos da palavra. Os gráficos de dispersão para quatro índices de consistência são apresentados na Figura 8.3. Todos eles evidenciam uma forte correlação positiva entre o número de vizinhos e a consistência (os índices não mostrados na figura exibem o mesmo comportamento). As palavras com maior número de vizinhos nas redes tendem a ser mais consistentes. É possível que as palavras com muitos vizinhos (provavelmente aquelas mais frequentes) sejam mais consistentes porque elas são usadas de uma forma mais uniforme já que elas são amplamente utilizadas por diferentes escritores. Por outro lado, as palavras com menos vizinhos provavelmente são palavras menos freqüentes, e portanto estão mais propensas à influência das experiências pessoais de cada escritor.

Uma vez que a distribuição do número de vizinhos para as 2.880 palavras é bastante ampla, dividimos as palavras em quatro classes: C_A , C_B , C_C e C_D , de acordo com a quantidade de vizinhos distintos, conforme mostra a Tabela 8.4. A correlação entre a consistência e os seguintes atributos foram calculados^a: (i) idade de aquisição (idade necessária para que a criança comece a usar a palavra como parte do seu vocabulário falado),

^a As quantidades (i)-(iv) foram obtidas do dataset MRC (125).

Figura 8.3 – Correlação entre o número de vizinhos distintos e os índices de consistência. Em todos os índices existe uma forte correlação entre consistência e o número de vizinhos.

Tabela 8.4 – Número de palavras em cada grupo e variação do número de vizinhos distintos. Enquanto a classe A contém palavras com poucos vizinhos (no máximo 150), a classe D engloba palavras com muitos vizinhos distintos.

Classe	# Palavras (η)	# Vizinhos
Classe A	$0 \leq \eta \leq 150$	1,070
Classe B	$151 \leq \eta \leq 300$	791
Classe C	$301 \leq \eta \leq 500$	463
Classe D	$501 \leq \eta \leq 800$	256

(ii) familiaridade, (iii) concretude, (iv) frequência de uso no Inglês e (v) ambiguidade (i.e., o número de significados distintos obtidos da WordNet (126)). As correlações para cada classe e os índices de consistência que forneceram as maiores correlações estão mostrados na Tabela 8.5.

O único caso em que a correlação foi alta foi para o atributo de familiaridade nas classes C_C e C_D , sendo que as demais correlações assumiram valores inferiores a 0,3. Apesar disso, as tendências de correlação podem ser inferidas a partir do sinal das correlações. Por exemplo, a correlação negativa com a idade de aquisição sugere que palavras pouco consistentes levam mais tempo para ser aprendidas. Este resultado de certa forma era esperado pois o uso heterogêneo de palavras pouco consistentes as torna mais difíceis de serem incorporadas ao vocabulário. Um raciocínio análogo se aplica à familiaridade e concretude (capacidade de imaginar o conceito), pois palavras familiares tendem a induzir os mesmos conceitos (ou seja, palavras amplamente conhecidas geralmente trazem remetem aos mesmos conceitos). Com relação à frequência das palavras na língua inglesa, a correlação positiva indica que palavras amplamente utilizadas são mais consistentes. Na verdade, o uso generalizado de uma palavra provavelmente faz com que ela seja utilizada de uma maneira mais homogênea. Finalmente, a ambigüidade das palavras correlacionou negativamente com sua consistência. Este resultado é uma consequência do fato de que palavras ambíguas tendem a aparecer em mais contextos, e portanto sua vizinhança tende a ser mais heterogênea.

Para entender por que algumas palavras são mais consistentes do que as outras, examinamos a vizinhança das palavras com maiores e menores valores de consistência. A título de ilustração, vamos analisar as seguintes palavras altamente consistentes em C_D : *sleep*, *silence*, *happy*, *hair*, *God* e *cold*. Para cada vizinho dessas palavras, contamos o número de autores que executaram essa associação. Os histogramas com as frequências na Figura 8.4 indicam que as palavras altamente consistentes tendem a possuir vizinhos semanticamente relacionados. Por exemplo, todos os oito autores associaram *sleep* a *night*, *wink* e *dream*. Para confirmar esta hipótese de proximidade semântica, comparamos os vizinhos encontrados nos textos escritos pelos oito autores com os vizinhos da rede semântica derivada do *Edinburgh Associativo Thesaurus* (EAT) (127), que conecta conceitos semanticamente relacionados. Para cada uma das palavras selecionadas com valores

Tabela 8.5 – Correlação r obtida pela comparação de atributos linguísticos com a consistência das palavras para as classes de palavras C_A , C_B , C_C e C_D .

Atributo linguístico	$r(C_A)$	Índice
Idade de aquisição	-0,09	Histogram
Familiaridade	+0,18	LHN
Concretude	+0,07	Pearson
Frequência	+0,12	Sørensen
Ambiguidade	-0,13	LHN
Atributo linguístico	$r(C_B)$	Índice
Idade de aquisição	-0,15	Histogram
Familiaridade	+0,27	Histogram
Concretude	+0,13	LHN
Frequência	+0,13	Sørensen
Ambiguidade	-0,13	LHN
Atributo linguístico	$r(C_C)$	Índice
Idade de aquisição	-0,24	Pearson
Familiaridade	+0,37	Histogram
Concretude	+0,14	Sørensen
Frequência	+0,03	Cosine
Ambiguidade	-0,14	LHN
Atributo linguístico	$r(C_D)$	Índice
Idade de aquisição	-0,20	Sørensen
Familiaridade	+0,43	Sørensen
Concretude	+0,07	Histogram
Frequência	+0,20	Sørensen
Ambiguidade	-0,20	LHN

Figura 8.4 – Frequência da associação de vizinhos para as palavras *sleep*, *silence*, *help*, *hair*, *God* e *cold*. Note que as palavras mostradas nesta figura (palavras altamente consistentes) apresentam vizinhanças que parecem estar restritas a contextos limitados.

elevados de consistência, computamos a sua distância do vizinho da rede de autores na rede semântica EAT. Em seguida, calculamos a razão entre o número de vizinhos que estão a uma distância d e o número esperado de vizinhos que estão à mesma distância, supondo que os nós fossem escolhidos aleatoriamente na rede EAT. Os resultados na Tabela 8.6 mostram que o número de vizinhos imediatamente relacionados na rede EAT é muito mais elevado do que o esperado, confirmando assim que a consistência está relacionada à limitação do contexto semântico.

8.5 Usando a consistência para identificar autoria

As correlações com os atributos lingüísticos confirmaram a adequação dos índices de consistência para quantificar aspectos semânticos no texto. Nesta seção investigaremos se esses índices podem ser usados para identificar a autoria, já que autores distintos podem usar palavras mais ou menos consistentes. No caso extremo em que comparamos autores de gêneros literários distintos, esperamos uma boa distinção. Esta hipótese foi investigada

Tabela 8.6 – Razão entre o número de vizinhos da rede de autores que estão a uma distância d do conceito central na rede EAT e o número esperado de vizinhos a uma distância d assumindo associações aleatórias. Interessantemente, a vizinhança imediata das palavras mais consistentes na rede de autores também está concentrada na vizinhança imediata ($d = 1$) da rede EAT. Isto significa que palavras consistentes induzem contextos específicos.

Palavra	d=1	d=2	d=3	d=4
<i>sleep</i>	15,74	0,66	0,00	0,18
<i>silence</i>	20,95	2,82	0,30	0,36
<i>happy</i>	19,92	1,75	0,05	0,00
<i>hair</i>	26,14	1,31	0,15	0,00
<i>God</i>	32,79	1,18	0,06	1,21
<i>die</i>	18,20	2,28	0,30	0,86
<i>cold</i>	9,19	0,69	0,00	0,60
<i>chair</i>	18,31	1,93	0,40	0,00

através da definição da consistência média de cada autor como

$$\langle \mathbb{C} \rangle = n_f^{-1} \sum_{i=1}^{n_p} \mathbb{C}(w) f_w, \quad (8.19)$$

onde n_p e n_f representam respectivamente o número de palavras distintas e o número de *tokens* no *corpus* de um dado autor e $\mathbb{C}(w)$ e f_w representam respectivamente a consistência e a frequência da w -ésima palavra. Usando $\mathbb{C}^{(hist)}$ para calcular $\langle \mathbb{C} \rangle$ para cada autor na equação 8.19, comparamos a consistência média para todos os pares de 28 autores distintos. A Figura 8.5 ilustra em escalas de cinza os p -values obtidos na comparação dos valores de $\langle \mathbb{C} \rangle$. Enquanto alguns pares de autores podem ser facilmente distinguidos (p.e. Stocker e Darwin em C_A , Wolf e Hardy em C_B , Darwin e Doyle em C_C , e Wodehouse e Dickens em C_D), outros pares de autores apresentam consistência média bastante similares (Stocker e Wolf em C_D). Além disso, a partir da contagem dos níveis mais escuros de escalas de cinza é possível notar que as palavras em C_A são melhores para distinguir autores do que as palavras nas outras classes. As palavras em C_D parecem fornecer a pior capacidade discriminativa.

Finalmente, para examinar como os autores são agrupados em termos de consistência, usamos o algoritmo de agrupamento UPGMA (128) baseado na similaridade cosseno. Neste caso, cada autor é caracterizado por um vetor de forma que cada elemento represente a frequência da palavra corresponde no livro do autor. A Figura 8.6 mostra que quatro *clusters* podem ser identificados a partir da escolha adequada de um limiar de distâncias. Significativamente, os autores de romances (como Wodehouse) estão separados dos autores de trabalhos científicos (como Darwin), o que indica que as diferenças em estilo estão refletidas no uso de palavras com consistência distintas. Enfim, os índices de consistência são de fato capazes de reconhecer autoria. Esta descoberta agora pode ser combinada com outros métodos convencionais para aperfeiçoar a habilidade de distinção de autores.

Figura 8.5 – Comparação da consistência média para os índices de consistência nas classes C_A , C_B , C_C e C_D . Quanto mais clara a escala de cinza, menor o p -value e maior a diferença na consistência média.

Figura 8.6 – Agrupamento hierárquico usando a consistência média. Note que os textos de Darwin podem ser considerados como *outliers*, já que são textos predominantemente científicos contendo palavras consistentes.

8.6 Conclusões

Neste capítulo estudamos o problema de quantificar a complexidade da escrita considerando a consistência de palavras. Assumindo que palavras consistentes induzem contextos gramaticalmente/semanticamente limitados, definimos vários índices para medir a tendência de uma palavra para ser usada de forma homogênea, isto é, preservando o contexto. Usando os vários índices propostos, verificamos que a consistência das palavras pode ser descrita por uma distribuição log-normal. Curiosamente, verificamos que a consistência pode ser modelada a partir de um processo de crescimento multiplicativo. Cada novo escritor usando a palavra modifica a consistência atual através da adição ou remoção de uma fração do valor de consistência atual. As palavras mais frequentes e familiares tendem a ser usadas de forma mais consistente, enquanto que as palavras ambíguas e palavras que demandam mais tempo para ser aprendidas tendem a ser menos consistentes. Finalmente, confirmamos que a quantificação de complexidade em termos de caracterização por consistência é capaz de distinguir autores, especialmente aqueles especializados em gêneros diferentes.

Investigações futuras poderiam introduzir novas métricas para a quantificar a consistência, a fim de verificar se os nossos resultados são preservadas. Como ponto de partida, a medida Katz (129) poderia ser empregada para quantificar a distância entre conceitos. Analogamente, o estudo da consistência poderia ser extrapolado para vizinhanças maiores

que a vizinhança imediata. Uma vez que a consistência das palavras está relacionada com a semântica, esta medida pode ser combinada com recursos relacionados ao estilo de escrita (p.e. medidas topológicas de redes complexas) para verificar se a distinguibilidade é reforçada. Por fim, os índices de consistência poderiam ser úteis em aplicações para quantificar subjetividade (em textos escritos ou no discurso transcrito), pois palavras subjetivas potencialmente assumem valores de consistência mais baixos pois pessoas diferentes provavelmente associam vizinhos distintos a conceitos subjetivos.

Capítulo 9

Tradução automática

A tarefa de tradução automática, que basicamente converte um texto produzido em uma linguagem fonte para um texto em uma linguagem alvo (130), foi provavelmente uma das primeiras aplicações não numéricas da computação. Os pesquisadores da área têm proposto diversos métodos e paradigmas de tradução (131): métodos diretos e indiretos, com ou sem transferência linguística, seguindo paradigmas superficiais ou profundos. Outras possibilidades incluem o uso de regras, traduções armazenadas como exemplos e outras estatísticas. A área se iniciou a partir de regras de traduções obtidas manualmente para cada par de línguas fonte-alvo. Apesar de representar uma tarefa altamente custosa (com relação ao esforço e tempo), ela ainda é usada em algumas aplicações com resultados bastante limitados. Atualmente, o estado da arte tem sido obtido com modelos estatísticos relativamente simples, como aqueles propostos pelo grupo de pesquisa da IBM (132–137). Tais modelos, que necessitam de grandes quantidades de textos durante a fase de treinamento, usualmente geram tabelas de traduções que são automaticamente “aprendidas” em um intervalo de tempo relativamente curto. Outra vantagem dos modelos estatísticos refere-se ao fato de que novos pares de traduções podem ser facilmente produzidos utilizando novos *corpora* bilíngues. Atualmente, tais sistemas têm sido capazes de gerar traduções de alta qualidade até mesmo para línguas bastante distintas.

Apesar da sua história relativamente longa, o processo de avaliação de traduções automáticas tem sido controversa. Avaliações manuais de traduções são obviamente mais custosas, demandam mais tempo e estão propensas a erros e inconsistências humanas. Em vista destes problemas, métricas automáticas têm sido propostas para prover avaliações objetivas, reproduzíveis e imparciais. Uma das medidas mais representativas é denominada como *Bilingual Evaluation Understudy* (BLEU) (138, 139). Esta medida basicamente conta o número de n-gramas comuns presentes nas traduções automáticas e nas traduções (manuais) de referência. Quanto maior o valor obtido pelo BLEU, maior é a qualidade da tradução. Os autores desta medida mostraram que ela é tão precisa quanto a avaliação humana.

O Instituto Nacional de Padrões e Tecnologia (NIST) conduz competições anuais de tradutores automáticos para medir o avanço do estado da arte na área. A competição adota o BLEU como principal medida de qualidade e considera pares de línguas bastante distintas, tais como Chinês-Inglês e Árabe-Inglês. Geralmente, os pesquisadores na área assumem que um método de tradução que funciona para línguas bastante distintas irá funcionar para línguas mais similares. Particularmente, o Chinês e o Árabe têm sido utilizados como línguas exóticas pois as traduções envolvendo tais idiomas demandam reordenações de longo-alcance em suas palavras. Recentemente, os resultados das competições (são disponibilizadas publicamente na página do NIST (140)) mostram que os métodos estatísticos apresentam o melhor desempenho. Além disso, eles têm sido significativamente melhores que outros métodos baseados em regras.

A motivação para se empregar métodos baseados em redes para analisar a qualidade de traduções vem do fato que as conexões entre conceitos (vértices) também são levados em consideração nas estatísticas de avaliações automáticas. Considerando estas conexões, podem-se utilizar as estatísticas das redes para estudar a co-ocorrência de conceitos e até mesmo estabelecer domínios semânticos. Neste capítulo, estudamos como os fatores estilísticos subjacentes ao texto podem ser quantificados com métricas de redes complexas para distinguir traduções manuais de automáticas, e traduções de qualidade distintas. Daremos atenção especial à aplicação de medidas hierárquicas (141), pois tais medidas fornecem uma quantificação compreensiva da topologia da rede ao considerar contextos mais profundos ao redor de cada vértice.

9.1 Distinção de diferentes tipos de traduções

As traduções automáticas e manuais foram modeladas e caracterizadas segundo a metodologia definida nos Capítulos 2 e 3. Além das medidas básicas, variações destas medidas foram tomadas para aperfeiçoar o processo de avaliação de traduções. Três tipos de caminhos mínimos foram utilizados. O primeiro leva em consideração todas as arestas com peso unitário. O segundo tipo é calculado considerando o complemento dos pesos nas arestas. Em outras palavras, o peso de cada aresta é substituído pela subtração do peso máximo encontrado na rede pelo peso da aresta. Finalmente, o terceiro tipo considera as arestas com pesos invertidos: o peso w_{ij} para cada aresta é substituído por w_{ij}^{-1} .

Uma nova medida relacionada ao grau e à entropia da rede foi utilizada. Esta medida, conhecida como informação de busca, quantifica a dificuldade de se encontrar um vértice em uma dada caminhada. A dificuldade é determinada considerando-se caminhadas aleatórias da rede e a probabilidade de um caminho mínimo ser seguido aleatoriamente. Seja $\mathbb{L}(i, b)$ o caminho mínimo ligando os vértices i e b . Um caminho aleatório é escolhido aleatoriamente

a partir do vértice i . Se $p_i(b)$ representa a probabilidade do caminho mínimo ser seguido, ele pode ser definido conforme:

$$p_i(b) = \frac{1}{k_i} \prod_{j \in \mathbb{L}(i,b)} \frac{1}{k_j - 1}, \quad (9.1)$$

onde a multiplicação é feita ao longo do caminho, exceto para os vértices iniciais e finais. Em uma rede não direcionada, k_i é escolhido como o grau do vértice. Uma vez que utilizamos grafos direcionados, k_i é utilizado como o grau de saída do vértice i . A informação de busca é relacionada com a entropia de todos os caminhos mínimos envolvendo estes dois vértices e é dada por:

$$\mathcal{H}(i, b) = - \sum_{\mathbb{L}(i,b)} p_i(b) \log_2 p_i(b). \quad (9.2)$$

Para complementar a caracterização topológica da rede usamos a medida de coeficiente cíclico, que representa uma extensão da medida de aglomeração pois considera ciclos envolvendo mais que três vértices. Trata-se de uma medida local, portanto, cada vértice possui um valor específico. Ela é definida como

$$CY_i = \frac{2}{k_i(k_i - 1)} \sum_j \sum_k \frac{a_{ij}a_{ik}}{T_{ijk}}, \quad (9.3)$$

onde o termo T_{ijk} representa o tamanho mínimo do ciclo que possui os vértices i , j e k . T_{ijk} assume seu valor mínimo quando i , j e k estão conectados em um ciclo, formando um triângulo. Dessa forma, o valor mínimo para T_{ijk} é $\min T_{ijk} = 3$. Esta fórmula local representa a média do inverso dos menores ciclos entre um vértice e seus vizinhos. Portanto, esta medida é diretamente proporcional ao número de ciclos de comprimento curto.

Para os experimentos reportados neste capítulo, textos contendo aproximadamente 500 vértices foram utilizados. Estes textos foram obtidos da revista de divulgação científica Pesquisa FAPESP (142), sendo o Português a língua fonte. Estes textos foram selecionados aleatoriamente da seção “Laboratório”. As traduções para o Espanhol (língua alvo) foram realizadas por especialistas humanos e por dois sistemas de tradução automática, Apertium (143) e Intertran (144). Para o Inglês, a tradução foi realizada com os sistemas Free Translation (145) e Intertran, além das traduções manuais. O alinhamento das duas versões (línguas fonte e alvo) foi realizado com o sistema LIHLA (146), que permitiu que o mapeamento dos vértices fosse feito para testar a preservação da topologia das redes para os textos alvo e fonte.

Após a modelagem dos textos fonte e alvo como redes complexas, as redes foram caracterizadas com medidas básicas e com as medidas de informação de busca e coeficiente cíclico. O mapeamento de vértice para vértice entre textos fonte e alvo foi utilizado como métrica de distinção de traduções. Para cada métrica (p.e. k , L , $\mathcal{H}$), gráficos foram construídos de forma que as medidas locais para o texto fonte fossem representadas nas

abscissas e as medidas locais da rede alvo fossem representadas nas ordenadas. O método dos mínimos quadrados foi utilizado para se obter a melhor reta, cujo coeficiente angular é um dos parâmetros utilizados na quantificação da qualidade das traduções. O outro parâmetro utilizado foi o coeficiente de Pearson, calculado a partir das medidas das redes representando os textos originais e as traduções. Este coeficiente quantifica a relação entre as medidas locais dos vértices fonte (original) e alvo (traduzido). Portanto, para cada par de tradução e para cada par de medida de rede, obtivemos o coeficiente de correlação de Pearson e o coeficiente angular. Adotando esta metodologia, verificamos que a melhor medida foi a terceira versão da medida de caminhos mínimos. A Figura 9.1 mostra que existe uma alta tendência de preservação desta medida para o Apertium, já que ambos os coeficientes estão próximos de 1. Diferentemente, o Intertran apresentou um baixo valor de coeficientes de Pearson e angulares. Os resultados obtidos para a tradução humana se encontram entre os dois sistemas de tradução automática, o que pode ser explicado pelo fato de que a preservação das medidas dos vértices não é tão alta como em sistemas de tradução automática de qualidade porque os humanos tendem a rephrasear textos durante o processo de tradução. Os valores de CY foram preservados nos textos fonte e alvo, independentemente do sistema de tradução. Com relação à métrica de informação de busca $\mathcal{H}$, observamos uma fraca preservação dos valores locais. Esses resultados sugerem que o uso isolado destas duas medidas não é útil para identificar a qualidade de traduções.

Quando todas as medidas foram combinadas em uma análise multivariada usando a plataforma *Projection Explorer* (PEx) (147) (ver Figura 9.2), uma discriminação razoável pôde ser obtida entre tradução humana e automática. No entanto, a mesmo poder discriminativo não foi obtido entre tradutores automáticos, uma vez que os grupos gerados usando a ligação de Ward (148) na Figura 9.3 aglomerou no mesmo grupo a tradução humana (H) e as traduções do Intertran (I). Quando o grau (de saída e entrada) e o coeficiente de aglomeração orientados foram também considerados, a capacidade de distinção aumentou significativamente, conforme ilustrado nas Figuras 9.4 e 9.5. Em experimentos subsidiários, notamos que se medidas alternativas fossem utilizadas, haveria uma pequena mudança na habilidade de distinção, já que as duas métricas escolhidas para a Figura 9.4 estão entre as melhores para a tarefa. Portanto, as melhores métricas para a distinção são o grau (de saída e entrada), o comprimento médio dos caminhos mínimos (a versão com peso w^{-1}) e o coeficiente de aglomeração.

Utilizando aprendizado de máquina, uma distinção igualmente boa foi obtida. A Tabela 9.1 mostra que com o algoritmo Expectation Maximization (149), as traduções humanas e as traduções do Apertium se concentraram principalmente no primeiro *cluster*, enquanto que as traduções do Intertran se organizaram totalmente no segundo *cluster*. Este resultado confirma que o uso das medidas de redes complexas permite a distinção de traduções automáticas de baixa e alta qualidade. Além disso, uma análise realizada com o algoritmo C4.5 mostrou ser possível identificar o sistema de tradução com uma taxa de

Figura 9.1 – Coeficiente angular e de Pearson para a medida de comprimento médio dos caminhos mínimos. Enquanto na tradução de alta qualidade (Apertium) existe uma alta tendência de preservação, na tradução de baixa qualidade a preservação das medidas não é clara.

Tabela 9.1 – Taxa de acerto na distinção de traduções.

Tradução	Cluster 1	Cluster 2
Apertium	95%	5%
Humana	80%	20%
Intertran	0%	100%

acerto de 95%, conforme ilustrado na matriz da Tabela 9.2.

Para o par Inglês-Português, a mesma abordagem foi seguida. Neste caso, os textos em inglês foram obtidos manualmente e com dois sistemas de tradução automática: Free Translation e Intertran. Os resultados sumarizados na Tabela 9.3 representam a matriz de confusão obtida com o algoritmo C4.5. Embora a acurácia obtida na distinção do par Inglês-Português tenha sido inferior, com uma média de 83%, ainda é possível distinguir entre os dois sistemas de tradução automática já que existe um erro de apenas 5% na classificação entre Intertran e Free Translation.

Enfim, os métodos de aprendizado de máquina em combinação com medidas de redes complexas parecem ser promissores na tarefa de distinção de diferentes tipos de traduções automáticas, e consequentemente na avaliação da qualidade de sistemas de tradução. Tais resultados sugerem que nossos métodos podem ser combinados com métricas tradicionais

Figura 9.2 – Análise multivariada usando o software PEx para distinção de textos traduzidos do Português para o Espanhol, em três tipos de traduções: Apertium (azul), Intertran (preto) e tradução humana (vermelho).

Figura 9.3 – Análise de agrupamento hierárquico usando a ligação de Ward na distinção de textos traduzidos do Português para o Espanhol, em três tipos de traduções: Apertium (A), Intertran (I) e Humana (H).

Tabela 9.2 – Matriz de confusão para a classificação de textos traduzidos em Espanhol.

Apertium	Humana	Intertran	Classificada como ($\leftarrow$)
100%	0%	0%	Apertium
5%	90%	5%	Humana
0%	5%	95%	Intertran

Tabela 9.3 – Matriz de confusão para a classificação de textos traduzidos em Inglês.

Intertran	Free Translation	Humana	Classificada como ($\leftarrow$)
90%	5%	5%	Intertran
5%	90%	5%	Free Translation
15%	15%	70%	Humana

9.2 Avaliação de traduções automáticas com métricas hierárquicas

A qualidade de traduções automáticas também foi avaliada com medidas hierárquicas (141). As medidas hierárquicas foram calculadas para cada vértice da rede fonte (v_f) e da rede alvo (v_a), com os níveis hierárquicos variando de $h = 1$ até o tamanho máximo das redes analisadas. Cada vértice na rede fonte e alvo foi representado pelas seguintes matrizes $\Phi_v^x = \{u_{hj}\}$, onde x representa a rede fonte (f) ou a rede alvo (a), h representa o nível hierárquico e j representa a j -ésima métrica obtida para o vértice v . Dado um vértice i da rede fonte, a relação binária $\Psi : i \mapsto j$ é definida como

$$\Psi_h(i) = j \mid j \in \text{rede alvo}, \quad (9.4)$$

considerando que o erro do mapeamento $i \mapsto j$ é mínimo, segundo um dado critério. Matematicamente, o erro do mapeamento entre os dois vértices i e j é dado por

$$\epsilon_h(i, j) = \sum_m \left(\phi_i^f(h, m) - \phi_j^a(h, m) \right)^2 \quad (9.5)$$

e o erro acumulado considerando todos os níveis é

$$\epsilon_h^{AC}(i, j) = \sum_h \sum_m \left(\phi_i^f(h, m) - \phi_j^a(h, m) \right)^2. \quad (9.6)$$

A cada passo, o vértice j na rede alvo escolhido com o melhor mapeamento do vértice i é aquele que satisfaz a relação

$$\hat{j}^{(R)}(i) = \arg \min_j \sum_m \left(\phi_i^f(h, m) - \phi_j^a(h, m) \right)^2, \quad (9.7)$$

caso apenas um nível seja considerado. Caso contrário,

$$\hat{j}^{(R)}(i) = \arg \min_j \sum_h \sum_m \left(\phi_i^f(h, m) - \phi_j^a(h, m) \right)^2. \quad (9.8)$$

Para evitar que as métricas assumindo maiores valores absolutos apresentem uma maior contribuição na definição de ϵ e ϵ^{AC} , cada componente da soma é normalizada. A acurácia do alinhamento é calculada a partir da comparação do alinhamento obtido com redes complexas $\hat{j}^{(R)}(i)$ e com o alinhador automático LIHLA, $\hat{j}^{(L)}(i)$:

$$T_h = M^{-1} \sum_{i=1}^M \delta(\hat{j}^{(R)}(i), \hat{j}^{(L)}(i)), \quad (9.9)$$

onde M na equação 9.9 refere-se ao número de vértices na rede fonte e δ é calculado como

$$\delta(x, y) = \begin{cases} 1 & \text{se } x = y, \\ 0 & \text{caso contrário.} \end{cases}$$

Note que a abordagem descrita é conceitualmente fundada no princípio de que os vértices correspondentes nas redes originais e traduzidas tendem a possuir o máximo de similaridade contextual, calculada pelas medidas adotadas. Neste contexto, as medidas hierárquicas são particularmente interessantes pois elas fornecem uma caracterização mais detalhada dos padrões de conectividade ao redor de cada vértice (150).

A motivação para se utilizar a taxa de acerto com medidas hierárquicas vem dos resultados descritos em (150), onde o uso de medidas hierárquicas permitiu um melhor mapeamento da rede de proteínas do organismo eucariota *S. Cerevisiae* (151). Nos experimentos realizados para o par Português-Espanhol, 50 textos foram utilizados, sendo o Português a língua fonte e as traduções sendo realizadas por humanos e pelos sistemas Apertium e Intertran. Similarmente aos experimentos descritos na seção anterior, o mapeamento de referência foi obtido a partir do alinhador automático LIHLA.

A Figura 9.6 mostra a taxa de acerto obtida na associação do vértice correto no texto alvo que corresponde a uma dado vértice no texto fonte. Os valores são consideravelmente altos, considerando que uma associação aleatória levaria a uma taxa aproximada de $1/M \simeq 1/500 \simeq 0,2\%$ (chance de se encontrar o vértice correto entre $M = 500$ vértices da rede alvo em uma escolha aleatória). O mapeamento foi particularmente preciso para o tradutor de alta qualidade (Apertium). Além disso, analogamente ao estudos realizados com a rede do organismo *S. Cerevisiae*, a precisão aumentou para os primeiros níveis hierárquicos, antes de se manter constante. Quando muitos níveis hierárquicos foram considerados, houve influência de ruídos relacionados com informações não correlatas. Consequentemente, a acurácia diminuiu, conforme ilustra a Figura 9.6.

Uma maior precisão no mapeamento com o uso de níveis hierárquicos profundos foi

Figura 9.6 – Taxa de acerto no mapeamento do vértice fonte (Português) no vértice alvo (Espanhol) em função do nível hierárquico aplicado para cada métrica. Uma maior taxa de acerto foi encontrada para o Apertium.

Tabela 9.4 – Matriz de confusão para método hierárquico

Google	Intertran	Classificada como ($\leftarrow$)
88%	12%	Google
20%	80%	Intertran

também observada para o par Português-Inglês, sendo que os textos alvos em Inglês foram obtidos com os sistemas de tradução automática do Google e Intertran. Em cada caso, 50 textos com aproximadamente 500 palavras distintas foram utilizados. A taxa de acerto obtida neste caso não foi tão alta como no caso dos pares Espanhol-Português, novamente refletindo as diferenças estruturais mais evidentes no par Inglês-Português, o que impede a preservação dos vértices na rede alvo. Além disso, a distinção entre os sistemas de tradução automática não foi tão evidente, como pode ser inferido a partir de uma inspeção visual na Figura 9.7.

A capacidade para distinguir entre diferentes tipos de tradução também foi verificada com métodos de aprendizado de máquina em conjunto com as medidas hierárquicas em níveis superiores. Em especial, altas taxas de acerto foram encontradas na distinção das traduções Espanhol-Português (98%) e Inglês-Português (84%) com o método C4.5. A matriz de confusão da classificação do par Inglês-Português é ilustrada na Tabela 9.4.

Enfim, o uso de apenas um atributo (a taxa de acerto no mapeamento vértice a

Figura 9.7 – Taxa de acerto no mapeamento do vértice fonte (Português) no vértice alvo (Inglês) em função do nível hierárquico aplicado para cada métrica.

vértice) foi suficiente para distinguir traduções de qualidade distintas. Mesmo quando as medidas foram calculadas com 10 níveis hierárquicos, os resultados obtidos com métodos de reconhecimento de padrão indicaram que os primeiros 3 ou 4 níveis já são suficientes para gerar as maiores taxas de acerto no mapeamento.

9.3 Conclusões

Neste capítulo exploramos diversas medidas de redes complexas para avaliar a qualidade de traduções automáticas. As melhores distinções foram obtidas com as medidas de grau, caminhos mínimos e coeficiente de aglomeração. As medidas de coeficiente cíclico e informação de busca apenas foram úteis quando utilizadas em combinação com outras métricas. Em especial, a tendência de preservação destas medidas parece não depender do tipo de tradução. Os métodos de reconhecimento de padrões foram capazes de distinguir traduções humanas de automáticas com alta precisão para o par Português-Espanhol, e com uma precisão menor para o par Português-Inglês.

As estratégias desenvolvidas nesta tese se diferenciam das estratégias tradicionais que usam similaridade lexical para distinguir traduções automáticas de humanas. Em trabalhos relacionados, como o desenvolvido por Patel e Radev (152), traduções manuais

e automáticas foram discriminadas a partir do uso de medidas de similaridade tradicionais, como a distância Levenshtein (136) e BLEU (138). Neste sentido, nossos resultados podem ser entendidos como complementares às descobertas de Patel e Radev, já que mostramos que quando informações topológicas (medidas da rede) e léxicas (alinhamento lexical) são combinadas, não apenas traduções humanas e automáticas podem ser diferenciadas umas das outras, mas também traduções automáticas de qualidades distintas.

Dois resultados importantes puderam ser obtidos durante a aplicação de níveis hierárquicos na análise de redes complexas representando textos. Primeiramente, observamos que os primeiros 3-4 níveis hierárquicos são suficientes para aperfeiçoar a habilidade de mapeamento da rede fonte na rede alvo. Também verificamos que o uso da taxa de acerto no mapeamento forneceu novos atributos que podem ser utilizados em sistemas de avaliação de qualidade. Os resultados obtidos neste capítulo sugere a possibilidade de associação dos nossos métodos com outros métodos tradicionais, já que a natureza dos atributos são complementares. De fato, os bons resultados obtidos na avaliação da qualidade mostrou, como prova de princípio, que a topologia das redes é útil para inferir a qualidade de traduções. Dessa forma, a abordagem proposta nesta tese pode ser estendida para um número maior de línguas. Possivelmente, nossa abordagem também será útil para avaliar traduções humanas.

Capítulo 10

Complexidade textual

O processo de simplificação textual é definido como o conjunto de operações aplicadas a um texto para desacentuar a sua complexidade linguística através da redução do léxico e atenuação de construções sintáticas complexas, com nenhuma ou pouca perda da informação principal. Trata-se, portanto, de uma tarefa relevante tanto para pessoas quanto para máquinas. No primeiro caso, pessoas com baixo nível de letramento, com deficiências cognitivas (p.e. afasia e dislexia), crianças e aprendizes de línguas estrangeiras são beneficiadas pela apresentação simples de um texto. Já no segundo caso, a simplificação textual pode ser utilizada como um pré-processamento para outras tarefas a fim de simplificar uma etapa posterior cuja eficiência é dependente do grau de complexidade da entrada.

Existem vários tipos de simplificação, desde a sumarização (para diminuir o tamanho do texto) até o aumento da inteligibilidade (eliminação de palavras pouco frequentes e estruturação sintática pouco complexa) e da legibilidade (estrutura de apresentação do texto). Neste capítulo, avaliamos o uso de redes complexas para caracterizar a complexidade de um texto. Em especial, esta tarefa pode ser útil não apenas para quantificar a complexidade da escrita humana, mas também pode ser útil para identificar automaticamente textos que necessitem de simplificação na Web a fim de aumentar a acessibilidade. Também, neste capítulo, criamos um método para simplificação textual que é baseado em sumarização extrativa.

10.1 Abordagens tradicionais

A quantificação da complexidade textual representa uma tarefa desafiadora para linguistas e estudiosos da área, já que o grau de complexidade de um texto pode depender de inúmeros fatores. Provavelmente não existe uma regra simples capaz de estabelecer uma relação direta e infalível entre a percepção de complexidade e métricas quantitativas.

Apesar disso, muitas facetas da complexidade podem ser avaliadas satisfatoriamente quando avaliadas isoladamente. Aspectos específicos da complexidade têm sido avaliados na literatura, com destaque para as características léxicas, sintáticas e estruturais, como discutidas brevemente a seguir.

Com relação à dimensão léxica, o uso de palavras de baixa frequência faz com que a complexidade dos textos aumente (153), especialmente para textos em uma língua não-nativa do leitor. Por esta razão, a manipulação de termos linguísticos incomuns (p.e. palavras raras) tem sido utilizada como uma das estratégias de simplificação (154). No entanto, esta manipulação nem sempre garante uma maior facilidade na leitura, já que esta característica tipicamente depende também de outros fatores (155). As técnicas e metodologias provenientes da teoria da informação também têm sido utilizadas para estimar a complexidade em análises léxicas, onde textos são tratados como sinais conduzindo alguma informação. A complexidade textual, segundo esta teoria, está associada com a quantidade de informação transmitida. Matematicamente, um dado manuscrito é representado como um sinal transportando palavras que ocorrem com uma certa probabilidade, que por sua vez é utilizada para calcular a entropia de Shannon.

Para estimar a complexidade de acordo com atributos sintáticos, as estruturas linguísticas que são intuitivamente mais complexas são ponderadas com maior peso na análise. Entre as principais medidas da análise sintática estão o comprimento da sentença (110, 156, 157) e o número de nós na árvore de *parsing* (158). De acordo com esta última medida, árvores envolvendo muitos vértices demandam um tempo de processamento maior. De fato, estudos similares indicaram que ambas as medidas sintáticas (comprimento da sentença e número de nós na árvore de *parsing*) estão diretamente correlacionadas com a percepção humana de complexidade. Uma estratégia alternativa considera a complexidade segundo a ordem de aparecimento das palavras nas sentenças. Esta complexidade estrutural é diretamente proporcional à raridade das ordenações de classes gramaticais empregadas.

Talvez as medidas mais conhecidas para estimar a complexidade em textos sejam os índices de legibilidade, que estão detalhados na Tabela 10.1. As fórmulas empregadas para calcular estes índices normalmente empregam os mesmos atributos, tais como número de palavras, sentenças e sílabas. De acordo com (158), este é um sinal de que a percepção de complexidade é compartilhada por vários autores, já que estes propuseram medidas similares de forma independente.

Todas as métricas de quantificação de complexidade apresentadas na Tabela 10.1 consideram palavras ou sentenças. O relacionamento global das palavras nos textos não é levado em consideração. Até onde sabemos, a estrutura global dos textos ainda não foi utilizada para quantificar complexidade, apesar dos vários trabalhos sugerindo que fenômenos linguísticos e cognitivos podem ser identificados pela análise da estrutura das redes. Por este motivo, nesta tese estudamos como os textos podem ser classificados com

Tabela 10.1 – Índices de legibilidade. (a) R: índice de legibilidade (0-100); (b) G: nível de estudo (0-12) necessário para responder 50% das questões a respeito do texto; e (c) ML: nível de estudo (0-12) necessário para responder a 100% das questões sobre o texto. Os atributos utilizados nas fórmulas são: (a) W: número de palavras; (b) T: número de sentenças; (c) D: número da palavras na lista *Dale Long* (159); (d) S: número de sílabas; (e) M: número de palavras monossílabas; e (f) P: número de palavras com três sílabas ou mais.

Autor	Equação	Referência
Flesch	$R = 206,835 - 84,6 S/W + 1,015 W/T$	(160)
Coleman	$R = -37,95 + 116 M/W + 148 T/W$	(161)
Dale, Chall	$G = 19,4265 - 15,79 D/W + 0,0496 W/T$	(159)
Holquist	$G = 14,862 - 11,42 D/W + 0,0512 W/T$	(162)
Gunning	$G = 3,0680 + 9,84 P/W + 0,0877 W/T$	(163)
McLaughlin	$ML = 3,1291 + 5,7127 \sqrt{P/T}$	(164)

relação à sua complexidade usando medidas que emergem da análise topológica de redes.

10.2 Projeto PorSimples

O banco de dados utilizado no estudo de quantificação dos índices de complexidade consiste num conjunto de 339 textos compilados pelo projeto Projeto PorSimples (165). Os textos são classificados em textos originais, simplificações naturais e fortes. A diferença entre as simplificações refere-se à quantidade de operações de simplificações aplicadas. As possíveis operações de simplificação são fornecidas na Tabela 10.2, enquanto que as estatísticas do banco de dados são mostradas na Tabela 10.3. Um exemplo ilustrando os dois tipos de simplificações é mostrado na Figura 10.1.

Tabela 10.2 – Operações de simplificação em função do número de sentenças antes (A) e depois (D) da operação correspondente. O símbolo (+) representa uma ou mais sentenças.

Operação	A	D
Reescrita	1	1
Sujeito-Verbo-Objeto	1	1
Mudança de voz	1	1
Inversão de ordem	1	1
Divisão	1	+
Junção	+	1
Remoção	1	0
Remoção parcial	1	1
Substituição léxica	1	1

Tabela 10.3 – Estatística do *corpus* de simplificação, incluindo o número médio de palavras, parágrafos e caracteres nos textos originais e simplificados (simplificação natural ou forte).

<i>Corpus</i>	Palavras	Parágrafos	Caracteres
Original	350,47	8,01	1.725,16
Natural	362,18	8,36	1.774,47
Forte	366,79	8,42	1.802,81

EUA enfrentam nova denúncia de massacre	EUA enfrentam nova denúncia de massacre	EUA enfrentam nova denúncia de massacre
<p>Para complicar ainda mais a situação dos militares dos EUA no Iraque, uma nova denúncia de massacre de civis por suas tropas veio à tona ontem.</p> <p>Um vídeo gravado após uma operação militar dos EUA em 15 de março na vila de Ishaqi, 80 quilômetros ao norte de Bagdá, mostra o resultado da operação: corpos de crianças e adultos iraquianos todos furados de balas. Alguns corpos têm tiros na cabeça. A vila de Ishaqi fica 80 quilômetros ao norte de Bagdá.</p>	<p>Uma nova denúncia de matança de civis por tropas dos EUA surgiu ontem, para complicar ainda mais a situação dos militares americanos no Iraque.</p> <p>Um vídeo gravado após uma operação militar dos EUA em 15 de março na vila de Ishaqi mostra o resultado da operação: corpos de crianças e adultos iraquianos todos furados de balas. Alguns corpos têm tiros na cabeça. A vila de Ishaqi fica 80 quilômetros ao norte de Bagdá.</p>	<p>Uma nova denúncia de matança de civis por tropas dos EUA surgiu ontem. A nova denúncia complica ainda mais a situação dos militares americanos no Iraque.</p> <p>Um vídeo gravado após uma operação militar dos EUA em 15 de março na vila de Ishaqi mostra o resultado da operação: corpos de crianças e adultos iraquianos todos furados de balas. Alguns corpos têm tiros na cabeça. A vila de Ishaqi fica a 80 quilômetros ao norte de Bagdá.</p>
(A)	(B)	(C)

Figura 10.1 – Exemplo de um extrato de texto (a) original; e suas respectivas simplificações (b) natural e (c) forte. Note que a diferença entre a simplificação natural e forte se concentra no nível de simplificação aplicado ao texto.

10.3 Diferenças entre textos originais e simplificados

Para averiguar a mudança na topologia induzida pelo processo de simplificação, calculamos várias medidas para as redes dirigidas e ponderadas. Para cada medida X , computamos a diferença relativa $dX = (X^{(s)} - X^{(o)})/X^{(o)}$ entre os valores obtidos para o texto simplificado ($X^{(s)}$) e os valores obtidos para o texto original ($X^{(o)}$). Significativamente, verificamos que o *strength* de saída apresenta uma tendência em aumentar seu valor quando um texto é simplificado, enquanto que os caminhos mínimos tendem a diminuir, conforme ilustra a Figura 10.2. As mudanças locais na topologia da rede foram analisadas graficando-se as medidas locais extraídas dos textos originais e das versões simplificadas correspondentes, vértice a vértice, de forma que cada medida distinta fosse representada em um *scatter plot*. Em outras palavras, para uma dada métrica X do vértice v , $X_v^{(o)}$ é o valor de X obtido para v na rede original e $X_v^{(s)}$ é o valor de X para o mesmo vértice na rede do texto simplificado. Neste caso, o ponto $(X_v^{(o)}, X_v^{(s)})$ pertence ao gráfico analisado. Dois descritores foram extraídos de cada *scatter plot*: o coeficiente angular e o coeficiente de correlação de Pearson, obtidos da melhor reta aproximando os dados. Estes descritores são importantes para fornecer informação a respeito da possível preservação de topologia. Por exemplo, se uma medida tende a se preservar os gráficos de dispersão se

Figura 10.2 – Distribuição da diferença relativa dX para (a) $X = s$; (b) $X = L$; e (c) $X = L$ (calculado com pesos complementares); para simplificações naturais e (d) $X = s$; (e) $X = L$; e (f) $X = L$ (calculado com pesos complementares); para simplificações fortes. Enquanto s tende a aumentar, o comprimento médio dos caminhos mínimos tendem a diminuir quando textos são simplificados.

aproximarão de uma linha rede com ambos os coeficientes próximos de 1 (ver exemplo na Figura 10.3). A partir da análise da distribuição destes coeficientes, foi possível identificar um padrão similar ao padrão encontrado para a análise de métricas globais. Para as medidas $X = \{s, L\}$, a correlação se distribui ao redor do valor unitário. Portanto, a distribuição de pontos no gráfico de dispersão se aproxima de uma linha reta. Com relação ao coeficiente angular, $X = s$ mostrou uma correlação predominantemente acima do valor unitário, enquanto que $X = L$ apresentou um histograma com valores predominantemente abaixo de 1, conforme ilustra a Figura 10.4. Portanto, o processo de simplificação tende a aumentar o *strength* local. Consequentemente, as distâncias geodésicas diminuem.

Os padrões encontrados nas mudanças globais e locais de redes complexas indicam que quanto maior o nível de simplificação, maior é a tendência de conectividade dos conceitos. Além disso, em textos simplificados, as distâncias separando conceitos é menor. Estas descobertas são consistentes com a expectativa de que em um texto simplificado as mesmas palavras devem aparecer mais vezes, pois menos palavras intermediárias são utilizadas

Figura 10.3 – Correlação de métricas locais para a medida de comprimento médio de caminhos mínimos. Os três gráficos de dispersão indicam que $X = L$ é preservado pois há um mapeamento um a um entre vértices correspondentes de redes distintas. O coeficiente de correlação de Pearson r e o coeficiente angular m obtido em cada caso foram: (a) $r = 0,988$ e $m = 1,013$; (b) $r = 0,998$ e $m = 1,003$; (c) $r = 0,998$ e $m = 1,001$. Note que, nos três casos, ambos os coeficientes assumem valores próximos do valor unitário.

Figura 10.4 – Histograma do coeficiente angular para (a) *strength* (simplificação natural); (b) caminhos mínimos (simplificação natural); (c) *strength* (simplificação forte); (d) caminhos mínimos (simplificação forte). Enquanto o *strength* tende a aumentar conforme os textos sofrem maiores simplificações, os caminhos mínimos diminuem.

Tabela 10.4 – Porcentagem de pares de textos cujo coeficiente de variação é maior para os textos originais do que para as simplificações natural e forte. Em seis destas nove medidas as redes obtidas para as simplificações são mais regulares. O p -value corresponde à probabilidade da porcentagem observada ser obtida aleatoriamente. A última coluna indica qual tipo de texto apresentou uma estrutura mais regular: o texto original (não simplificação) ou sua versão simplificada (natural ou forte).

X	Natural	p -value	Forte	p -value	Mais regular
acc	53.98%	2.3×10^{-1}	61.94%	7.0×10^{-3}	Simplificado
acc ⁽²⁾	75.22%	3.7×10^{-8}	80.53%	1.8×10^{-11}	Simplificado
acc ⁽³⁾	73.45%	3.0×10^{-7}	77.88%	1.0×10^{-9}	Simplificado
s	14.16%	1.2×10^{-15}	7.96%	1.0×10^{-15}	Original
$s^{(2)}$	79.64%	7.4×10^{-11}	86.72%	1.0×10^{-15}	Simplificado
C	79.64%	7.4×10^{-11}	81.41%	4.4×10^{-12}	Simplificado
$C^{(2)}$	13.27%	1.0×10^{-15}	16.81%	2.0×10^{-13}	Original
$C^{(3)}$	73.45%	3.0×10^{-7}	71.68%	2.3×10^{-6}	Simplificado

para manter um vocabulário simples e conciso.

10.4 Simplicidade textual e regularidade topológica

Uma correlação entre complexidade textual e a complexidade de redes pode ser obtida através da investigação de como as métricas variam durante o processo de simplificação. Examinamos a hipótese de que textos simplificados induzem uma estrutura de rede que também é simples, de forma que a estrutura ao redor dos vértices seja estável. A variação de cada medida de rede foi calculada com o coeficiente de variação v de cada medida. Basicamente, v é uma medida normalizada de dispersão de uma distribuição de probabilidade que fornece altos valores para distribuições irregulares. Primeiramente, analisamos cada medida individualmente, identificando as medidas que são capazes de distinguir simplificações de textos originais de acordo com o coeficiente de variação das medidas. Para cada par (original-simplificado), o coeficiente de variação v foi calculado para avaliar a fração de textos na qual existe um padrão revelando que v é maior no texto original do que no texto simplificado. A porcentagem de casos no qual houve uma predominância de valores maiores de v para uma dada classe de textos é mostrada na Tabela 10.4. Note que, em 6 de 9 casos, os padrões de textos simplificados são mais regulares, já que v foi maior em textos originais. Também, a acessibilidade acc⁽²⁾ de segundo nível, o *strength* de segundo e terceiro níveis ($s^{(2)}$ e $s^{(3)}$) e o coeficiente de aglomeração são particularmente importantes já que estas medidas possuem alta tendência de se estabilizar em textos simplificados, especialmente quando simplificações fortes são consideradas.

A análise de regularidade topológica também foi realizada usando uma abordagem multivariada. Cada vértice representando uma palavra no texto foi caracterizado com

Figura 10.5 – Análise de componentes principais de uma rede. Cada ponto representa um vértice da rede e a distância entre vértices é proporcional à dissimilaridade entre eles.

diversas métricas, que foram projetadas em um plano bidimensional usando PCA (ver Figura 10.5). Decidimos projetar os dados porque este procedimento reduz a complexidade do problema já que menos atributos são considerados no particionamento. A Figura 10.5 ilustra a projeção obtida para uma rede fictícia, onde cada ponto representa um vértice distinto. Os dois grupos evidentes são delimitados por linhas pontilhadas. Para identificar os grupos, a distribuição de pontos no plano foi estimada com o método de Parzen, que associa uma função gaussiana para cada ponto no espaço bidimensional. Então, a função densidade de probabilidade foi gerada a partir da soma da contribuição de cada gaussiana no espaço considerado. Os desvios padrões adotados foram $\sigma_{xx} = \sigma_{yy} = 0.5$. Os picos da distribuição (ver Figura 10.6) foram então identificados como máximos locais e associados a grupos distintos, conforme ilustra a Figura 10.5.

O número de grupos induzido por cada texto foi tomado como medida de regularidade das redes. Este procedimento foi adotado porque redes contendo vários grupos tendem a apresentar vértices com características bastante variadas, enquanto que redes contendo poucos grupos são caracterizadas por poucos padrões de conectividade, dando origem a uma topologia mais regular. Seguindo este raciocínio, testamos a hipótese de que textos simplificados apresentam a tendência de gerar menos grupos, já que assumimos que textos simplificados geram redes mais regulares. Para comparar o número de grupos, dois métodos foram empregados. No primeiro, o número absoluto de grupos ζ_{abs} foi tomado. No segundo, consideramos o número normalizado de grupos $\zeta_{\text{norm}} = (\zeta_{\text{abs}} - \langle \zeta_{\text{rand}} \rangle) / \sigma_{\zeta_{\text{rand}}}$, onde ζ_{abs} foi

Figura 10.6 – Estimação da densidade de probabilidades da distribuição da Figura 10.5 usando o método de janelas de Parzen. Cada máximo local é associado a um *cluster* na identificação de agrupamentos não-supervisionado.

normalizado pela média $\langle \zeta_{\text{rand}} \rangle$ e pelo desvio padrão $\sigma_{\zeta_{\text{rand}}}$ do número de grupos obtidos em 20 redes aleatórias equivalentes (normalização *z-score*).

A distribuição da diferença relativa do número absoluto e normalizado de grupos entre textos originais e simplificados mostrou que $\zeta_{\text{abs}} > 0$ em 72% dos textos ($p\text{-value} = 2.3 \times 10^{-6}$) e $\zeta_{\text{norm}} > 0$ em 73% ($p\text{-value} = 9.0 \times 10^{-7}$) (ver Figura 10.7). Portanto, textos originais (não simplificados) tendem a apresentar vértices aglomerados em uma maior quantidade de grupos em relação aos textos simplificados. Em outras palavras, textos simplificados tendem a gerar menos padrões de conectividade, o que sugere que a simplicidade linguística é mapeada em simplicidade topológica. Note que os padrões similares encontrados para ambos ζ_{abs} e ζ_{norm} confirmam que nossa análise não foi afetada pelo tamanho das redes.

10.5 Classificação de textos complexos

Para distinguir entre textos originais e simplificados, as medidas topológicas foram calculadas para os textos do banco de dados. Além da abordagem tradicional, excepcionalmente, nesta tarefa, ligamos as palavras com distância de 1 até 6 no texto para aperfeiçoar a identificação de ligações sintáticas mais distantes. Os dados extraídos de textos originais

Figura 10.7 – Diferença no número de *clusters* obtidos em textos originais e respectivas simplificações considerando (a) valores absolutos (normalizados pelo valor máximo); e (b) o *z-score* relativo. Em ambos os casos, os histogramas estão deslocados para o semi-eixo positivo. Estes resultados sugerem que textos originais possuem mais *clusters* que textos simplificados e portanto a topologia de textos não simplificados é mais irregular.

e simplificados foram projetados para identificar as diferentes classes. Os resultados estão ilustrados nas Figuras 10.8 e 10.9. Interessantemente, a primeira variável canônica já foi suficiente para separar as duas classes. Isto significa que é possível discriminar textos de acordo com sua complexidade usando apenas uma combinação linear de métricas de redes. A habilidade para separar os dois tipos de textos foi maior quando simplificações fortes foram consideradas e maiores comprimentos de janelas na formação das redes foram utilizadas. Portanto, a modelagem de vizinhos imediatos não é completamente capaz de capturar relacionamentos distantes que caracterizam a complexidade textual. Na verdade, nossos resultados sugerem que a complexidade tem relação com a distância entre conceitos na rede.

A capacidade para distinguir textos originais de simplificados foi quantificada usando todas as medidas topológicas como atributos para descrever os textos, ao invés de apenas as duas variáveis utilizadas na análise canônica. As Tabelas 10.5 e 10.6 fornecem os resultados das classificações original-natural e original-forte, respectivamente. O *p*-value representa a probabilidade de se obter a mesma taxa de acerto em classificações totalmente aleatórias. De modo análogo ao observado nas projeções, é mais fácil identificar as simplificações fortes. Novamente, as maiores taxas de acerto não ocorreram nas vizinhanças mais próximas. Os resultados obtidos com rede de adjacência de adjacência $\mathcal{J}$ palavras foram comparados com modelos mais sofisticados de rede, onde as ligações são estabelecidas de acordo com as relações de dependência sintática (166). Os resultados são fornecidos nas Tabelas 10.8 e 10.9. Interessantemente, os esforços adicionais para realizar o *parsing* não são úteis, já que as taxas de acerto para os modelos sintáticos são menores que as taxas de acerto obtidas

Figura 10.8 – Análise de variáveis canônicas para textos originais e simplificações naturais considerando as seguintes janelas de ligação (a) $\mathcal{J} = 1$; (b) $\mathcal{J} = 2$; (c) $\mathcal{J} = 3$; (d) $\mathcal{J} = 4$; (e) $\mathcal{J} = 5$; e (f) $\mathcal{J} = 6$. Uma melhor distinção foi obtida quando janelas mais longas foram consideradas.

para os modelos baseados em vizinhança. Portanto, para este tipo de aplicação, redes sintáticas não são úteis para caracterizar a complexidade textual.

A fim de comparar o nosso método de identificação de textos simplificados com métodos que fazem uso de recursos linguísticos profundos, caracterizamos cada texto do *corpus* com um conjunto de 15 atributos, de acordo com (165). A lista de atributos utilizada está ilustrada na Tabela 10.7. Os resultados da separação obtida entre os pares original-natural e original-forte estão ilustrados nas Tabelas 10.8 e 10.9, respectivamente. Como esperado em casos onde há disponibilidade de recursos linguísticos mais profundos, uma taxa de acerto superior foi obtida em relação ao método baseado em redes complexas. Apesar disso, a diferença em desempenho é de aproximadamente 10%. Portanto, o uso de métricas de redes complexas *apenas* não representa a melhor abordagem para identificar a complexidade. Apesar disso, nossa estratégia ainda assim mantém sua utilidade prática, principalmente em ocasiões em que recursos linguísticos caros não estão disponíveis. Nossa estratégia poderia também ser combinada com atributos linguísticos para aperfeiçoar a discriminabilidade dos dados.

Figura 10.9 – Análise de variáveis canônicas para textos originais e simplificações fortes considerando as seguintes janelas de ligação (a) $\mathcal{J} = 1$; (b) $\mathcal{J} = 2$; (c) $\mathcal{J} = 3$; (d) $\mathcal{J} = 4$; (e) $\mathcal{J} = 5$; e (f) $\mathcal{J} = 6$. Uma melhor distinção foi obtida quando janelas mais longas foram consideradas.

10.6 Sumarização extrativa

Uma das possibilidades de simplificação textual é através da sumarização automática. Devido à vasta quantidade de informação eletrônica disponível atualmente, a tarefa de sumarização extrativa automática se tornou uma das ferramentas mais populares em processamento de línguas naturais. Basicamente, esta tarefa está relacionada com a produção de um texto reduzido em conteúdo através da seleção de fragmentos (essencialmente sentenças) provenientes de um texto longo (167). Dessa forma, o texto original é comprimido de forma a manter a informação mais relevante. De modo análogo à outras aplicações em processamento de línguas naturais, a sumarização extrativa tem se beneficiado do uso de grafos na representação de conceitos. Por exemplo, Antiqueira et al. (168) mostrou que o uso da representação de grafos pode levar a bons resultados de informatividade, mesmo

Tabela 10.5 – Taxa de acerto para os pares de textos originais e simplificações fortes. Todas as classificações são estatisticamente significativas, o que sugere uma forte correlação entre a complexidade dos textos e as métricas topológicas das redes complexas. $\mathcal{J}$ representa o tamanho da janela de ligação e as taxas de acerto referem-se às melhores taxas obtidas testando os vários classificadores supervisionados.

$\mathcal{J}$	Acerto	p -value
1	82.30%	2.0×10^{-13}
2	84.96%	7.3×10^{-15}
3	84.07%	4.0×10^{-14}
4	85.40%	3.5×10^{-15}
5	86.28%	1.2×10^{-15}
6	85.40%	3.5×10^{-15}

Tabela 10.6 – Taxa de acerto para os pares de textos originais e simplificações naturais. Todas as classificações são estatisticamente significativas, o que sugere uma forte correlação entre a complexidade dos textos e as métricas topológicas das redes complexas. $\mathcal{J}$ representa o tamanho da janela de ligação e as taxas de acerto referem-se às melhores taxas obtidas testando os vários classificadores supervisionados.

$\mathcal{J}$	Acerto	p -value
1	71.81%	2.3×10^{-6}
2	73.12%	3.0×10^{-7}
3	75.33%	3.6×10^{-8}
4	71.81%	2.3×10^{-6}
5	72.68%	9.0×10^{-7}
6	74.89%	3.8×10^{-8}

quando os recursos linguísticos são limitados. Nesta tese, estendemos o estudo conduzido em (168) propondo novas métricas e novos modelos para mostrar que as recentes medidas projetadas para analisar as redes são capazes também de identificar conceitos relevantes em textos escritos. Além disso, mostramos o efeito de se adicionar conhecimento linguístico mais profundo.

10.7 Sumarizadores extrativos automáticos

As métricas de redes complexas que formam a base dos sumarizadores extrativos são descritas a seguir:

10.7.1 Acessibilidade

Quatro métodos baseados na medida de acessibilidade foram utilizados, sendo diferenciados entre si pelo comprimento h da caminhada aleatória auto-excludente. Para

Tabela 10.7 – Lista de atributos linguísticos e símbolo ($\mathcal{S}$) correspondente.

$\mathcal{S}$	Atributo
a	Número de palavras
b	Número de sentenças
c	Número de sujeitos
d	Índice Flesch
e	Número de palavras antes do verbo principal
f	Número de sentenças por parágrafo
g	Número de sílabas por palavra
h	Número de conectivos causais positivos
i	Número de conectivos causais negativos
j	Número de conectivos lógicos positivos
k	Número de conectivos lógicos negativos
l	Indicidência de palavras de conteúdo
m	Taxa de ambiguidade verbal
n	Taxa de ambiguidade nominal

$h = \{1, 2, 3\}$, criamos os métodos ACC1, ACC2 e ACC3, respectivamente. Além destes métodos, criamos o método ACCM que leva em consideração a diversidade média para $h = \{1, 2, 3\}$. No primeiro passo, o número de sentenças do sumário é estimado de acordo com a taxa de compressão τ . Então, para cada vértice (palavra), calcula-se sua acessibilidade. Em seguida, uma ponderação ω é associada para cada sentença sent no conjunto de sentenças $\mathcal{S}$, de acordo com o peso de cada palavra $\rho \in \text{sent}$:

$$\omega(\text{sent}) = \frac{1}{\|\text{sent}\|} \sum_{\rho \in \text{sent}} \text{acc}^{(h)}(\rho). \quad (10.1)$$

Em seguida, os valores de ω são ordenados em ordem decrescente para determinar o limiar de corte como o valor na posição $(1 - \tau) \cdot \theta$, onde θ é o tamanho do texto original (em número de sentenças). Finalmente, as sentenças do texto original são verificadas uma a uma na mesma ordem em que aparecem no texto original. Se uma sentença apresenta ω acima do limiar, ela é adicionada ao sumário. Aqui assumimos que as sentenças com um alto valor de acessibilidade média provavelmente possuem conceitos-chave e portanto elas conferem um bom nível de informatividade ao sumário. Esta abordagem parece razoável já que em experimentos preliminares, alto níveis de diversidade aparecem correlacionados com conceitos chaves do texto.

10.7.2 Acessibilidade e *strength*

Neste método, identificado como ACCG, ambas as medidas de acessibilidade e *strength* são calculadas para cada palavra individualmente. Basicamente, tentamos tirar vantagem

Tabela 10.8 – Comparando métodos baseados em conhecimento linguístico com métodos baseados em redes complexas. Textos originais e suas simplificações naturais foram comparados. Os resultados referem-se à taxa de acerto obtida com o classificador C4.5. A combinação de atributos é detalhada pelas letras da Tabela 10.7.

Métodos	Acurácia
adefghijklmno	86.94%
abcdeefghijklmno	78.32%
abcdeefghijklno	75.92%
bcddefghijklmno	75.44%
Redes com 3 vizinhos	75.33%
abcdeefghjlmno	74.96%
Redes com 6 vizinhos	73.57%
Redes com 2 vizinhos	71.81%
Redes com 4 vizinhos	71.81%
Redes com 1 vizinho	71.37%
Redes com 5 vizinhos	68.72%
Redes sintáticas	58.14%
acdeefghjlmno	52.40%

do cálculo de centralidade usando ambas as medidas. Os mesmos passos iniciais descritos na Seção 10.7.1 foram utilizados para obter dois vetores ordenados de acordo com as medidas de acessibilidade e *strength*. Em seguida, verificamos a classificação de cada sentença nestes vetores para selecionar aquelas aparecendo nas melhores posições de ambos os vetores. Em outras palavras, escolhemos as palavras com os maiores valores de *strength* e acessibilidade. Dessa forma, o peso $\omega(\text{sent})$ é associado à sentença *sent*. Se *sent* é classificada na posição $p_1(\text{sent})$ no primeiro vetor (relativo à acessibilidade) e $p_2(\text{sent})$ no segundo vetor (relativo ao *strength*), então $\omega(\text{sent})$ será dado por

$$\omega(\text{sent}) = p_1(\text{sent}) + p_2(\text{sent}). \quad (10.2)$$

Finalmente, selecionamos as sentenças com os menores valores de ω , já que quanto menor é a ponderação, maior será a relevância. Por exemplo, se uma sentença é classificada na primeira e terceira posições e outra sentença é ordenada na segunda e quarta posições, a primeira é considerada como mais informativa pois está mais bem classificada nas duas listas.

10.7.3 Caminhos mínimos

Este método, identificado como SP, é similar ao método de acessibilidade. A diferença aqui é que empregamos a medida de caminhos mínimos para quantificar a relevância das sentenças. Um vértice v será considerado relevante quanto um caminho mínimo começando

Tabela 10.9 – Comparando métodos baseados em conhecimento linguístico com métodos baseados em redes complexas. Textos originais e suas simplificações fortes foram comparados. Os resultados referem-se à taxa de acerto obtida com o classificador C4.5. A combinação de atributos é detalhada pelas letras da Tabela 10.7.

Método	Acurácia
bcdeghijlmno	95.19 %
abcdfghjklno	91.82 %
abcefgghjklmo	90.38 %
abcdeghjklmno	89.90 %
Redes com 6 vizinhos	84.95 %
Redes com 3 vizinhos	82.74 %
Redes com 2 vizinhos	82.30 %
Redes com 1 vizinho	81.86 %
Redes com 4 vizinhos	81.41 %
Redes com 5 vizinhos	80.97 %
Redes sintáticas	75.23 %
adeefghjklmno	75.23 %
bdefghjklmno	70.19 %

em v é curto, isto é, qualquer outro vértice é facilmente alcançável a partir de v . Esta abordagem é justificada pelo ponto de vista de que vértices relativamente próximos a outros vértices podem representar *hubs* ou pelo menos podem estar próximos a *hubs*. Dessa forma, é provável que tais vértices representem conceitos com alta informatividade.

10.7.4 *Betweenness*

Como no método anterior, o método denominada BTW utiliza uma medida de centralidade (o *betweenness* B) para encontrar palavras-chave. De forma análoga a outros métodos tradicionais de centralidade, esta métrica é baseada no cálculo dos caminhos mínimos para classificar a relevância dos vértices. Neste caso, um vértice é candidato a palavra-chave se o seu *betweenness* é alto; isto é, se existem muitos caminhos mínimos passando por ele. Uma vez que vértices com os maiores valores de *betweenness* tendem a ser mais acessíveis por caminhos eficientes, este método busca selecionar como palavras-chave aquelas palavras com uma tendência de encurtar a distância entre conceitos. Após o cálculo da centralidade, as sentenças são escolhidas de modo similar ao método de acessibilidade, usando o *betweenness* ao invés da acessibilidade no cálculo de ω na equação 10.1.

10.7.5 Vulnerabilidade

Este método (VUL) usa a medida de vulnerabilidade localmente para estabelecer uma relação entre relevância e resiliência da rede. Assume-se que uma palavra é importante se sua remoção reduz a eficiência global da rede. Portanto, vértices que mantêm a estrutura da rede ou que representam pontos de articulação (169) assumirão altos valores de vulnerabilidade e portanto serão classificados como palavras-chave. Similarmente aos métodos descritos nas Seções 10.7.2 e 10.7.4, este método emprega ω para detectar a relevância de sentenças, de acordo com a vulnerabilidade calculada para cada palavra na sentença examinada.

10.8 Avaliação da qualidade de sumários

A avaliação da qualidade de sumarizações extrativas representa uma tarefa tão difícil quanto o próprio processo de sumarização. Embora existam vários métodos de avaliação, ainda não existe um consenso sobre o melhor método. Uma das dificuldades envolvidas é a necessidade de avaliar os sumários de acordo com vários critérios, incluindo informatividade, coerência, coesão, legibilidade, gramaticalidade e textualidade. Outra abordagem para a avaliação é o emprego de métricas automáticas, que são vantajosas no sentido de serem mais padronizadas. No entanto, tais métricas usualmente não permitem a avaliação de vários dos critérios citados anteriormente. Entre as medidas mais amplamente utilizadas estão a cobertura Γ e a precisão ς (170), definidas como a porcentagem de sentenças pertencentes ao sumário de referência que aparece no sumário gerado e a porcentagem de sentenças pertencentes ao sumário gerado que aparece no sumário de referência. A chamada medida-F também é muito popular, sendo definida como um fator de ponderação entre Γ e ς :

$$\mathcal{F} = \frac{2 \cdot \Gamma \cdot \varsigma}{\Gamma + \varsigma}. \quad (10.3)$$

Note que, de acordo com a equação 10.3, o valor da medida-F varia entre 0 e 1.

Nesta tese, as métricas utilizadas foram a precisão, a cobertura e a medida-F disponíveis no pacote de avaliação automática ROUGE (170). Esta avaliação é baseada na co-ocorrência de unidades (n-gramas) entre sumários criados automaticamente e os sumários de referência. Tais métricas foram escolhidas porque elas apresentam forte correlação com a avaliação humana (170). Os sumarizadores foram avaliados usando o *corpus* TEMARIO, que contém textos divididos em dois grupos: textos originais a partir do qual os sumários são construídos e textos de referência, que são úteis para avaliar o sumário gerado. O *corpus* foi construído especialmente para auxiliar a tarefa de sumarização automática e

contém mais de 60.000 palavras. Os textos originais compreendem um conjunto de 100 artigos de jornais, dos quais 60 pertencem ao jornal brasileiro Folha de São Paulo (versão online). Os outros 40 textos foram extraídos do Jornal do Brasil. Os temas abordados nos textos foram variados, pois eles foram selecionados de várias seções: notícias do mundo, editoriais e reportagens especiais (Folha de São Paulo), política e notícias internacionais (Jornal do Brasil). O gênero jornalístico foi escolhido nas análises devido à heterogeneidade lingüística presente em tais documentos. Outro fator que influenciou a escolha deste gênero foi sua ubiquidade em sistemas de avaliação de sumarização automática em competições internacionais, como a TAC (*Textual Analysis Conference*) (171).

10.9 Avaliação dos sumarizadores propostos

Os sumarizadores extrativos foram construídos utilizando ambos os modelos baseados em adjacência e ligações sintáticas e comparados em termos de qualidade, primeiro dentro de cada modelo. Adicionalmente, a correlação entre os os sumarizadores foi quantificada a fim de verificar semelhanças entre sumários gerados dentro de um único modelo. A Tabela 10.10 mostra os valores ROUGE-1 (1-grama) para o modelo convencional, enquanto que a Tabela 10.11 mostra os valores de Rouge-1 para o caso de ligações sintáticas. Ambas as tabelas apresentam pontuações mais altas para os métodos baseados nas métricas de acessibilidade ($acc^{(1)}$, $acc^{(2)}$ e $acc^{(3)}$), o que indica que esta medida não só é adequada para detectar bordas em redes complexas (37), mas também para identificar palavras-chave em textos informativos. Significativamente, vários dos métodos baseados na acessibilidade alcançaram valores de cobertura superiores aos melhores já alcançados (168) (o melhor valor de cobertura para os sumarizadores baseados em redes complexas derivadas em (168) foi $\Gamma = 0,5031$). Diferentemente, os métodos que empregaram a vulnerabilidade, os caminhos mínimos e o *betweenness* levaram a resultados piores, mesmo quando informações sintáticas foram utilizadas.

Tabela 10.10 – Valores de ROUGE-1 para os métodos de sumarização baseados em redes de adjacência.

Método	Cobertura	Precisão	Medida-F
ACC1	0,5085	0,3847	0,4305
ACC3	0,5058	0,379	0,4245
ACC2	0,5046	0,3810	0,4265
ACCM	0,5045	0,3796	0,4254
SP	0,5032	0,3761	0,4228
ACCG	0,4983	0,3942	0,4323
BTW	0,4954	0,3859	0,4268
VUL	0,4882	0,3766	0,4178

Tabela 10.11 – Valores de ROUGE-1 para os métodos de sumarização baseados em redes sintáticas.

Método	Cobertura	Precisão	Medida-F
ACC1	0,5089	0,3865	0,4318
ACC3	0,5073	0,3815	0,4272
ACC2	0,5051	0,3722	0,4210
SP	0,5047	0,3800	0,4253
ACCM	0,5041	0,3816	0,4265
ACCG	0,4989	0,3899	0,4304
BTW	0,4901	0,3834	0,4223
VUL	0,4887	0,3807	0,4202

Também vale a pena mencionar o fato de que o desempenho dos sumarizadores foi aperfeiçoado quando conhecimento lingüístico foi incorporado, mas este aumento foi muito pequeno. Ao adicionar informação sintática, os valores de cobertura foram menores do que os obtidos com os sumarizadores baseados em conhecimento lingüístico profundo (neste caso, a cobertura alcança valores iguais a $\Gamma = 0,58$ (172)) . Conclui-se, portanto, que dificilmente os métodos estatísticos serão aperfeiçoados. Deste modo, investigações futuras devem se basear em abordagens híbridas que utilizem tratamento lingüística sofisticado.

Em nossos experimentos, também examinamos a correlação entre sumarizadores para verificar se há redundância nas informações fornecidas pelos diferentes sistemas. A correlação entre a classificação das sentenças geradas foi calculada com o coeficiente de correlação de Spearman (173). Este coeficiente assume valor máximo (unitário) quando ambos os métodos para classificar a relevância de sentenças selecionam as mesmas sentenças. O valor mínimo é alcançado quando os *ranks* das sentenças dado por um método é o inverso do outro método (a sentença principal para um é a menos importante para outro e assim por diante). A correlação foi obtida (i) entre sumarizadores baseados na adjacência de palavras; (ii) entre sumarizadores baseados em redes de ligação sintática; e (iii) entre sumarizadores pertencentes a ambos os modelos anteriores. Os resultados são mostrados na Figura 10.10 e na Tabela 10.12. Uma inspeção visual da Figura 10.10 revela uma alta correlação entre os métodos baseados na acessibilidade (ACC1, ACC2, ACC3 e ACCM), bem como entre aqueles baseados em distâncias geodésicas (SP, BTW e VUL). O método ACCG não parece correlacionar-se fortemente com qualquer outro método, provavelmente porque é um método híbrido que considera a acessibilidade e as métricas geodésicas. Portanto, há na verdade três tipos de sumarizadores: a versão com fundamentação geodésica, a versão baseada na acessibilidade e a versão híbrida. Analisando-se a correlação entre os sumarizadores na Tabela 10.12, é possível observar que as mudanças introduzidas pelas “bordas” sintáticas são realmente pequenas, uma vez que todos os valores de correlação entre o método de adjacência e o método sintático foram superiores a 0,90.

Figura 10.10 – Coeficiente de Spearman entre os sumarizadores construídos com (a) o modelo de adjacência; (b) o modelo de ligação sintática.

Tabela 10.12 – Correlação de Spearman para os mesmos sumarizadores em modelos distintos.

Método	Spearman
BTW	$0,991 \pm 0,010$
VUL	$0,983 \pm 0,021$
ACC1	$0,982 \pm 0,012$
ACCM	$0,980 \pm 0,015$
ACC3	$0,973 \pm 0,022$
SP	$0,970 \pm 0,031$
ACCG	$0,958 \pm 0,025$

10.10 Identificação de palavras-chave

Motivados pelos excelentes resultados obtidos com a medida de acessibilidade na seção anterior, aplicamos esta métrica para determinar as palavras-chave de textos baseados num único tema. Especificamente, este experimento utilizou um conjunto de 300 redações produzidas por estudantes no ENEM (Exame Nacional do Ensino Médio), cujo tema foi baseado na seguinte questão: “O direito ao voto: como usar esta ferramenta para promover as mudanças sociais que o país necessita?”. Portanto, espera-se que os conceitos-chave sejam semanticamente relacionados com o processo de votação, eleições, candidatos etc.

A estratégia desenvolvida aqui identifica para cada texto e para cada valor de $h = \{1, 2, 3\}$ as palavras – i.e. os vértices – com os 10% maiores e menores valores de acessibilidade. Para cada texto, uma dada palavra pode aparecer no máximo h vezes dentro dos 10% maiores valores, já que os cálculos foram realizados para cada nível separada-

mente. Portanto, a frequência de cada palavra vai variar entre 0 e 900, já que a contagem é realizada no *corpus* de 300 redações. As palavras mais frequentes na lista de maior acessibilidade estão sumarizadas na Tabela 10.13, que mostra as palavras relacionadas ao tema e palavras com alta frequência, como os verbos “ser”, “ter”, “fazer” e “ir”. A Tabela 10.14 ilustra as palavras mais comuns na lista de menor diversidade. Note que a maioria das palavras possuem baixa correlação com o tema das redações, o que confirma que os vértices próximos à borda possuem baixa relevância.

Tabela 10.13 – Número de vezes (N^+) que uma dada palavra apareceu entre as 10% mais importantes segundo a medida de acessibilidade.

Palavra	N^+
ser	732
não	447
votar	407
ter	384
voto	379
poder	255
país	246
direito	246
dever	162
Brasil	162
fazer	132
pessoa	104
candidato	104
estar	90
ir	80
político	74
melhor	63
brasileiro	58
povo	58
saber	57

10.11 Conclusões

Neste capítulo descrevemos métodos para avaliar e gerar textos simplificados. Na análise, textos originais e textos simplificados puderam ser distinguidos através da representação de textos como redes complexas. As métricas que permitiram uma maior distinção foram o *strength*, os caminhos mínimos, a acessibilidade e as medidas hierárquicas. As melhores classificações alcançaram taxas de acerto iguais a 75% e 85%, respectivamente para os pares original-natural e original-forte. Acreditamos que a metodologia desenvolvida para a análise de complexidade possa ser especialmente útil para aplicações relacionadas à

Tabela 10.14 – Número de vezes (N^-) que uma dada palavra apareceu entre as 10% menos importantes segundo a medida de acessibilidade.

Palavra	N^-
ano	125
bem	99
cidadão	82
ajudar	74
terminar	72
bom	72
ainda	66
acreditar	64
fazer	62
pessoa	61
candidato	61
estar	60
ir	52
político	52
melhor	50
brasileiro	45

simplificação textual, onde a detecção de estruturas complexas é essencial para a posterior obtenção de textos concisos.

Talvez a principal implicação dos resultados da análise de complexidade não esteja relacionada com as taxas de acerto dos classificadores induzidos por técnicas de reconhecimento de padrões. Na verdade, os resultados mais importantes estão relacionados à emergência de dois padrões específicos em textos complexos: as longas distâncias de conexão de conceitos e os padrões heterogêneos caracterizados por um grande número de grupos topológicos na redes de textos não simplificados. Ambos os padrões são importantes porque eles estão correlacionados com os fatores cognitivos que tornam um texto complexo, tais como a quantidade de informação que precisa ser memorizada no processo de cognição (174).

Neste capítulo também utilizamos redes complexas para o desenvolvimento de sumarizadores extrativos. Através de testes exaustivos realizados sobre várias medidas, descobrimos que as medidas de vulnerabilidade, caminhos mínimos e *betweenness* apresentaram um desempenho regular na seleção de sentenças informativas. Por outro lado, as medidas de acessibilidade se mostraram bastante informativas, alcançando valores de ROUGE competitivos. Outra contribuição importante refere-se à introdução de informação linguística para o processamento textual. Como esperado, a adição de informação sintática foi capaz de aperfeiçoar o desempenho dos sistemas de sumarização, mas o aumento observado foi pequeno. Portanto, embora exista espaço para uma melhora no desempenho através da introdução de informações linguísticas profundas, é necessário ter em mente que uma melhora significativa pode necessitar de um tratamento sofisticado (p.e. com a introdução de informação semântica). Os bons resultados obtidos com a medida de acessibilidade

motivou sua aplicação na tarefa de detecção de palavras-chave em um pequeno conjunto de textos focado em um tema único. Os resultados desta análise mostraram que as palavras com alta diversidade estão relacionadas com palavras-chave. De maneira geral, este último resultado abre novas possibilidades para a aplicação de conceitos de redes complexas em tarefas de recuperação de informação pois medidas de acessibilidade foram pouco exploradas até o momento e portanto podem ser úteis em áreas que demandem o estabelecimento de taxonomias e/ou ontologias.

Capítulo 11

Modelagem de redes de citações

Avaliações quantitativas de pesquisadores, instituições, regiões geográficas, revistas científicas e áreas da ciência e tecnologia são bastante comuns atualmente, especialmente com a grande disponibilidade de informação em bancos de dados científicos. O número de citações e o fator de impacto estão entre os parâmetros mais comuns usados para avaliação. Consequentemente, tais índices representam um dos principais fatores na decisão de promoções, concessões e na identificação de tendências de pesquisas. Nos últimos anos, a ciência tem se guiado de certa forma pela cienciometria (175–177, 181, 182), que representa uma das principais motivações para estudar padrões de citações e outros parâmetros de interesse. Redes de citações, por exemplo, têm sido modeladas com conceitos e metodologias de redes complexas (183, 184). O grau de entrada destas redes (isto é, o número de citações recebidas pelos artigos científicos) segue uma lei livre de escala (185), o que equivale a dizer que a probabilidade de um artigo ser citado depende do seu número de citações (185). Algumas evidências apontam que a probabilidade de um artigo ser citado depende também de sua similaridade semântica com outros artigos científicos (186), do tempo decorrido desde a sua publicação (187, 188), da reputação dos autores, revistas e instituições, e até mesmo da língua materna dos autores, já que a qualidade de escrita influencia na qualidade de manuscritos científicos (176).

Devido à grande variedade de possíveis fatores, a modelagem de redes citações não tem sido simples. Métodos tradicionais que consideram apenas um fator por vez podem ser úteis para explicar a dinâmica deste fator. No entanto, tais modelos podem omitir padrões importantes que seriam obtidos caso mais fatores fossem considerados (189). O modelo de conexão preferencial, por exemplo, prevê o grau da distribuição das redes, mas não condiz com fatores ligados às características internas dos artigos científicos (189). Outros métodos são capazes de explicar também a distribuição de grau ou o coeficiente de aglomeração (190), mas não consideram o conteúdo dos artigos científicos em conjunção com fatores ligados à idade dos artigos. Portanto, na modelagem de redes de citações é importante considerar o maior número de fatores possíveis.

Neste capítulo fazemos uso das técnicas de classificação textual para modelar o processo de citação de acordo com fatores topológicos, semânticos e relativos à data de publicação dos manuscritos científicos. Devido à complexidade de se quantificar a reputação ou visibilidade de revistas científicas ou instituições, estes fatores não foram incluídos no modelo proposto. Alternativamente, desenvolvemos experimentos com redes artificiais contendo comunidades de autores com visibilidades distintas a fim de verificar a influência da visibilidade em índices de produtividade de autores (175, 191, 192).

11.1 Modelo de rede de citações

Nesta seção propomos um modelo para descrever alguns atributos de redes de citações. Três parâmetros são considerados: a topologia, a similaridade de conteúdo e a idade dos artigos científicos. Este modelo foi utilizado para gerar redes com citações sendo selecionadas de acordo com um destes três critérios. A similaridade de conteúdo foi calculada através da análise de conteúdo de uma coleção de textos do repositório arXiv em dois tópicos distintos: “*complex networks*” e “*graphenes*”, gerando as redes denominadas como CN e GF, respectivamente. Neste caso, cada artigo foi caracterizado pela frequência das palavras lematizadas após a eliminação das *stopwords* (ver Capítulo 3). Assumindo que a frequência das palavras nos artigos a e b são representados respectivamente pelos vetores $\vec{v}_a$ e $\vec{v}_b$, então a similaridade entre os dois documentos é dada por:

$$\text{sim}(a, b) = \frac{\vec{v}_a \cdot \vec{v}_b}{\|\vec{v}_a\| \cdot \|\vec{v}_b\|}. \quad (11.1)$$

Uma vez que $\text{sim}(a, b)$ quantifica o cosseno do ângulo formado entre os vetores $\vec{v}_a$ e $\vec{v}_b$, $0 \leq \text{sim}(a, b) \leq 1$. No conjunto de artigos do arXiv, a similaridade de conteúdo segue a distribuição

$$P(\text{sim}(a, b)) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(\text{sim}(a, b) - \mu)^2}{2\sigma^2}\right), \quad (11.2)$$

onde μ e σ representam a média e desvio da distribuição. Outro critério para selecionar citações refere-se à regra de conexão preferencial baseada no grau de entrada atual de um artigo científico. Desta forma, artigos com um grande número de citações são mais prováveis de serem citados novamente, de acordo com uma lei de potência

$$P(k) \propto k^{-\gamma_k}, \quad (11.3)$$

onde γ_k é calculado de acordo com a metodologia desenvolvida em (60). Com relação ao critério de idade do artigo, a probabilidade de um artigo ser citado é considerada como

inversamente proporcional à diferença Δt entre o tempo atual e a data de publicação do artigo. De acordo com a observação nos bancos de dados do arXiv, a função

$$P(\Delta t) \propto \Delta t^{-\gamma_t} \quad (11.4)$$

pode ser utilizada para caracterizar a probabilidade de um artigo ser citado Δt meses após sua data de publicação.

Considere X uma variável aleatória assumindo o valor $X = 1$ quando o artigo é citado e $X = 0$ caso contrário. Dada as distribuições de probabilidade definidas respectivamente nas equações 11.2, 11.3 e 11.4, definimos a quantidade $P(X = 1|k, \text{sim}, \Delta t)$ como a probabilidade de o artigo ser citado possuindo os parâmetros k , sim e Δt como:

$$P(X = 1|k, \text{sim}, \Delta t) = \frac{\alpha P(k) + \beta P(\text{sim}) + \lambda P(\Delta)}{\alpha + \beta + \lambda}, \quad (11.5)$$

onde $\lambda = 1 - \alpha - \beta$. Para verificar a adequabilidade do modelo usamos a seguinte função de erro:

$$\begin{aligned} \epsilon^2 = & \quad 1/6 \left(\frac{\lambda_{k,m} - \lambda_{k,r}}{\lambda_{k,r}} \right)^2 + 1/12 \left(\frac{\mu_{l,m} - \mu_{l,r}}{\mu_{l,r}} \right)^2 + 1/12 \left(\frac{s_{l,m} - s_{l,r}}{s_{l,r}} \right)^2 \\ & + \quad 1/6 \left(\frac{\mu_{\text{sim},m} - \mu_{\text{sim},r}}{\mu_{\text{sim},r}} \right)^2 + 1/6 \left(\frac{s_{\text{sim},m} - s_{\text{sim},r}}{s_{\text{sim},r}} \right)^2 + 1/3 \left(\frac{\gamma_{t,m} - \gamma_{t,r}}{\gamma_{t,r}} \right)^2, \end{aligned} \quad (11.6)$$

cujos parâmetros são resumidos na Tabela 11.1. Estes pesos foram distribuídos de forma a atribuir a mesma ponderação para os três fatores^a Uma vez que uma busca exaustiva é impraticável, utilizamos a estratégia conhecida como *simulated annealing* (193) para minimizar o erro ϵ^2 na equação 11.6.

As redes simuladas obtidas com apenas um dos critérios (topologia, semântica ou tempo) exibiram características que diferenciam consideravelmente das redes reais CN e GF extraídas do repositório arXiv. Este resultado é mostrado quantitativamente através da determinação do erro ϵ^2 na Tabela 11.2. Uma excelente concordância do modelo com as redes reais foi observada, no entanto, quando os três critérios foram combinados em um procedimento de otimização, como ilustram as Figuras 11.1 e 11.2 para as redes CN e GF, respectivamente. Com relação à ponderação atribuída a cada um dos fatores em ambos as redes, o fator predominante foi o fator semântico ($\beta = 52,5\%$ para CN e $\beta = 45,0\%$ para GF). O fator topológico foi o segundo fator mais proeminente na rede CN, enquanto que o fator temporal foi o segundo fator mais proeminente na rede GF. Embora a similaridade de conteúdo represente o fator mais importante, isto não significa que os autores estejam selecionando os artigos mais similares para serem citados. Isto pode ser

^a O total dos pesos atribuídos foram: $1/6 + 1/12 + 1/12 = 1/3$ para a topologia, $1/6 + 1/6 = 1/3$ para semântica e $1/3$ para o fator temporal.

Tabela 11.1

Símbolo	Descrição
$\gamma_{k,m}$	Coefficiente da lei de potência para o grau de entrada da rede obtida do modelo
$\gamma_{k,r}$	Coefficiente da lei de potência para o grau de entrada da rede real
$\mu_{l,m}$	Comprimento médio dos caminhos mínimos da rede obtida do modelo
$\mu_{l,r}$	Comprimento médio dos caminhos mínimos da rede real
$s_{l,r}$	Desvio padrão do comprimento dos caminhos mínimos da rede obtida a partir do modelo
$s_{l,r}$	Desvio padrão do comprimento dos caminhos mínimos da rede real
$\mu_{\sigma,m}$	Similaridade de conteúdo média no modelo
$\mu_{\sigma,r}$	Similaridade de conteúdo média nas redes reais
$s_{\sigma,m}$	Desvio padrão na similaridade de conteúdo no modelo
$s_{\sigma,r}$	Desvio padrão na similaridade de conteúdo nas redes reais
$\gamma_{t,m}$	Coefficiente da lei de potência para Δt nas redes do modelo
$\gamma_{t,r}$	Coefficiente da lei de potência para Δt nas redes reais

observado nas distribuições das Figuras 11.1(c) e 11.2(c), que mostram que poucos dos artigos citados são bastante similares. Este resultado também é consistente com a baixa correlação encontrada entre a lista de referências real e a lista de artigos mais similares em um banco de dados (186).

Tabela 11.2 – Estratégias de modelagem de redes de citações.

Rede	α	β	λ	$\epsilon^2/\epsilon_{min}^2$
Rede CN	100,0%	0,0%	0,0%	3,125
Rede CN	0,0%	100,0%	0,0%	2,104
Rede CN	0,0%	0,0%	100,0%	7,982
Rede GF	100,0%	0,0%	0,0%	2,617
Rede GF	0,0%	100,0%	0,0%	1,945
Rede GF	0,0%	0,0%	100,0%	3,445

Uma questão importante a ser observada é a necessidade de se empregar parâmetros distintos para reproduzir as características de redes reais. Isto indica que o modelo de três parâmetros não é universal, pois provavelmente não descreve todas as características das redes de citações. Esta limitação é, de certa forma, esperada pois outros critérios são importantes no processo de seleção de referências. Talvez o critério mais importante seja a reputação (ou visibilidade) de autores e revistas (194, 195), que está parcialmente (mas não inteiramente) implícito no grau de entrada incorporado no modelo. O critério de visibilidade não foi incluído no modelo proposto porque este tipo de informação não está prontamente disponível. Por exemplo, nem todos artigos no arXiv possuem informação a respeito da revista publicada (seja porque o artigo não foi publicado ou porque a informação foi simplesmente omitida). Desta forma é impossível usar o fator de impacto das revistas. Com relação às instituições, não existe nenhum índice bem estabelecido que é capaz de

Figura 11.1 – Distribuição para as redes CN nas redes reais (vermelho) e nas redes obtidas com o modelo (azul). As propriedades previstas pelo modelo incluem topologia das redes (grau de entrada em (a)) e o comprimento médio dos caminhos mínimos em (b)), semântica (similaridade de conteúdo entre os artigos e suas referências em (c)) e tempo.

Figura 11.2 – Distribuição para as redes GF nas redes reais (vermelho) e nas redes obtidas com o modelo (azul). As propriedades previstas pelo modelo incluem topologia das redes (grau de entrada em (a)) e o comprimento médio dos caminhos mínimos em (b)), semântica (similaridade de conteúdo entre os artigos e suas referências em (c)) e tempo.

quantificar a notoriedade ou reputação. Com relação aos autores, poderíamos ter feito uso do banco de dados *ISI Highly Cited* (196). Infelizmente, apenas um pequeno número de autores são listados neste banco de dados. O efeito da visibilidade nas redes de citações com relação ao índice-h é discutido na próxima seção.

11.2 Efeito da visibilidade na evolução do índice-h

A fim de analisar como a visibilidade interfere na dinâmica das redes de citações, estudamos a evolução do índice-h (175, 191, 192) dos autores pertencendo a duas comunidades artificiais com visibilidade distintas, assumindo que uma comunidade possui o dobro de visibilidade da outra comunidade. Quatro modelos foram considerados em termos de distribuição e visibilidade. Em todos os modelos, assumimos que o número de artigos publicados por um autor a cada ano segue uma lei de potência $p(y) = cy^{-\gamma}$, onde $p(y)$ representa a distribuição de probabilidade de y e c e γ são parâmetros reais. Estes valores foram determinados através da definição dos pontos $(y, p(y))$ da distribuição: $(1, m)$ e $(s, 1)$. Consequentemente, assumimos que m autores publicam um artigo por ano e apenas um autor publica s papers por anos. Com estes limites, $c = m$ e

$$\gamma = \frac{\log(m)}{\log(s)}. \quad (11.7)$$

Nos experimentos, assumimos $m = 15$ e $s = 30$. Note que, a partir de $p(y) = cy^{-\gamma}$, é necessário amostrar alguns valores de y . Sem perda de generalidade, selecionamos os seguintes valores: $y = (1, 2, 3, 5, 10, 15, 30)$. Com estes valores, obtemos que $p(y) = (15, 9, 6, 4, 2, 2, 1)$. Em outras palavras, 15 autores publicam 1 artigo por ano, 9 autores publicam 2 artigos por ano e assim por diante. Portanto, o número total de autores e o número total de artigos são dados respectivamente por $N_a = \sum p(y) = 39$ e $N_p = \sum yp(y) = 151$. Esta distribuição foi assumida para cada uma das comunidades (denominadas como comunidades A e B). Portanto, um total de 302 artigos foram produzidos por 78 autores.

No modelo proposto, novos artigos (vértices da rede) são criados dentro de um período de 25 anos. A distinção das duas comunidades A e B com relação à visibilidade (i.e., a probabilidade de receber novas citações) é feita através da definição dos índices de visibilidade f_A e f_B para as comunidades A e B, respectivamente. Assumimos que a comunidade A possui o dobro da visibilidade da comunidade B, isto é, $f_A = 2f_B$. Portanto, os artigos da comunidade A possuem o dobro de chance de serem citados por novos artigos em comparação com os artigos da comunidade B. Estes valores diferentes de visibilidade foram adotados para simular as diferenças provenientes dos fatores como impactos das revistas, instituição dos autores, dentro outros (176). O crescimento de cada rede foi

obtido a cada ano. Os modelos utilizados nas simulações foram:

1. **Modelo UNI:** trata-se de um modelo uniforme, com seleção de artigos de forma aleatória.
2. **Modelo PREF:** seleção preferencial dos artigos segundo a visibilidade dos artigos.
3. **Modelo PREFC:** seleção preferencial de artigos com maior número de citações.
4. **Modelo DBPREF:** seleção preferencial dos artigos com alta visibilidade e alto número de citações.

No modelo UNI, cada artigo incluído cita w outros artigos selecionados aleatoriamente dentre os artigos publicados. Analogamente, no modelo PREF, artigos selecionados aleatoriamente são citados. No entanto, a visibilidade é considerada. O modelo PREFC também é preferencial, mas cada uma das w citações de cada artigo é escolhida de forma a dar preferência aos artigos mais citados (as citações são contabilizadas do primeiro ano até o ano atual). Portanto, este modelo é similar ao modelo BA (ver Capítulo 2), diferindo apenas no fato de que os graus de entradas são atualizados ao final de cada ano e não após a inclusão de cada novo artigo. Finalmente, o modelo DBPREF é preferencial a ambos os termos de visibilidade e grau de entrada (número de citações). Mais especificamente, uma lista é mantida onde o número total de entradas na lista do artigo é proporcional ao número de citações vezes a visibilidade. Novas citações são escolhidas aleatoriamente através de uma seleção uniforme nesta lista.

Cada uma das configurações foi realizada 20 vezes a fim de fornecer representatividade estatística. O índice- h e o número total de citações foram computados para cada autor a cada ano. Os resultados na Figura 11.3 indicam que a inclusão do modelo preferencial (PREF) com relação à visibilidade apenas exerce pouco efeito para a comunidade A. No entanto, um grande efeito foi observado na comunidade B. De fato, o índice- h de todos os autores na comunidade B cresceu com uma menor taxa e, após 25 anos, esses índices- h assumiram valores consideravelmente menores do que os índices dos autores na comunidade A. Essas observações se aplicam para $w = \{5, 20\}$, embora os índices- h para $w = 20$ sejam maiores.

Com relação à importância do número total de citações, a Figura 11.4 evidencia um pequeno aumento no índice- h para $w = 5$ em comparação com o modelo UNI (Figura 11.3a). Por outro lado, os valores de índice- h são muito menores quando a regra de ligação preferencial é aplicada para $w = 20$ na Figura 11.4d em comparação com o caso aleatório (modelo UNI na Figura 11.3d). Quando a visibilidade e o grau são considerados (DBPREF), observa-se um pequeno aumento no índice- h para a comunidade A, enquanto os efeitos para a comunidade B são bastante intensos. Para os autores desta última comunidade, os valores de índice- h são muito menores. A única exceção parece ser o autor com o maior

número de artigos para $w = 20$. Por alguma razão, existe um efeito de compensação neste caso, e o índice-h para este autor não é muito menor. Note também que através da aplicação da conexão preferencial baseada no grau de entrada (para PREFC e DBPREF), a distribuição assimétrica de citações entre artigos fez com que os índices-h assumissem valores consideravelmente menores do que no modelo UNI ou PREF, para um número fixo de referências.

Figura 11.3 – Dinâmica do índice-h para (a) modelo UNI com $w = 5$; (b) modelo PREF para comunidade A com $w = 5$; (c) modelo PREF para comunidade B com $w = 5$; (d) modelo UNI para $w = 20$; (e) modelo PREF para comunidade A com $w = 20$; e (f) modelo PREF para comunidade B com $w = 20$.

11.3 Conclusões

O uso combinado de três fatores distintos (similaridade de conteúdo, grau de entrada e data de publicação) foi útil para modelar e reproduzir diversas características topológicas de redes de citação. O modelo proposto, portanto, representa um progresso considerável na descrição da dinâmica de redes de citação. O modelo com três atributos foi aplicado a duas redes reais obtidas do repositório arXiv para os tópicos “*graphenes*” e “*complex networks*”. A importância relativa de cada parâmetro nestas duas redes não foi o mesmo.

Figura 11.4 – Dinâmica do índice-h usando (a) modelo PREFC com $w = 5$; (b) modelo DBPREF para a comunidade A com $w = 5$; (c) modelo DBPREF para comunidade B com $w = 5$; (d) modelo PREFC com $w = 20$; (e) modelo DBPREF para a comunidade A com $w = 20$; (f) modelo DBPREF para a comunidade B com $w = 20$

Enquanto a similaridade de conteúdo foi o fator mais relevante para ambas as redes, os fatores de tempo e número atual de citações variou de rede pra rede. Esta dependência com a rede provavelmente está relacionado ao fato da dependência do modelo com outros fatores, já que mesmo artigo bastante similares podem ser esquecidos (186). De fato, a reputação dos autores e instituições devem desempenhar um papel fundamental nestes modelos. O efeito da reputação é parcialmente levado em consideração quando o grau de entrada é considerado, porque artigos científicos com mais citações possuem uma tendência maior de receber novas citações. No entanto, este efeito é uma manifestação indireta da reputação, que não considera a maior visibilidade obtida por manuscritos escritos por pesquisadores renomados, principalmente quando o número de citações atuais é pequena.

Devido a importância do fator de visibilidade (ou reputação), decidimos verificar o efeito de tal fator na evolução do índice-h de autores considerando redes de citações artificiais. Mostramos que a comunidade com maior visibilidade (com a maior probabilidade de ter um manuscrito citado) beneficia apenas marginalmente – em termos do índice-h – em comparação com uma rede de citação de controle com visibilidade uniforme. Este aumento do índice-h de autores proeminentes provavelmente ocorre porque esta medida

segue uma distribuição *Lotkaian* (197) e, portanto, o efeito de concentração pode ser um efeito de reforço (197). Por outro lado, as comunidades com menor visibilidade podem ser duramente afetadas, já que seus valores de índice-h podem ser consideravelmente menores do que os estimados para uma rede de controle imparcial. Esta descoberta confirma a observação em redes reais de que os valores de índice-h dependem da produtividade e das práticas de citação de determinadas áreas de pesquisa (198). Portanto, o uso do índice-h para comparar autores de comunidades distintas não é aconselhável.

Capítulo 12

Cienciometria virtual

O avanço do conhecimento depende fundamentalmente da ampla divulgação de novas abordagens e resultados, o que permite que outros cientistas e profissionais possam analisar os resultados, validando e complementando as investigações. O objetivo de qualquer manuscrito científico, além de certificar/autenticar o conhecimento (199), é ser *lido*, *validado* e *citado* pelo maior número de pessoas possível (200). Enquanto artigos têm sido avaliados em termos do número de citações recebidas, as revistas têm sido normalmente classificadas de acordo com os fatores de impacto que refletem o número de citações de seus artigos. O número de citações tem sido um fator importante desde 1920 (201), sendo um dos objetos de estudo da cienciometria, i.e., a ciência que investiga a relação entre as áreas do conhecimento e a evolução da ciência (202–207). Obviamente, o sucesso de um manuscrito científico depende de vários fatores, incluindo a notoriedade da revista e eminência dos autores e suas instituições. Estritamente falando, tal sucesso deveria depender não apenas da qualidade, originalidade, integridade e clareza de um manuscrito, mas também do grau de relação e sobreposição com outras investigações na literatura. Todos os manuscritos fortemente relacionados com o manuscrito sendo escrito deveriam ser lidos e potencialmente citados. No entanto, devido à disponibilidade limitada de pesquisadores para realizar uma busca bibliográfica completa, alguns trabalhos relacionados são filtrados de acordo com critérios pessoais. Devido à esta imperfeição na busca bibliográfica, algumas publicações importantes inevitavelmente são negligenciadas (208, 209), comprometendo assim a eficiência de todo o sistema, no sentido de que os esforços gastos em buscas bibliográficas tendem a ser repetidos e/ou ignorados.

Neste capítulo usamos a tarefa de classificação textual (por similaridade semântica) para auxiliar o procedimento de seleção adequada de uma lista de referência para artigos científicos. Analisamos dois critérios para avaliar a qualidade das citações selecionadas em manuscritos científicos. O primeiro critério afirma que trabalhos fortemente relacionados deveriam ser lidos e, potencialmente, citados. Já o segundo critério sugere uma busca sistemática sobre publicações relacionadas e suas citações/referências. Em especial, verificamos

se esses dois critérios são seguidos usando uma metodologia baseada em redes complexas e formalismos de processamento de línguas naturais. Dois conjuntos de artigos contendo 900 artigos cada foram usados para obter duas redes para cada área analisada: (i) as redes de citação tradicionais, onde cada artigo é um vértice e as citações são representadas por arestas dirigidas, e (ii) uma rede de similaridade de conteúdo de pares de artigos. Estas redes são denominadas como *citação* e *similaridade*, sendo a primeira dirigida e a segunda não dirigida. Para calcular a similaridade entre os artigos, foram utilizados o conteúdo dos artigos postados no repositório arXiv (210). Para recuperar as citações dos artigos, o repositório do arXiv e a base de dados *Web of Science* (211) foram utilizados. Com relação à rede de similaridades, cada artigo foi modelado como uma rede complexa a fim de quantificar a similaridade entre pares de artigos. Vários índices de similaridades tradicionais foram utilizados para mostrar que a análise realizada é fracamente dependente do índice empregado. Depois de definir estas duas redes, quantificamos o número de: (a) artigos que estão relacionados e citados, (b) artigos que estão relacionados, mas não citados, e (c) artigos que são pouco relacionados, mas citados. Vamos mostrar que a análise desses números indica que o critério de similaridade para a seleção de referências não é obedecido. A partir de uma caminhada aleatória nas redes de citação para simular uma busca sistemática por um autor, inferimos que o segundo critério também não é obedecido. Além de discutir as possíveis causas e implicações desses resultados, neste capítulo sugerimos uma abordagem de citações virtuais, onde citações são artificialmente criadas de acordo com a semelhança de conteúdo entre documentos e referências. Este modelo pretende complementar o processo de citação manual custosa, dando origem a uma cienciometria virtual.

12.1 Redes de citação e similaridade

Nos experimentos, as relações (similaridade ou citação) entre dois artigos são modeladas como redes. As redes foram construídas usando um *corpus* contendo 900 artigos sobre “*complex networks*” e 900 artigos sobre “*graphenes*”, os quais foram obtidos do repositório arXiv. Os primeiros 900 artigos retornados pela busca no banco de dados do arXiv foram selecionados para compor nosso banco de dados. As palavras-chave empregadas na busca foram “*complex networks*”, “*scale-free networks*” e “*graphenes*”. Apenas a versão mais atual dos manuscritos foram consideradas.

A rede de citação $W^{(cit)}$, que pode ser considerada como uma rede social acíclica modificada (o relacionamento entre dois indivíduos é substituído pelo relacionamento entre artigos) (212) e segue uma lei de potência, armazena toda a informação a respeito das citações entre artigos, onde cada artigo é um vértice e cada aresta representa uma

citação. Estas redes têm atraído o interesse de muitos pesquisadores. Por exemplo, índices imparciais foram utilizados para quantificar a relevância de artigos científicos (215) e redes de citações foram empregadas para revelar padrões de conectividade entre autores e redes de co-autoria (214). Na sua forma mais simples, uma rede de citação é construída da seguinte maneira. Se o artigo i cita o artigo j , então existirá uma aresta direcionada representada como $i \rightarrow j$. Para construir a rede de citações, além do identificador do arXiv, coletamos na lista de referência o título dos artigos. Isto foi feito porque os artigos já publicados não são citados com o identificador do arXiv. Dessa forma, usando a relação entre título e identificador do arXiv, fomos capazes de recuperar a maioria das citações sem identificadores, mas que de fato pertencem ao banco de dados do arXiv. Além disso, usamos o banco de dados *Web of Science* para recuperar as citações de forma robusta. Não consideramos as mudanças realizadas em um documento após sua publicação no arXiv e aparecimento na revista, já que usamos o *Web of Science* apenas para recuperar referências adicionais. Para ilustrar uma rede de citações, a Figura 12.1 mostra a rede obtida após a análise da lista de referência de dois artigos, cujos identificadores do arXiv são 0902.3068 e 0805.2298.

Figura 12.1 – Visualização de uma subrede de citação para o tema “*complex networks*”. Enquanto o artigo 0902.3068 cita quatro artigos no banco de dados do arXiv, o artigo 0805.2298 cita apenas dois artigos neste mesmo banco de dados.

Para construir as redes de similaridade, cada arquivo representando um artigo no banco de dados foi pré-processado para que as *tags* da linguagem de marcação $\text{\LaTeX}$ fossem removidas, de forma que os termos provenientes da linguagem $\text{\LaTeX}$ não afetassem o cálculos dos índices de similaridade (ver abaixo). A rede de similaridade $W^{(sim)}$ é não direcionada, e contém vértices representando os artigos (como na rede de citações) e arestas que conectam dois vértices com um dado grau de similaridade. Para cada aresta, um peso proporcional à similaridade entre os dois vértices é associado. Para visualizar as redes de

similaridade, eliminamos todos os pesos abaixo de um dado limiar (ver exemplo na Figura 12.2). Para obter a similaridade entre dois artigos (vértices), um procedimento de dois passos foi adotado: (i) modelagem dos textos como redes complexas (ver Capítulo 3); e (ii) comparação dos textos modelados como redes.

12.2 Similaridade baseada na topologia (TSI) e na consistência (CI)

Dois textos são considerados similares se eles compartilham um grande número de conceitos (representados por palavras). Para quantificar o grau de similaridade entre dois artigos, propomos um índice que é calculado através da comparação das redes dos textos em consideração. A quantificação não é feita apenas comparando-se a quantidade de conceitos compartilhados, pois a vizinhanças das palavras na rede também é considerada. Formalmente, o índice de similaridade entre dois textos foi calculado da seguinte forma. Sejam $N_p^{(i)}$ e $N_p^{(j)}$ as frequências de ocorrência da palavra p nos textos pré-processados i e j . O fator α_p , definido como

$$\alpha_p = \frac{N_p^{(i)} + N_p^{(j)} - |N_p^{(i)} - N_p^{(j)}|}{N_p^{(i)} + N_p^{(j)}} \quad (12.1)$$

representa uma primeira medida de similaridade. A idéia por trás desta equação é bastante simples. Ela quantifica a porcentagem de frequência das palavras compartilhada pelos dois textos, já que o numerador é o dobro de $\min\{N_p^{(i)}, N_p^{(j)}\}$ e o denominador é o dobro da frequência máxima compartilhada, que ocorre quando $N_p^{(i)} = N_p^{(j)}$.

A informação da redes de adjacência foi obtida da seguinte maneira. Sejam $v_p^{(i)}$ e $v_p^{(j)}$ os vértices associados à palavra p na rede do texto. Se $v_p^{(i)}$ possui κ vizinhos em comum com os vizinhos de $v_p^{(j)}$, o fator β_p pode ser representado como

$$\beta_p = \frac{\sum_{v=1}^{\kappa} (W^{(i)}(v_p, v) + W^{(j)}(v_p, v) - |W^{(i)}(v_p, v) - W^{(j)}(v_p, v)|)}{\sum_{v=1}^{\kappa} W^{(i)}(v_p, v) + \sum_{v=1}^{\kappa} W^{(j)}(v_p, v)}, \quad (12.2)$$

onde $W^{(i)}$ e $W^{(j)}$ representam as redes dos textos i e j , respectivamente. De modo análogo à equação 12.1, o numerador quantifica o dobro do peso que é compartilhado (i.e., o dobro de $\min\{W^{(i)}(v_p, v), W^{(j)}(v_p, v)\}$, onde v representa um vizinho arbitrário de v_p) e o denominador quantifica a quantidade de peso compartilhado. Portanto, similarmente a α_p , β_p está limitado ao intervalo $0 \leq \beta_p \leq 1$. Com estas definições, o índice de similaridade

para a palavra p , denominado como $IS(p)$, é dado por

$$IS(p) = \alpha_p \beta_p. \quad (12.3)$$

O índice global de similaridade (TSI) entre os pares de textos é então definido como

$$TSI = \frac{1}{M} \sum_{p=1}^M IS(p) \quad (12.4)$$

onde M é o tamanho do vocabulário dos dois textos concatenados.

O processo de quantificação da similaridade entre dois textos é bastante subjetivo. Apesar disso, acreditamos que o índice TSI é razoável, como indicado na Figura 12.2, onde foi possível diferenciar claramente a área de “*complex networks*” da área de “*graphenes*”. Quando ligados como uma rede única com arestas entre vértices sendo estabelecidas de acordo com o grau de similaridade, os vértices da mesma cor se agruparam nas mesmas comunidades. Dado que a cor representa um dado assunto, e um baixo *overlapping* de cores foi observado, uma separação efetiva foi possível com a utilização do índice TSI.

Além do índice de similaridade proposto nesta tese, empregamos três outros índices tradicionais: Jaccard, Cosseno e Dice (216). O procedimento para calcular a similaridade é diferente para estes três índices, no entanto, todos eles estão relacionados entre si porque são baseados nas frequências de termos comuns. Em outras palavras, se dois textos compartilham muitos termos, a similaridade correspondente será alta.

Uma vez que consideramos que um dos critérios para consultar e potencialmente citar um dado artigo deve ser baseado na similaridade de conteúdo com o manuscrito sendo escrito, verificamos se este critério é satisfeito usando índices de consistência. Entende-se por consistência a citação de artigos fortemente relacionados semanticamente. O índice de consistência CI foi concebido para variar entre zero e um, onde o limitante superior corresponde ao caso em que todas as referências citadas são bastante similares. O valor zero corresponde ao caso em que as citações dos artigos não pertencem ao banco de dados de artigos mais similares selecionados. Portanto, nos referimos como banco de dados total ao banco contendo os η artigos do arXiv usados na rede de citações (usamos $\eta = 900$), enquanto que o banco de dados selecionado inclui os artigos citados que pertencem ao banco de dados total. O procedimento utilizado para calcular CI foi o seguinte. Para cada artigo i no banco de dados, obtemos o limiar de similaridade para o qual o número de artigos similares no banco de dados total é igual ao número de referências no artigo i . Os manuscritos do banco de dados do arXiv com similaridade acima do limiar são denominados de banco de dados selecionado. Como exemplo, considere que o manuscrito i possui 20 referências. Este manuscrito irá contribuir com 20 artigos – os mais similares ao artigo i – para o banco de dados selecionado. Seja x_i o número de referências do manuscrito i que pertencem ao banco de dados selecionado. Seja y_i o número de referências do manuscrito

Figura 12.2 – Visualização da rede de similaridades (maior componente) para os artigos pertencendo aos assuntos “*complex networks*” (laranja) e “*graphenes*” (amarelo). A presença de apenas algumas conexões entre vértice com cores diferentes indica que o índice de similaridade conecta de fato artigos similares.

i que pertencem ao banco de dados total do arXiv. A razão x_i/y_i , portanto, fornece a fração de artigos citados que são similares ao manuscrito original. Por exemplo, suponha que 8 dentre das 20 referências no manuscrito i pertençam ao banco de dados do arXiv (banco de dados completo com η manuscritos). Então, $y_i = 8$. Se 4 destas 8 referências aparecem no banco de dados selecionado (o banco de dados de artigos mais similares), então $x_i = 4$, e a razão $x_i/y_i = 0,5$. O índice de consistência será dado por

$$CI = \frac{1}{\eta} \sum_{i=1}^{\eta} \frac{x_i}{y_i}. \quad (12.5)$$

Altos valores deste índice indicarão que as referências são de fato similares ao manuscrito sendo analisado. Neste ponto, é importante mencionar que na busca dos artigos mais similares consideramos apenas aqueles publicados antes do manuscrito original, já que

artigos concomitantes ou publicados depois não poderiam ser citados na prática.

Existe uma possibilidade de que $CI=1$ embora a lista de referência não seja tão similar. Para visualizar este problema, considere o seguinte exemplo. Suponha que um manuscrito com $x_i = 3$ e $y_i = 3$, mas as 3 referências são as menos similares no banco de dados selecionado. A fim de considerar este efeito, criamos um outro parâmetro CI' , que considera a ordem em que os manuscritos aparecem no banco de dados selecionado. Neste índice, os valores variarão entre zero (no caso em que as referências estão na cauda da lista) e um (no caso em que as referências estão no topo da lista). Formalmente, CI' é definido como na equação 12.6, a partir da adição de uma ponderação w_i em CI , dependendo do *rank* r_k de cada um dos x_i artigos na lista de similaridade contendo y_i artigos. Este peso foi concebido de forma que *rank*s maiores forneçam pesos maiores e vice-versa. Para conhecer a variação da soma de r_k , calculamos primeiro o máximo valor de Γ_{max} (ocorre quando todos os x_i artigos estão na cauda da lista) e o mínimo valor de Γ_{min} (ocorre quando todos os x_i artigos estão no topo da lista), conforme as equações 12.7 e 12.8, respectivamente. Então, a amplitude da soma de r_k é calculada com a equação 12.9. A diferença entre a soma de r_k e Γ_{min} também é computada a fim de se quantificar a distância desta soma do pior caso (i.e., $\Delta\Gamma$). Finalmente, com esses valores calculados, é possível calcular o peso w_i de acordo com a equação 12.10.

$$CI' = \frac{1}{\eta} \sum_{i=1}^{\eta} \frac{x_i \omega_i}{y_i} \quad (12.6)$$

$$\Gamma_{min} = \sum_{k=1}^{x_i} k = \frac{x_i(x_i + 1)}{2} \quad (12.7)$$

$$\begin{aligned} \Gamma_{max} &= \sum_{k=0}^{x_i-1} (y_i - k) = x_i y_i - \sum_{k=0}^{x_i-1} k \\ &= \frac{2x_i y_i - x_i(x_i - 1)}{2} = \frac{x_i}{2} (2y_i - x_i + 1) \end{aligned} \quad (12.8)$$

$$\begin{aligned} \Delta\Gamma &= \Gamma_{max} - \Gamma_{min} \\ &= \frac{x_i}{2} (2y_i - x_i + 1) - \frac{x_i}{2} (x_i + 1) \\ &= x_i(y_i - x_i) \end{aligned} \quad (12.9)$$

$$\begin{aligned}
\omega_i &= 1 - \frac{\sum_{k=1}^{x_i} r_k - \Gamma_{min}}{\Delta\Gamma} = 1 - \frac{2\sum_{k=1}^{x_i} r_k - x_i(x_i + 1)}{2x_i(y_i - x_i)} \\
&= \frac{x_i(2y_i - 3x_i - 1) - 2\sum_{k=1}^{x_i} r_k}{2x_i(y_i - x_i)} \tag{12.10}
\end{aligned}$$

Em suma, nesta seção definimos duas medidas para quantificar a consistência de citações em artigos científicos, i.e., a tendência para citar artigos cujo conteúdo está relacionado com o manuscrito em preparação. Definimos a consistência CI de um artigo como a fração de suas referências que estão entre os artigos mais similares no banco de dados. CI' modificou CI a fim de considerar a posição relativa dos artigos na lista de artigos mais similares. Com estas duas medidas, estamos aptos para quantificar a consistência de dois conjuntos de artigos pertencentes a duas áreas distintas.

12.3 Buscas sistemáticas em redes de citações

A fim de quantificar a frequência de acesso teórico dos artigos, as redes de citações foram analisadas usando o mecanismo de caminhadas aleatórias, onde a próxima aresta é escolhida a partir de uma distribuição uniforme dentre as arestas adjacentes. Adicionalmente, uma análise de caminhadas aleatórias foi realizada considerando o peso das arestas na escolha do próximo vértice. Em especial, o próximo vértice foi escolhido com um peso proporcional ao seu número de citações atuais. Desta forma, este segundo tipo de caminhada aleatória tende a selecionar primeiramente os vértices mais relevantes. Para calcular a frequência de acesso de cada vértice da rede, usamos a matriz $W^{(cit)}$ e criamos uma cadeia de Markov. A cada passo, a partícula efetuando uma caminhada aleatória pode mudar seu estado atual para outro estado (ou permanecer no mesmo estado), de acordo com uma distribuição de probabilidade associada previamente. Assumimos que as probabilidades de transição estão associadas com a probabilidade de um leitor seguir uma aresta, que é tomada como sendo proporcional ao seu peso peso (caminhada tradicional) ou proporcional ao grau de entrada do vizinho no fim da aresta. Desta forma, artigos com alta probabilidade de acesso no estado estacionário (quando o tempo tende ao infinito) são mais prováveis de serem vistos em uma caminhada aleatória. Nossa hipótese é que se autores de fato realizam uma busca sistemática na rede de citações, então os artigos com as maiores frequências de acesso no estado estacionário serão provavelmente mais consultados (e portanto mais citados). Para verificar se este critério é obedecido, calculamos o coeficiente de correlação de Pearson entre o número real de citações (ou *downloads*) e o acesso esperado dado pela frequência da caminhada aleatória.

12.4 Análise dos índices de consistência

O principal objetivo da análise realizada neste capítulo é fornecer formalismos para verificar se os critérios possíveis na seleção de referências para um artigo sendo escrito estão sendo seguidos. Sugerimos dois importantes princípios a serem seguidos por autores numa busca literária e numa seleção de publicações para serem analisadas e possivelmente citadas: (i) seleção de manuscritos fortemente relacionados (217, 218), o qual pode ser feito através da análise da similaridade de conteúdo; e (ii) busca sistemática na rede de citação envolvendo artigos de destaque relacionados. Obviamente, outros critérios poderiam ter sido utilizados. Por exemplo, os autores poderiam buscar as publicações consideradas como as mais relevantes no tópico sendo estudado, os quais poderiam ser identificadas pelo número de citações ou reputação dos autores/jornais em que os manuscritos científicos são publicados. Infelizmente, o uso deste critério introduziria inevitavelmente um viés, o qual desejamos eliminar. Além disso, estes manuscritos relevantes provavelmente são encontrados em buscas sistemáticas. Outra possibilidade é que os autores possam precisar citar artigos pouco relacionados, que podem por exemplo fornecer uma metodologia de outra área de pesquisa. Enquanto este tipo de citação seria totalmente justificável, ela ocorre muito raramente. Portanto, para o propósito de nossa análise, não precisamos considerar este último critério.

Para verificar se os autores selecionam artigos semelhantes durante a preparação da lista de referências, definimos os índices de consistência CI e CI', os quais assumem valores máximos quando todas as referências possuem a maior similaridade semântica com o texto sendo produzido. Em outras palavras, os artigos citados estão entre os mais similares do banco de dados. Estes índices foram calculados para o *corpus* relacionado aos temas de “*complex networks*” e “*graphenes*” usando o texto integral ou extratos do resumo ou introdução. Estes extratos foram utilizados porque tanto as seções introdutórias como o resumo geralmente representam as seções mais representativas dos tópicos de um manuscrito científico. Assim, ao examinarmos apenas estas seções, investigamos o conteúdo que expressa a essência do texto, bem como sua inserção na área de pesquisa. Como será demonstrado a seguir, os resultados não dependem da escolha no uso de extratos ou do texto integral. Os textos foram representados como redes complexas, usando a topologia destas redes para obter a similaridade de conteúdos entre dois manuscritos. Os resultados apresentados nas Tabelas 12.1 e 12.2 revelam índices de consistência relativamente baixos para ambas as áreas, independentemente do uso de extratos ou textos completos no cálculo dos índices de consistência. De fato, em média, apenas uma a cada quatro referências bibliográficas de um artigo eram significativamente similares ao artigo analisado. Curiosamente, o valor de CI' é aproximadamente metade do valor de CI, o que significa que as citações mais similares foram distribuídas regularmente na lista de artigos mais

Tabela 12.1 – Índice de consistência CI calculado para duas áreas. R refere-se ao uso do resumo, I refere-se ao uso da introdução e T refere-se ao uso do artigo completo.

Área	Cosseno	Dice	Jaccard	TSI
Redes complexas (R)	0,15	0,15	0,14	0,11
Redes complexas (R+I)	0,18	0,18	0,19	0,17
Redes complexas (T)	0,22	0,22	0,23	0,20
Grafenos (R)	0,13	0,14	0,14	0,11
Grafenos (R+I)	0,11	0,11	0,10	0,16
Grafenos (T)	0,25	0,21	0,21	0,22

Tabela 12.2 – Índice de consistência CI' calculado para duas áreas. R refere-se ao uso do resumo, I refere-se ao uso da introdução e T refere-se ao uso do artigo completo.

Área	Cosseno	Dice	Jaccard	TSI
Redes complexas (R)	0,09	0,09	0,09	0,08
Redes complexas (R+I)	0,12	0,12	0,11	0,11
Redes complexas (T)	0,15	0,14	0,17	0,15
Grafenos (R)	0,06	0,08	0,08	0,07
Grafenos (R+I)	0,07	0,07	0,06	0,07
Grafenos (T)	0,17	0,14	0,14	0,15

similares. Ou seja, elas não estão concentradas nem entre as mais similares nem entre as menos similares. Estas conclusões não foram afetadas pelo uso de outros métodos na quantificação de similaridade. Se as palavras vizinhas não são consideradas no cálculo da similaridade, i.e. β_p é ignorado, então a similaridade é essencialmente baseada no número de palavras compartilhadas, desconsiderando a topologia da rede. Estes resultados confirmam a hipótese de que outros fatores são mais relevantes do que a similaridade (a data de publicação é um desses fatores, pois manuscritos antigos tendem a ser esquecidos (188)).

12.5 Análise da busca sistemática

O segundo critério avaliado foi a busca sistemática que os autores deveriam realizar na rede de citações contendo artigos relevantes ou relacionados com a pesquisa sendo conduzida. Simulamos a busca sistemática através de uma caminhada aleatória na rede de citações para as áreas de “*complex networks*” e “*graphenes*”. Para determinar a frequência de acesso relativa para cada vértice na rede de citação, a matriz de transição τ é criada a partir da redefinição da matriz $W^{(cit)}$ (maior componente) como $A^{(cit)}$ para representar a rede de busca sistemática. O elemento $a_{ij}^{(cit)}$ assume valor 1 se existe uma aresta $i \rightarrow j$ na rede de citações $W^{(cit)}$. Assumimos que a aresta $j \rightarrow i$ (sentido contrário da rede original) possui um peso igual a 0.2, já que é mais fácil para o leitor seguir uma referência em um artigo do que descobrir quem citou este artigo. Dessa forma, associamos o menor peso

à direção menos provável. A matriz τ é uma matriz estocástica, derivada de $A^{(cit)}$. Sua definição depende do tipo de caminhada aleatória desejada. Para a caminhada aleatória tradicional, a probabilidade de transição do vértice i ao vértice j depende somente das arestas de saída de i :

$$\tau_{ij} = \frac{a_{ij}^{(cit)}}{\sum_j a_{ij}^{(cit)}}. \quad (12.11)$$

Por outro lado, se a caminhada aleatória considera a importância dos vizinhos para realizar o próximo passo, então τ dependerá do grau de entrada k_i dos vizinhos:

$$\tau_{ij} = \frac{a_{ij}^{(cit)} \sum_{k=1} a_{kj}^{(cit)}}{\sum_j a_{ij}^{(cit)} \sum_k a_{kj}^{(cit)}}. \quad (12.12)$$

Em ambos os casos, τ mede a probabilidade do agente executando a caminhada aleatória visitar o vértice j depois de ter visitado o vértice i . Estamos interessados na distribuição de probabilidades do estado estacionário do agente estar localizado em cada vértice. Em outras palavras, desejamos saber a probabilidade de se estar em um dado vértice após um número infinito de passos. Esta distribuição, denotada como π , pode ser obtida através da resolução da equação

$$\tau\pi = \pi. \quad (12.13)$$

Segundo a equação 12.13, π é o auto-vetor associado ao auto-valor unitário de τ , onde a soma dos elementos de π é igual a um para refletir a soma das probabilidades. A condição necessária e suficiente para que π tenha solução única exige que τ seja irredutível, i.e. a rede precisa ser fortemente conectada, o que acontece quando cada vértice pode ser alcançável a partir de qualquer outro vértice através de um número finito de passos. Felizmente, esta condição é satisfeita em nossa rede, já que selecionamos o maior componente de $W^{(cit)}$ na criação da matriz $A^{(cit)}$.

O coeficiente de correlação de Pearson foi calculado para inferir a correlação entre a frequência esperada de visitas aos vértices (artigos) e o número real de citações e *downloads* para os manuscritos no repositório do arXiv. Especificamente, o número de *downloads* e citações foram obtidos do banco de dados Citebase (219). Os resultados ilustrados nas Figuras 12.3 e 12.4 revelam baixas correlações com relação aos dois índices nas duas áreas distintas analisadas.

Os baixos valores de correlação provavelmente indica que os pesquisadores não consultam as referências dos artigos para citar; em outras palavras, o critério de uma busca sistemática na rede de citação não é satisfeito. Mesmo quando uma busca eficiente de trabalhos relacionados é feita, os autores ainda assim podem esquecer de incluir artigos que de fato foram consultados na busca sistemática (220).

Figura 12.3 – Gráficos de dispersão mostrando a correlação entre a medida espectral calculada para a rede de citação e o frequência real de acesso, para a área de redes complexas. Com exceção do número de citações comparado à frequência de acesso preferencial na caminhada aleatório, todas as correlações assumiram valores baixos.

12.6 Cienciometria virtual

Os resultados obtidos em nossa análise claramente indicam, pelo menos no caso do banco de dados do arXiv, que os dois critérios mais importantes que deveriam guiar citações em manuscritos científicos não são seguidos na prática. Este resultado é preocupante pois a premissa básica do método científico baseada na análise de artigos relacionados para evitar a reprodução de esforços e garantir originalidade, não é seguida. Existe também um efeito colateral causado nas estatísticas das citações de autores e jornais. Estas conclusões também parecem ser válidas para outras áreas da ciência. Através de

uma pesquisa realizada com editores de jornais de Psicologia, Croner (221) mostrou que 80% dos participantes acreditavam que os autores dos manuscritos falhavam em citar artigos relevantes, pois algumas citações pareciam ser forjadas para chamar a atenção dos editores (222). Existe também o problema relacionado com a data de publicação, pois existe uma tendência de que os artigos científicos mais recentes sejam mais citados do que artigos relacionados publicados há um longo tempo (223, 224). Outros fatores que influenciam a escolha de referências têm sido relatados amplamente na literatura. Por exemplo, a probabilidade de citação de um artigo pode depender da área de pesquisa (225, 226) (áreas maiores recebem mais citações do que áreas menores), do fator de impacto da revista (222), do grau de amizade dos autores (227), do tipo de artigo (p.e., *survey*, artigo regular ou *letter*) (228), da língua (229) e da disponibilidade (230) (p.e., se o manuscrito está disponível gratuitamente *online*). As citações não são, portanto, apenas baseadas em influências intelectuais dos pares científicos, mas também baseadas em fatores não-científicos (231). Acreditamos que medidas deveriam ser tomadas a fim de complementar e/ou corrigir os procedimentos de citação normalmente adotados. Embora seja difícil mudar os hábitos de citação já estabelecidos, os contínuos avanços da informática e Internet permitem que novas abordagens sejam implementadas. Nesta seção, descrevemos uma ferramenta computacional que poderia ser útil na preparação das referências bibliográficas. Além disso, esta ferramenta poderia ser utilizada na avaliação do impacto de pesquisadores e artigos com relação ao conceito de citações virtuais, i.e., citações que seriam artificialmente estabelecidas de acordo com o critério de similaridade de conteúdo. Este conceito está detalhado a seguir.

A ferramenta automática para auxiliar os autores na identificação de referências relevantes poderia ser feita com uma simples implementação dos métodos propostos na seções anteriores. Mais especificamente, um texto sendo preparado poderia ser transformado em uma rede complexa e seu conteúdo poderia ser comparado com o conteúdo de todos os outros documentos em um banco de dados (p.e. o arXiv) usando os índices de similaridade definidos neste trabalho. A saída seria uma lista de trabalhos relacionados classificados em ordem decrescente de similaridade. O autor poderia então avaliar esta lista e identificar trabalhos potencialmente relevantes. Esta ferramenta poderia ser usada também na obtenção de citações virtuais para um manuscrito já publicado da seguinte maneira. Dada uma área de pesquisa na qual um conjunto de artigos poderiam ser selecionados, o número de citações incluídas poderia ser fixa ou dependente da similaridade de conteúdo entre os artigos. As citações virtuais para um artigo poderiam ser escolhidas acima de um dado limiar de similaridade, que poderia ser definido como aquele capaz de gerar um número médio de referências que é igual ao número real médio de referências dos artigos no banco de dados. Portanto, um artigo que é similar a uma grande quantidade de artigos no banco de dados receberia muitas citações, já que supostamente ele lida com um tópico de interesse geral. Usando a rede de citações obtidas deste banco de dados, como

discutido na seções anteriores, outra possibilidade seria avaliar o impacto de um artigo através do cálculo da frequência de visitas em uma caminhada aleatória nesta rede.

As duas medidas obtidas com as citações virtuais e caminhadas aleatória poderiam ser utilizadas para comparar os fatores de impacto de um autor (ou de um artigo específico) com os fatores de outros autores (ou artigos), com as seguintes vantagens:

1. Não haveria nenhum um viés relacionado com a escolha pessoal de cada autor na seleção das referências, embora outros vieses eventualmente poderiam permancer. Por exemplo, apenas os artigos escritos na mesma língua do artigo sendo preparado seria incluído na lista de referências. Além disso, os artigos não pertencentes ao banco de dados nunca seriam citados.
2. A comparação com artigos da mesma área seria simples e direta.

Neste ponto, deve-se enfatizar que de maneira alguma sugerimos que as estratégias tradicionais de citação sejam substituídas por citações virtuais e métricas decorrentes da frequência de visitas em caminhadas aleatórias realizadas em redes de citação. Tal estratégia não seria razoável porque, além da similaridade, a qualidade do manuscrito deveria ser considerada. Apesar disso, acreditamos que as novas métricas seriam úteis na complementação da avaliação do impacto de um dado texto científico ao refletir, de modo compreensivo, seu relacionamento com a literatura.

12.7 Conclusões

Analisando um conjunto considerável de 1.800 artigos de duas áreas de pesquisa, descobrimos que os dois procedimentos de uma pesquisa bibliográfica não são seguidos pelos autores. Aparentemente, outros fatores são importantes, os quais podem não considerar o mérito científico. As conclusões deste capítulo são similares às conclusões de outros estudos, que relata que o número de citações e seu respectivo impacto podem depender do país dos autores (232) ou do lugar de publicação (233). Com relação ao método sugerido para avaliar as contribuições dos cientistas, o qual chamamos de cienciometria virtual, ele é vantajoso pois não é propenso ao viés dos autores. Tal método também está de acordo com propostas recentes de métricas para capturar a importância de conferências (234). Além disso, propomos o uso de ferramentas computacionais para auxiliar os autores durante a pesquisa bibliográfica na literatura. De fato, comunidades de médicos e engenheiros de software (235) já possuem métodos bem estabelecidos para conduzir pesquisas bibliográficas (formulação do problema, seleção de estudos e coleção de dados são alguns exemplos), os quais são similares aos propostos aqui – embora com motivações distintas.

Figura 12.4 – Gráficos de dispersão confirmando a correlação entre a medida espectral calculada para a rede de citação e a frequência real de acesso, para a área de grafenos.

Capítulo 13

Desambiguação em redes de colaboração

O paradigma *e-Science* pode ser explorado para transformar uma grande quantidade de dados disponíveis eletronicamente para aquisição de informação em diversos campos. Em Ciência e Tecnologia, por exemplo, grandes bases de dados como redes de citações (176, 180, 187, 188, 190), periódicos (236), arXiv (210), CiteSeer (237), DBLP (238), *Web of Science* (211) e *Google Scholar* (239) estão disponíveis para análises úteis aos processos de decisão de financiamento de agência e instituições acadêmicas. As redes de citação, em particular, têm sido estudadas em uma variedade de contextos, p.e. na identificação dos manuscritos mais relevantes para a preparação de uma *survey* e na quantificação do impacto de periódicos, conferências, pesquisadores e instituições. A aplicabilidade destes banco de dados, no entanto, é prejudicada na presença de ruídos em forma de ambiguidades. Nos bancos de dados científicos, dois problemas principais ocorrem na lista de autores de manuscritos científicos: (i) o mesmo autor pode ser referenciado de diversas formas, e (ii) autores distintos podem compartilhar nomes idênticos. Em especial, o caso (ii) é bastante ubíquo em publicações envolvendo pesquisadores chineses e coreanos.

Diversos métodos têm sido utilizados para resolver a tarefa de desambiguação de nomes de autores, o qual está relacionada com os problemas de mapeamento (215) e detecção de duplicatas (240). Estes métodos são baseados principalmente em mineração de textos e técnicas de processamento de línguas naturais, já que os pesquisadores podem ser diferenciados pelo seus campos de pesquisa através da análise textual do conteúdo de suas produções científicas. A lista de co-autores também tem sido utilizada como critério para desambiguação (241, 242) pois cientistas tendem a manter grupos de colaboração específicos. Outros critérios menos importantes incluem a análise do nome do periódico (243), das auto-citações (244) e dos meta-dados (244).

Uma abordagem complementar na detecção e correção de inconsistências em banco de dados é a representação de relacionamentos relevantes como redes complexas. Neste

capítulo, desenvolvemos um método inspirado na técnica de desambiguação de palavras desenvolvida no Capítulo 7 para desambiguar nomes de autores em artigos científicos. Embora estratégias tradicionais de redes de colaborações sejam úteis em áreas altamente colaborativas, elas podem falhar em áreas cujos manuscritos incluem menos de dois co-autores (245). Além disso, as estratégias tradicionais não consideram informações relevantes da rede, tais como padrões de conectividade de longo alcance. As abordagens desenvolvidas neste capítulo foram avaliadas no banco de dados do arXiv, onde nomes ambíguos foram deliberadamente introduzimos para gerar um banco de dados confiável. Propomos três abordagens distintas: (i) utilização de hierarquias profundas para analisar a conectividade das redes de colaboração; (ii) caracterização topológica de conectividade; e (iii) composição de estratégias através de um classificadores híbrido.

13.1 Redes de colaboração

As redes de colaboração foram geradas a partir de bancos de dados do repositório arXiv. Nestas redes, cada vértice representa um indivíduo, e arestas são estabelecidas caso eles co-participam em um documento. O processo de produção da rede colaborativa de autores está ilustrado na Figura 13.1 para uma pequena rede contendo sete autores fictícios. A rede completa de colaboração do arXiv é mostrada na Figura 13.2. Nos manuscritos fictícios da figura, temos como objetivo desambiguar os autores com o mesmo nome AA. Note que precisamos representar os mesmos autores com nomes ambíguos como vértices distintos na rede porque a desambiguação é feita em nível de artigo. De qualquer forma, não é necessário criar uma rede completamente diferente toda vez que um novo indivíduo precisa ser desambiguado. Uma rede comum refletindo todas as colaborações no banco de dados pode ser mantida e apenas alguns poucos vértices e arestas seriam adicionadas/removidas da rede comum para a análise de cada novo autor. Desta forma, o processo de desambiguação dependeria fundamentalmente do número de artigos escritos pelo autor sendo analisado.

A força da conexão entre dois vértices i e j é dada pelo peso w_{ij} :

$$w_{ij} = \sum_{\rho \in \Pi} \frac{\delta_{ij\rho}}{\|\rho\|}, \quad (13.1)$$

onde

$$\delta_{ij\rho} = \begin{cases} 1 & \text{se } i \text{ e } j \text{ aparecem em } \rho, \\ 0 & \text{caso contrário.} \end{cases} \quad (13.2)$$

Π representa o conjunto de todos os manuscritos do banco de dados e $\|\rho\|$ é o número de autores de um dado manuscrito ρ . O peso foi dividido por $\|\rho\|$ para considerarmos o

Figura 13.1 – Exemplo de uma rede colaborativa construída a partir de uma lista fictícia de autores. AX significa autor X. As seguintes autorias de artigos foram consideradas: paper 1 (AA, AB e AC), paper 2 (AA, AG e AD), paper 3 (AA, AF, AG e AI), paper 4 (AA e AG), paper 5 (AC e AD), paper 6 (AE e AF) e paper 7 (AF e AE).

Figura 13.2 – Componente gigante da rede colaborativa de uma fração do banco de dados do arXiv. As cores dos vértices representam seus graus de conectividade. Vértices em azul e vermelho possuem o menor e maior número de vizinhos, respectivamente.

fato de que os relacionamentos entre poucos autores tendem a ser mais fortes do que os relacionamentos envolvendo vários autores (246). O peso dos *links* não é mostrado na Figura 13.1, mas seu cálculo é direto. Por exemplo, o peso para a ligação da ligação $AB \leftrightarrow AC$ é $1/3$ e o peso da conexão $AE \leftrightarrow AF$ é 1 ($1/2$ proveniente do manuscrito 6 e $1/2$ proveniente do manuscrito 7).

13.2 Desambiguação hierárquica

Na estratégia baseada em co-autoria, cada ocorrência de uma entidade ambígua é caracterizada pelas relações de co-participação no mesmo documento. Seja e uma entidade ambígua e considere que o vetor $\vec{v}_e$ descreva os atributos de co-autoria de e . Cada elemento $v_e(i)$ representa uma das possíveis entidades do banco de dados. Dessa forma, se i e j aparecem no mesmo documento, então $v_e(i) = 1$. Caso contrário, $v_e(i) = 0$. A fim de reduzir a complexidade do problema, as técnicas de redução de dimensionalidade PCA e CVA foram utilizadas. Dado que esta última técnica apresentou um desempenho superior ao PCA nos experimentos, todos os resultados relatados neste capítulo foram obtidos com a análise de variáveis canônicas (CVA).

Nesta estratégia foram utilizados um conjunto de $N_A = 1.842$ atributos, onde N_A representa o número de autores do arXiv. Uma vez que os dados incluindo os N_A autores apresentam uma alta dimensão, técnicas de redução de dimensionalidade foram utilizadas antes da aplicação das técnicas de reconhecimento de padrão (ver Capítulo 4). A análise de vizinhos imediatos dos autores na rede colaborativa permitiu distinguir autores homônimos, sendo que a precisão média aumentou conforme o número de hierarquias utilizadas. A Figura 13.3 mostra o comportamento da medida-f obtido para a terceira hierarquia da rede do arXiv. Como esperado, as curvas de performance diminuem com o número de autores homônimos em todos os algoritmos testados (resultado não mostrado). O desempenho superior da análise considerando a segunda e terceira hierarquias pode ser comprovado pela análise da Figura 13.4, que mostra a porcentagem em que cada hierarquia alcançou o melhor desempenho. Uma análise estatística desta figura é fornecida na Tabela 13.1. Os resultados são consistentes com a melhor performance obtida na tarefa de tradução automática (ver Capítulo 9). Hierarquias maiores que $h = 3$ não foram verificadas devido ao alto custo computacional associado. No entanto, acreditamos que a performance não seria melhorada consideravelmente já que o uso de hierarquias elevadas geralmente causa mais perda do que ganho de informação.

O processo de desambiguação em redes colaborativas pode depender da densidade de arestas. Se a rede é completa, por exemplo, a desambiguação se deteriora. Neste caso, outros atributos devem ser utilizados. Felizmente, na prática, as redes colaborativas são

Figura 13.3 – Medida-f obtida para a tarefa de desambiguação hierárquica usando a terceira hierarquia. Os algoritmos utilizados foram (a) κ -NN, e (b) C4.5. Em ambos os casos, a capacidade discriminativa diminui com o aumento do número de autores ambíguos.

Figura 13.4 – Porcentagem de casos em que cada hierarquia alcança o valor máximo de medida-r para os seguintes algoritmos: (a) ligação completa (128); (b) Ward's (128); (c) k-means (128); e (d) Expectation Maximization (128). Na maioria dos casos, a terceira hierarquia apresenta um desempenho melhor que a primeira e segundas hierarquias. Um desempenho similar foi obtido para outros algoritmos.

Tabela 13.1 – p -value representando a probabilidade de que o modelo baseado na terceira hierarquia apresente um desempenho superior ao modelo baseado na primeira hierarquia considerando classificações aleatórias. Note que em todos os casos $p < 5 \times 10^{-2}$. Portanto, a hipótese acima é estatisticamente significativa. Os algoritmos analisados foram *Expectation Maximization* (EM), Wards e k-means.

$\#N_a$	EM	Wards	k-means
2	2.2×10^{-2}	7.8×10^{-5}	9.7×10^{-6}
3	9.7×10^{-6}	9.7×10^{-6}	9.7×10^{-6}
4	7.0×10^{-7}	3.0×10^{-8}	7.0×10^{-7}
5	3.0×10^{-8}	3.0×10^{-8}	7.0×10^{-7}
6	3.0×10^{-8}	7.0×10^{-7}	3.0×10^{-8}
7	3.0×10^{-8}	8.0×10^{-7}	3.0×10^{-8}
8	3.0×10^{-8}	3.0×10^{-8}	9.7×10^{-6}
9	9.7×10^{-6}	3.0×10^{-8}	8.0×10^{-7}
10	7.0×10^{-7}	3.0×10^{-8}	3.0×10^{-8}

organidas em módulos (comunidades) e portanto o coeficiente de aglomeração dentro das comunidades é alto. Dessa forma, o alto *clustering* é desejável quando autores pertencem a comunidades distintas. Quanto maior o coeficiente de aglomeração, menor é o número de ligações externas à comunidade. Consequentemente, os padrões de co-autoria dos autores serão bastante distintos dado que eles pertençam a comunidades distintas.

13.3 Desambiguação topológica

Além da estratégia baseada em co-autores, avaliamos a capacidade das medidas topológicas no processo de desambiguação. As principais medidas utilizadas foram k , que quantifica o número de colaboradores distintos; o *strength* s , que quantifica a soma dos pesos das ligações; o coeficiente de aglomeração C , que mede a densidade de ligações ao redor do vértice de interesse; o grau e *strength* médio dos vizinhos ($\langle k_n \rangle$ e $\langle s_n \rangle$); e os desvios dos graus e *strength* dos vizinhos (Δk_n e Δs_n). Os resultados na Figura 13.5 confirmam que a discriminabilidade média obtida com a topologia das redes é pior do que a obtida com a discriminação hierárquica (ver Figura 13.3). Apesar disso, a desambiguação baseada em atributos topológicos é estatisticamente significativa conforme mostra a Tabela 13.2. Estes resultados confirmam que autores possuem padrões de conectividade específicos nas redes de colaborações.

Tabela 13.2 – Número de casos no banco de dados do arXiv em que a classificação baseada em atributos *topológicos* foi estatisticamente significativa, i.e., p -value $< 5,0 \times 10^{-2}$. $\#N$ é o número de nomes ambíguos dos autores.

$\#N$	2	3	4	5	6	7	8	9	10
(%)	80,8	88,0	100,0	97,3	100,0	100,0	100,0	100,0	100,0

Figura 13.5 – Medida-f para a tarefa de desambiguação baseada na abordagem topológica. Os algoritmos utilizados foram (a) κ -NN e (b) C4.5. Em todos os casos, a capacidade para discriminar autores diminuiu conforme o número de entidades ambíguas aumentou.

Também investigamos quais medidas topológicas foram as mais eficientes para discriminar os autores. A análise das medidas foi realizada de acordo com os critérios KL e AUC (ver Capítulo 4). Enquanto o método KL apresenta como vantagem o fato de a avaliação ser independente do algoritmo, tal método pode desconsiderar importantes relações não-triviais entre atributos. Por esta razão, também utilizamos o método AUC, que leva em consideração tais relações não triviais^a. As medidas mais eficientes para discriminação foram $\langle k_n \rangle$ and $\langle s_n \rangle$, conforme ilustra a Tabela 13.3, que mostra a porcentagem de casos em que estas medidas apareceram como os melhores atributos no método de reconhecimento de padrões mais eficiente (κ -NN). A tabela também mostra os valores correspondentes de p -value, considerando uma ordenação aleatória de medidas como modelo nulo, em ambos os critérios de classificação. Significativamente, o coeficiente de aglomeração C não apareceu entre as medidas mais importantes para distinção. Isto significa que a discriminação de autores parece ser governada pelo número médio de co-autores de vizinhos, e com uma menor importância, pelo *strength* de tais conexões. Portanto, a informação mais relevante é, na verdade, o número de conexões com autores externos, i.e., o número de conexões dos co-autores com outros autores que não colaboraram com o autor em análise. Isto significa que a estrutura dos vizinhos permite uma melhor caracterização do que a estrutura local do vértice. Esta conclusão é consistente com os resultados obtidos na discriminação hierárquica.

^a Embora o método AUC a princípio dependa da estratégia de reconhecimento de padrões utilizada, em nossos experimentos, todos os algoritmos apresentaram praticamente os mesmos resultados.

Tabela 13.3 – Porcentagem de casos em que $\langle k_n \rangle$ e $\langle s_n \rangle$ apareceram na primeira posição dos *ranks* feitos usando os critérios KL com o algoritmo κ -NN. O *p*-value corresponde à probabilidade da porcentagem correspondente ser obtida considerando como modelo nulo um *rank* aleatório de medidas. $\#N_a$ representa o número de nomes ambíguos para um dado autor.

$\#N_a$	$\langle k_n \rangle$		$\langle s_n \rangle$	
	%	<i>p</i> -value	%	<i>p</i> -value
2	56,8%	$1,0 \times 10^{-15}$	31,1%	$9,6 \times 10^{-5}$
3	67,4%	$1,0 \times 10^{-15}$	23,1%	$3,9 \times 10^{-2}$
4	73,2%	$1,0 \times 10^{-15}$	20,6%	$7,0 \times 10^{-2}$
5	73,2%	$1,0 \times 10^{-15}$	23,1%	$3,9 \times 10^{-2}$
6	75,8%	$1,0 \times 10^{-15}$	20,6%	$7,0 \times 10^{-2}$
7	71,1%	$1,0 \times 10^{-15}$	25,8%	$1,0 \times 10^{-2}$
8	66,3%	$1,0 \times 10^{-15}$	30,0%	$3,0 \times 10^{-4}$
9	69,4%	$1,0 \times 10^{-15}$	26,3%	$7,2 \times 10^{-3}$
10	42,6%	$7,7 \times 10^{-13}$	52,6%	$1,0 \times 10^{-15}$

Tabela 13.4 – Porcentagem de casos em que $\langle k_n \rangle$ e $\langle s_n \rangle$ apareceram na primeira posição dos *ranks* feitos usando os critérios AUC com o algoritmo κ -NN. O *p*-value corresponde à probabilidade da porcentagem correspondente ser obtida considerando como modelo nulo um *rank* aleatório de medidas. $\#N_a$ representa o número de nomes ambíguos para um dado autor.

$\#N_a$	$\langle k_n \rangle$		$\langle s_n \rangle$	
	%	<i>p</i> -value	%	<i>p</i> -value
2	45,2%	$3,8 \times 10^{-14}$	50,0%	$1,0 \times 10^{-15}$
3	70,0%	$1,0 \times 10^{-15}$	29,5%	$1,2 \times 10^{-3}$
4	73,7%	$1,0 \times 10^{-15}$	26,3%	$2,1 \times 10^{-2}$
5	81,6%	$1,0 \times 10^{-15}$	18,4%	$7,3 \times 10^{-1}$
6	91,1%	$1,0 \times 10^{-15}$	8,9%	$1,0 \times 10^0$
7	86,3%	$1,0 \times 10^{-15}$	13,2%	$1,0 \times 10^0$
8	96,8%	$1,0 \times 10^{-15}$	3,2%	$1,0 \times 10^0$
9	82,6%	$1,0 \times 10^{-15}$	17,4%	$1,0 \times 10^0$
10	92,1%	$1,0 \times 10^{-15}$	7,9%	$1,0 \times 10^0$

13.4 Desambiguação híbrida intrínseca

As estratégias baseadas na conectividade (terceira hierarquia) e nos atributos topológicos foram combinadas na tarefa de desambiguação de uma banco de dados colaborativa obtido do banco de dados IMDb, que contém a relação de atores e suas participações conjuntas em filmes. Diferentemente da desambiguação híbrida realizada com palavras (ver Capítulo 7), nesta seção realizamos a combinação intrínseca de estratégias, na qual os atributos são combinados antes da classificação. Em outras palavras, cada instância ambígua é caracterizada por atributos hierárquicos e topológicos. Apenas após esta caracterização é que a classificação (de baixo nível tradicional) é realizada. A Tabela 13.5 mostra a taxa de acerto obtida para cada ator e o valor correspondente de p -value (assumindo como modelo nulo um sistema de desambiguação aleatório). Para cada ator, as taxas mostradas referem-se à estratégia topológica (TP), à estratégia de conectividade hierárquica (CN) e à estratégia híbrida (CN+TP). Para os atores Attila e Matt Hughes, os atributos da abordagem TP apresentam um desempenho inferior ao da estratégia CN, mas os melhores resultados foram obtidos com a combinação CN+TP. Analogamente, no caso dos atores Igor e Justin Long, a combinação CN+TP gerou a melhor desambiguação. Surpreendentemente, no caso de Bill Baley, a estratégia TP foi capaz de gerar o melhor sistema de desambiguação. No caso de Steve Austin, a combinação também foi capaz de gerar os melhores resultados, que já haviam sido obtidos com a estratégia CN. Finalmente, para o ator Christian, uma desambiguação aleatória já foi eficiente, e portanto não faz sentido comparar técnicas. Os resultados deste estudo são análogos à descoberta de desambiguação obtida com palavras no Capítulo 7, na qual a topologia se mostrou útil para revelar padrões relacionados à semântica das palavras.

13.5 Desambiguação híbrida extrínseca

Diferentemente da seção anterior, combinamos os atributos hierárquicos e topológicos de forma extrínseca, i.e., após o processo de classificação. Em outras palavras, um peso indicando a membresia de cada instância é fornecido por cada estratégia. Então, estes pesos são combinados a fim de gerar uma probabilidade de que a instância desconhecida pertença a uma dada classe. O algoritmo κ -NN *fuzzy* descrito no Capítulo 4 foi utilizado para construir o desambiguador híbrido da seguinte maneira. $u_{ij}^{(T)}$ representa a força com que o objeto x_j pertence à classe i quando atributos hierárquicos (colaborativos) são utilizados. Analogamente, $u_{ij}^{(C)}$ representa a força com que o mesmo objeto pertence à classe i de acordo com métricas topológicas. Desta forma, o peso final que representa a

Tabela 13.5 – Taxa de acerto e p -value obtido para a desambiguação baseada em topologia (TP) e na lista de vizinhos hierárquicos (CN). A combinação (CN+TP) de atributos também foi examinada. Para todos os atores, os atributos topológicos aparecem entre os melhores classificadores.

Ator	Técnica	Acerto	p -value	Atributos
Attila	C4.5	64.3 %	1.0×10^{-1}	TP
	κ -NN	78.6 %	2.8×10^{-2}	CN
	κ -NN	85.7 %	6.0×10^{-3}	CN + TP
M. Hughes	κ -NN	77.4 %	5.8×10^{-1}	TP
	κ -NN	80.7 %	8.5×10^{-2}	CN
	κ -NN	83.9 %	3.6×10^{-2}	CN + TP
Igor	C4.5	81.8 %	5.5×10^{-2}	TP
	κ -NN	81.8 %	5.5×10^{-2}	CN
	κ -NN	86.4 %	1.8×10^{-2}	CN+TP
J. Long	κ -NN	95.8 %	2.2×10^{-4}	TP
	κ -NN	93.6 %	3.0×10^{-2}	CN
	C4.5	95.8 %	2.2×10^{-4}	CN+TP
B. Bailey	C4.5	96.0 %	2.2×10^{-2}	TP
	κ -NN	87.8 %	6.1×10^{-1}	CN
	κ -NN	87.8 %	6.1×10^{-1}	CN+TP
S. Austin	Bayes	84.8 %	1.0×10^{-1}	TP
	Bayes	88.9 %	2.2×10^{-2}	CN
	κ -NN	88.9 %	2.2×10^{-2}	CN + TP
Christian	κ -NN	97.1 %	4.5×10^{-1}	TP
	κ -NN	97.6 %	2.9×10^{-1}	CN
	κ -NN	97.6 %	2.9×10^{-1}	CN + TP

força com que x_j pertence à classe i é dado por:

$$u_{ij}^{(H)} = \lambda u_{ij}^{(T)} + (1 - \lambda) u_{ij}^{(C)}, \quad (13.3)$$

onde $\lambda \in [0, 1]$. Note que, λ na equação 13.3, desempenha o papel de ponderação das diferentes estratégias. Quanto maior o valor de λ , maior importância é alocada à estratégia topológica e vice-versa. A classe i associada ao objeto x_j é aquela que satisfaz a relação

$$\tilde{u}_{ij}^{(H)} = \max_i u_{ij}^{(H)}. \quad (13.4)$$

A abordagem híbrida extrínseca para a tarefa de desambiguação foi testada com três redes de colaboração obtidas do repositório do arXiv. Estas três redes possuem manuscritos das seguintes áreas de pesquisa em Física: “*complex networks*”, “*topological insulators*” e “*liquid crystals*”, os quais foram abreviadas neste capítulo como CN, TI e LC, respectivamente. Em todos os experimentos, utilizamos $m = 2$ como parâmetro do classificador κ -NN *fuzzy*.

A abordagem tradicional baseada no uso de padrões de colaboração foi comparada com a estratégia híbrida da seguinte maneira. Para cada par de autores homônimos, empregamos o algoritmo *fuzzy* κ -NN para considerar a informação topológica e hierárquica. A taxa de acerto Γ_H foi obtida para cada valor de λ (ver equação 13.3) e então o melhor λ^* foi determinado. A Figura 13.6 ilustra o comportamento de Γ_H conforme λ varia dentro do intervalo $\lambda \in [0, 1]$ para três pares específicos de autores homônimos. Em todos os casos, a melhor discriminação ocorreu com $\lambda^* > 0$, confirmando que o uso de atributos topológicos é útil para aperfeiçoar a tarefa de desambiguação. Em casos particulares, é possível até mesmo ignorar a estratégia tradicional em favor da estratégia topológica, pois o valor ótimo de Γ_H incluem os casos em que $\lambda = 1$ (veja por exemplo as Figuras 13.6(a) e 13.6(c)). Em outros casos, o melhor Γ_H foi observado através de uma combinação convexa de estratégias, onde $0 < \lambda < 1$ (ver Figura 13.6(b)). A distribuição de λ^* para todos os pares de autores mostrada na Figura 13.7 confirma que a topologia desempenha um papel fundamental na tarefa pois $\lambda^* \neq 0$ em praticamente todos os pares de autores analisados.

As taxas de acertos individuais (Γ_C para a estratégia colaborativa tradicional e Γ_T para a estratégia topológica) foram comparadas com a estratégia híbrida na Tabela 13.6, cujas linhas representam um par específico de autores homônimos. Para cada banco de dados (CN, LC e TI), a tabela mostra os 15 melhores resultados onde a abordagem híbrida apresentou o maior ganho de performance em relação à estratégia tradicional. Em outras palavras, as linhas estão ordenadas em ordem decrescente da margem de ganho $\Gamma_{HC} = \Gamma_H - \Gamma_C$. De acordo com a tabela, a amplitude de Γ_{HC} depende da rede colaborativa analisada. Enquanto nas redes CN e LC os maiores ganhos foram de 32,5% e 30,0%, respectivamente; a rede TI apresentou um ganho máximo Γ_{HC} menor que 10,0%. Estes resultados sugerem que a discriminação desta última rede foi prejudicada por apresentar

Figura 13.6 – Variação da taxa de acerto Γ_H em função de λ para um par específico de autores homônimos nas redes (a) CN; (b) TI e (c) LC. Note que, em (a) e (b) o melhor Γ_H inclui $\lambda = 1$. Isto significa que a melhor taxa de acerto pode ser obtida usando apenas a estratégia topológica em casos particulares. O efeito híbrido pode ser observado em (c) pois existe há melhor combinação apresenta contribuições das estratégia tradicionais e topológicas.

Figura 13.7 – Distribuição de λ^* para as redes (a) CN; (b) TI; e (c) LC. Note que $\lambda^* > 0$ para a maioria dos casos, confirmando a adequabilidade das medidas topológicas para aperfeiçoar o desempenho da tarefa de desambiguação.

uma regularidade de topologia maior que as outras duas redes analisadas.

Tabela 13.6 – Valores de taxa de acerto obtidas com a classificação tradicional (Γ_C), topológica (Γ_T) e híbrida (Γ_H). As linhas estão ordenadas de acordo com a margem de ganho $\Gamma_{HC} = \Gamma_H - \Gamma_C$ de cada banco de dados. Os valores de λ associados ao melhor Γ são apresentados nas últimas colunas. O ganho Γ_{HC} varia de rede para rede e alcança um valor máximo de 32,5% na rede CN.

Redes complexas				Isolantes topológicos				Cristais líquidos			
Γ_C (%)	Γ_T (%)	Γ_H (%)	λ^* –	Γ_C (%)	Γ_T (%)	Γ_H (%)	λ^* –	Γ_C (%)	Γ_T (%)	Γ_H (%)	λ^* –
51.3	82.5	83.8	0.55	88.8	98.8	98.8	1.00	52.5	82.5	82.5	1.00
47.5	75.0	75.0	1.00	35.8	41.6	42.5	0.85	50.0	72.5	72.5	1.00
77.5	98.3	99.2	0.98	62.5	63.8	68.8	0.56	67.5	77.5	85.0	0.77
72.5	90.0	90.0	1.00	28.8	35.0	35.0	1.00	72.5	82.5	87.5	0.98
81.3	97.5	97.5	1.00	18.1	20.6	23.8	0.88	57.5	71.3	71.3	1.00
85.0	97.5	98.8	0.88	95.0	98.7	100.0	0.79	65.0	70.0	77.5	0.53
86.3	96.3	100.0	0.29	18.8	23.8	23.8	1.00	83.8	85.0	95.0	0.60
86.3	93.8	98.8	0.83	61.3	66.3	66.3	1.00	82.5	87.5	92.5	0.31
83.8	95.0	96.3	0.42	90.8	90.8	95.0	0.70	77.5	72.5	87.5	0.48
85.0	94.0	96.3	0.85	20.0	23.3	23.3	1.00	85.0	77.5	92.5	0.29
85.0	95.0	95.0	1.00	97.5	92.5	100.0	0.37	75.0	60.0	82.5	0.03
85.0	87.5	95.0	0.49	31.3	32.5	33.8	0.84	95.0	100.0	100.0	1.00
85.0	84.2	93.4	0.49	28.8	31.3	31.3	1.00	40.0	45.0	45.0	1.00
85.9	94.1	94.2	1.00	26.6	29.2	29.2	1.00	80.0	85.0	85.0	1.00
72.5	76.3	80.0	0.49	33.8	36.3	36.3	1.00	91.3	95.0	95.0	1.00

13.5.1 Relevância das medidas topológicas

Na seção anterior observamos que as medidas topológicas extraídas de redes colaborativas são úteis para aperfeiçoar o desempenho da tarefa de desambiguação. Nesta seção estudamos quais são os atributos topológicos são responsáveis pela ganho de desempenho da tarefa. Para isto, adotamos a estratégia AUC para cada uma das redes analisadas. Seleccionamos para análise apenas os casos cujo desempenho tradicional é fraco e portanto a topologia é de fato responsável por aperfeiçoar a tarefa. Para cada par de autores homônimos calculamos a relevância através da avaliação da taxa de acerto de todas as combinações de atributos possíveis. Dessa forma, se $\mathcal{F}$ representa o número total de medidas, então

$$\sum_{i=0}^{\mathcal{F}} \mathcal{F}! / (\mathcal{F} - i)! i! = 2^{\mathcal{F}} \quad (13.5)$$

classificadores são gerados. Com este procedimento, identificamos as medidas mais relevantes como sendo aquelas aparecendo entre os classificadores com a maiores taxas de acerto. A relevância foi então quantificada a partir da análise da matriz $\mathcal{M}$, composta de

$2^{\mathcal{F}}$ linhas e $\mathcal{F}$ colunas, cujos elementos m_{ij} são dados por :

$$m_{ij} = \begin{cases} 1 & \text{se a } i\text{-ésima melhor combinação usou o } j\text{-ésimo atributo,} \\ 0 & \text{caso contrário.} \end{cases} \quad (13.6)$$

Se um atributo j sempre aparece entre os melhores classificadores, então $f(x) = \sum_{i=1}^x m_{ij}$, $\{x \in \mathbb{N}^* | x \leq 2^{\mathcal{F}}\}$ crescerá rapidamente. Por outro lado, se j aparece entre os piores classificadores, $f(x)$ crescerá significativamente apenas para altos valores de x . Dessa forma, a relevância $AUC(j)$ de um atributo foi computado como a área abaixo da curva $f(x)$:

$$AUC(f) = \int_1^{2^{\mathcal{F}}} f(x) dx = \sum_{i=1}^{2^{\mathcal{F}}} \sum_{k=1}^{i-1} m_{kj} + \frac{1}{2} \sum_{i=1}^{2^{\mathcal{F}}} m_{ij}. \quad (13.7)$$

A relevância dos atributos em cada rede colaborativa, calculada de acordo com a equação 13.7, é mostrada na Tabela 13.7.

Tabela 13.7 – *Rank* (R) de relevância de medidas para as redes colaborativas CN, TI e LC. A distinguibilidade dos pares de autores homônimos foi avaliada de acordo com a quantidade $AUC(j)$ (ver equação 13.7) e o *rank* final de cada rede colaborativa foi tomado como a média geométrica de $AUC(j)$ em cada par de autores homônimos. Em todas as redes colaborativas, o fator topológico mais proeminente foi o comprimento médio dos caminhos mínimos.

Rank	1	2	3	4	5	6	7	8	9	10
Redes	L	$\langle k_n \rangle$	Δs_n	Δk_n	$\langle s_n \rangle$	$\mathcal{L}$	B	k	C	s
Isolantes	L	C	B	Δk_n	Δs_n	$\langle s_n \rangle$	k	$\langle k_n \rangle$	s	$\mathcal{L}$
Cristais	L	k	Δk	s	$\langle s_n \rangle$	B	C	Δs	$\langle k_n \rangle$	$\mathcal{L}$

- **Comprimento médio dos caminhos mínimos (L):** esta medida representa o atributo mais importante da tarefa de discriminação. Diferentemente de outras métricas tradicionais, esta medida considera não apenas os padrões de conectividade local, mas também padrões de conectividade mais distantes. Este resultado sugere, portanto, que padrões de conexões de longo alcance são essenciais na caracterização de autores.
- **Graus (k , $\langle k_n \rangle$ e Δk_n):** estas medidas são relevantes apenas em casos particulares. $\langle k_n \rangle$ aparece na segunda posição para a rede CN, enquanto que o número de vizinhos k representa o segundo atributo mais importante na rede LC. Para a rede TI, ambos k e $\langle k_n \rangle$ apareceram na parte inferior da tabela de relevância provavelmente porque as conexões entre vizinhos desempenham um papel fundamental na discriminação de autores ambíguos e esta característica somente é capturada com o coeficiente de aglomeração (ver discussão abaixo). O desvio padrão dos graus dos vizinhos não parece ser relevante porque esta medida se encontra em posições intermediárias na tabela de relevância.

- **Strength** (s , $\langle s_n \rangle$ e Δs_n): o *strength* calculado nos primeiros níveis hierárquicos foi classificado em posições intermediárias. A introdução do peso das arestas não parece melhorar significativamente a discriminabilidade pois as diferenças $AUC(k) - AUC(s)$ e $AUC(\langle k_n \rangle) - AUC(\langle s_n \rangle)$ são relativamente pequenas.
- **Betweenness** (B): o número de caminhos mínimos passando por vértices representando autores homônimos desempenham um papel inferior na tarefa. O melhor *rank* alcançado foi $r(B) = 3$ na rede TI. Na verdade, a fraca performance desta quantidade revela que o comprimento dos caminhos e não a sua quantidade é responsável por aperfeiçoar a capacidade discriminativa.
- **Coeficiente de aglomeração** (C): a conectividade dos vizinhos parece ser útil apenas para a rede TI. Este resultado é diferente daquele observado para a rede CN, onde o número de conexões saindo dos vizinhos foi relevante (note que $\langle k_n \rangle$ foi classificado na segunda posição do *rank* da rede CN). De acordo com o *rank* da rede TI, o número total de conexões externas deixando os vizinhos é menos informativo para a tarefa porque C é mais importante do que as outras medidas que consideram conexões externas (Δk , Δs e $\langle k_n \rangle$). Isto significa que, na rede TI, o atributo responsável pela distinguibilidade de autores é o grau de conectividade entre colaboradores. Em outras palavras, autores homônimos podem ser diferenciados entre si pois eles pertencem a comunidades com modularidades distintas.
- **Índice de localidade** ($\mathcal{L}$): definida como a porcentagem de conexões localizada no primeiro nível hierárquico, esta medida é pouco discriminativa pois alcançou, no melhor caso, apenas a sexta posição. Este resultado sugere que o índice de localidade dos autores em redes colaborativas não varia entre autores. Mesmo que o número total de ligações externas (k) e internas (C) sejam relevantes em alguns casos, a proporção de conexões localizadas no círculo concêntrico mais próximo ao vértice em análise não é informativa.

13.5.2 É possível prever o valor de λ^* ?

Durante a concepção do desambiguador híbrido extrínseco foi possível perceber que a escolha exata do parâmetro λ é capaz de fornecer um classificador com desempenho superior à técnica tradicional. A desvantagem associada com esta estratégia está no fato deste classificador não ser diretamente aplicável em problemas reais pois não existe nenhuma informação disponível de antemão para guiar a seleção do melhor λ , isto é, λ^* .

A estratégia que propomos para estimar λ^* é baseada na hipótese de que o uso de atributos topológicos se torna relevante especialmente quando a discriminação obtida

Figura 13.8 – Previsão do valor de λ^* a partir da análise de discriminabilidade obtida com a abordagem tradicional colaborativa. O melhor ajuste obtido para CN, TI e LC foi: (a) $\lambda^* = 3,3222 - 3,389 \times 10^{-2} \Gamma$ ($\text{Corr}(\Gamma_C, \lambda^*) = 0,46$ e $p = 4,8 \times 10^{-3}$); (b) $\lambda^* = 1,026 - 9,3 \times 10^{-3} \Gamma$ ($\text{Corr}(\Gamma_C, \lambda^*) = 0,79$ e $p \simeq 0$) e (c) $\lambda^* = 1,68433 - 1,633 \times 10^{-2} \Gamma$ ($\text{Corr}(\Gamma_C, \lambda^*) = 0,63$ e $p \simeq 0$).

com atributos colaborativos é incapaz de fornecer uma discriminação razoável. Mais especificamente, nossa estratégia associa os maiores valores de λ^* aos piores classificadores tradicionais da seguinte maneira. Inicialmente, avaliamos a discriminabilidade de autores homônimos a partir do cálculo da taxa de acerto Γ_C obtida com atributos colaborativos. Em seguida, o valor de λ^* é identificado com a abordagem híbrida de acordo com a equação 13.3. Finalmente, λ^* é correlacionado com Γ_C e a melhor linha obtida do ajuste dos dados é extraída. Em alguns casos foi necessário escolher um valor particular de λ^* pois o melhor λ ocorreu em intervalos (ver por exemplo Figura 13.6(a) e (c)). Para construir o modelo de regressão que maximiza a capacidade de previsão de λ^* escolhemos os valores de λ^* que maximizam a correlação entre Γ_C e λ^* através do uso da heurística de anelamento simulado. O relacionamento entre Γ_C e λ^* está ilustrada na Figura 13.8. Note que, nas três redes, a correlação entre Γ_C e λ^* alcançou valores significativos ($\text{Corr}(\Gamma_C, \lambda^*) = 0,46$, $\text{Corr}(\Gamma_C, \lambda^*) = 0,79$ e $\text{Corr}(\Gamma_C, \lambda^*) = 0,63$ respectivamente para CN, TI e LC), sugerindo que a ponderação ótima para a estratégia híbrida pode ser razoavelmente prevista a partir da análise da distinguibilidade tradicional.

13.6 Conclusões

Duas abordagens inovadoras foram introduzidas nesta tese para tratar o problema de desambiguação de nomes em redes colaborativas. Na primeira abordagem, estendemos o método tradicional baseado na conectividade de vizinhos imediatos em redes através da incorporação de hierarquias mais profundas. Mostramos que a terceira hierarquia levou a uma performance consideravelmente melhor do que a tradicional. Na segunda abordagem,

usamos pela primeira vez a topologia das redes para a tarefa de desambiguação. As duas medidas mais eficientes para a distinção foram o grau e o *strength* médio dos vizinhos, sugerindo que a distinção depende principalmente da conectividade de vizinhos. Esta conclusão reforça estudos anteriores sugerindo o uso de hierarquias profundas em redes de colaboração.

Os resultados descritos em nosso estudo foram obtidos da análise de subredes do repositório arXiv para as áreas de redes complexas, isolantes topológicos e cristais líquidos. Todas as ambiguidades foram introduzidas artificialmente. Esta opção por um sistema artificial foi escolhida para assegurar a confiabilidade e significância estatística dos dados. Além disso, a robustez da análise foi assegurada pelo emprego de vários métodos de reconhecimento de padrões. As abordagens inovadoras em redes artificiais também foram analisadas em uma rede real de atores de filmes. Em particular, notamos que a combinação de abordagens na estratégia híbrida intrínseca levou a um desempenho melhorado na tarefa de desambiguação.

Vale a pena ressaltar ainda que, em um estudo complementar, derivamos um esquema de classificação extrínseca que considera ao mesmo tempo fatores colaborativos (abordagem tradicional) e topológicos para caracterizar autores em artigos científicos, o que também permitiu uma melhor performance na desambiguação. Este esquema de classificação foi concebido a partir da definição de um classificador cuja inferência final é baseada em uma ponderação convexa de decisões baseadas em padrões colaborativos e topológicos de conexão. Neste estudo, diversos resultados interessantes puderam ser encontrados. Primeiramente, a ponderação relacionada com a estratégia topológica foi diferente de zero na maioria dos casos, o que implica que a abordagem topológica desempenha um papel complementar na caracterização de autores homônimos. Outra descoberta interessante refere-se ao fato de que quando a abordagem tradicional não é capaz de distinguir autores, a abordagem colaborativa serve como uma caracterização complementar. Nestes casos, a distância média entre autores de uma dada área mostrou ser o fator mais relevante para a discriminação, e portanto deve ser considerada em aplicações práticas como um atributo complementar.

Capítulo 14

Conclusões

A linguagem humana pode ser tratada como um sistema complexo (15), que é influenciado por fatores como o contexto histórico, o alvo comunicativo ou a criatividade do escritor. A fim de melhor entender os mecanismos da linguagem, esta tem sido alvo de investigação na busca de padrões universais que possam explicar sua origem, evolução e organização. Entre os padrões mais bem estabelecidos atualmente estão a lei de Zipf (9, 11, 13) e a distribuição log-normal do comprimento das palavras (178) e sentenças (179), que têm sido úteis para fins práticos de reconhecimento de autoria ou criptografia (89). Com a introdução do conceito de redes complexas, a linguagem passou a ser estudada em termos da interação de componentes básicos (palavras (15) ou sentenças (26), por exemplo). Nesta tese, empregamos esta representação para criar técnicas e metodologias de modelagem, processamento e classificação de textos com ênfase no paradigma estrutural. Enquanto aplicações tradicionais de processamento de línguas naturais fazem uso extensivo de informação semântica, no estudo desenvolvido nesta tese investigamos a linguagem focando nos padrões de relacionamento das palavras. Aplicações da modelagem desenvolvida também foram estudadas no contexto de cienciometria aplicada à rede de citações, de colaborações e de similaridade de conteúdo de artigos científicos.

As principais contribuições deste trabalho estão resumidas a seguir.

14.1 Principais contribuições

Nesta tese desenvolvemos técnicas e metodologias capazes de classificar documentos textuais de acordo com a estrutura topológica das redes subjacentes. Para ilustrar a metodologia aplicada, mostramos a rede complexa obtida a partir da modelagem desta tese (Capítulos 1 a 13) como rede. Entre as principais contribuições deste trabalho, podemos citar:

Figura 14.1 – Rede de co-ocorrência de palavras obtida a partir da modelagem do conteúdo deste tese (Capítulos 1 a 13) como rede complexa. As cores representam o grau de conectividade das palavras. Entre as palavras que são *hubs* da rede estão: “texto”, “rede” e “palavra”. O pequeno componente separado refere-se ao subconjunto desta tese que foi escrito no idioma inglês.

- **Análise de informatividade e natureza das métricas:** embora medidas de redes complexas têm sido usadas em várias aplicações textuais, uma análise dos fatores (sintáticos ou semânticos) captados por cada uma delas ainda não havia sido desenvolvida. Neste contexto apresentamos uma análise detalhada da natureza das principais métricas de redes complexas para estudar o estilo de textos. Esta investigação foi motivada pela inexistência de interpretação universal relacionando medidas e características textuais. Dentre os principais resultados verificamos a informatividade das métricas com relação a textos sem significado (textos aleatórios). Também verificamos que medidas calculadas em redes de co-ocorrência são basicamente dependentes da sintaxe da língua, com exceção das medidas de intermitência. As informações de informatividade e dependência na sintaxe são úteis, por exemplo, para identificar a veracidade e similaridade de idioma de textos com conteúdo desconhecido.
- **Análise de estilo de autores e livros:** do ponto de vista prático, merecem destaque os resultados encontrados para identificação de padrões de autores e livros. Ao contrário da metodologia tradicional que se baseia no estudo das propriedades das palavras funcionais, nesta tese verificamos que os padrões de associação de palavras de conteúdo são específicos para cada autor. Em especial, as métricas que mais dependem do estilo foram o caminho mínimo (relacionado às distâncias entre palavras distintas) e a intermitência (relacionada às distâncias de aparições de uma mesma palavra). Portanto, estes resultados sugerem que a distância de colocação de palavras em textos representa uma grandeza fundamental para o estilo. A mesma abordagem aplicada à análise de livros publicados entre 1590 e 1922 revelou que fatores sintáticos captados por medidas de redes estão fortemente ligados a fatores semânticos referentes a um estilo literário. Isto se tornou evidente quando estilos literários puderam ser identificados a partir da análise dos padrões de conectividade de palavras, sugerindo que existe forte correlação entre sintaxe e semântica em redes textuais.
- **Desambiguação de sentido das palavras:** as abordagens tradicionais de desambiguação de sentido das palavras baseiam-se na semântica para identificar o contexto e inferir o significado. Neste estudo propomos uma abordagem complementar que utiliza a conectividade das palavras para inferir o contexto semântico. Os resultados da análise topológica local mostraram ser possível deduzir com uma acurácia significativa o sentido de uma palavra sem usar o contexto semântico. Surpreendentemente, para algumas palavras, esta estratégia topológica foi mais eficiente do que usar os padrões de recorrência de palavras vizinhas. Também propomos uma abordagem topológica no espaço de atributos para desambiguação.

- **Avaliação de traduções automáticas:** neste estudo analisamos como a estrutura da rede textual é afetada pelo processo de tradução automática. Os resultados mostraram que medidas de rede também são úteis para discriminar tradutores automáticos. Foi possível encontrar padrões para discriminar automaticamente a qualidade das traduções.
- **Similaridade de textos:** métricas tradicionais de similaridade de textos são baseadas na sua maioria na semântica das palavras (39). Um exemplo é a tarefa de agrupamento de textos por tópicos. No entanto, em alguns casos é necessário usar uma abordagem estrutural. Neste estudo mostramos que a combinação da abordagem tradicional com a estrutural é mais eficiente para o reconhecimento de qualidade de tradutores automáticos e reconhecimento de autoria.
- **Sumarização extrativa:** neste trabalho estudamos a tarefa de sumarização extrativa baseada na modelagem de redes complexas. Através do uso de novas medidas como *betweenness*, vulnerabilidade (10) e diversidade (37), um desempenho superior a outras estratégias que não utilizam informação semântica profunda foi observado. Este estudo serviu para reforçar que redes são apropriadas para identificar sentenças relevantes no texto.
- **Análise de consistência das palavras:** neste estudo quantificamos a consistência de uso das palavras em textos, medindo-se a taxa de preservação da vizinhança em diferentes textos. Padrões interessantes foram observados. Primeiramente notamos que a distribuição de consistência segue uma log-normal. Também verificamos uma correlação negativa entre taxa de preservação da vizinhança (consistência) com a ambiguidade da palavra e idade de aquisição. Mostrou-se que palavras familiares usadas com frequência são geralmente mais consistentes. Uma comparação de índices de consistência para autores distintos mostrou que estes tendem a usar diferentes padrões de consistência em suas obras. Com isso, foi possível distinguir autores que usam em média palavras com consistência menor (autores de artigos científicos) de autores que usam palavras mais consistentes (autores de romances).
- **Análise de complexidade de textos:** textos com diferentes níveis de simplificação foram representados como redes de co-ocorrência para analisar a relação entre topologia da rede e complexidade textual. Como principal resultado verificou-se que existe uma correlação negativa entre regularidade topológica e a complexidade dos textos. Além disso, em textos de menor complexidade a distância entre conceitos, representados por vértices, tende a ser menor do que em textos complexos. Usando técnicas de análise multivariada, foi possível discriminar textos complexos de simplificados com uma taxa de acerto de 86%.

- **Abordagens de citação:** técnicas foram desenvolvidas para avaliar o processo de citação em artigos científicos considerando valores livres de vieses provenientes de revistas, notoriedade de autores etc. Os critérios desenvolvidos consideram (i) a similaridade de conteúdo com o tópico sendo investigado, a fim de evitar que trabalhos relacionados sejam reproduzidos ou ignorados; e (ii) a importância do artigo dado pela rede de citações. Usando formalismos de redes complexas para dois bancos de dados de artigos do arXiv e da *Web of Science* mostramos que nenhum dos critérios são seguidos na prática. Através da representação de textos como redes complexas, estimamos um índice de similaridade entre textos e descobrimos que a lista de referências não contém os artigos mais similares do banco de dados. A partir da simulação de uma busca sistemática na rede de citações, descobrimos que a frequência de visitas em um caminhada aleatória possui baixa correlação com o número de citações ou *downloads* dos artigos. Estes resultados indicaram que autores e usuários do repositório não seguem os critérios relacionados a uma busca sistemática. Com base neste resultados, desenvolvemos um conceito de cienciometria virtual, o qual acreditamos ser mais útil para avaliar e complementar citações de um dado autor.
- **Modelagem de redes de citação:** elaboramos um modelo que mostra que três fatores são capazes de modelar os padrões de redes de citações. Tal modelo foi capaz não apenas de reproduzir a topologia das redes, mas também as distribuições de similaridade e tempo de publicação. Descobrimos que o fator mais relevante utilizado no processo de citações é a similaridade de conteúdo, enquanto que os outros dois fatores – grau de entrada (número de citações) e idade da publicação – apresentaram importância variando com o tópico estudado. Esta dependência indica que fatores adicionais (como reputação e visibilidade) desempenham um papel importante na modelagem. Através da simulação do comportamento de uma rede de citações artificiais, prevemos o comportamento de índices-h em comunidades com visibilidade distintas. Em especial, verificamos que o índice-h para os autores na comunidade com maior visibilidade evolui com valores levemente maiores do que os valores de uma rede de controle (com visibilidades iguais). Por outro lado, um efeito drástico é causado nas comunidades com menores visibilidades.
- **Desambiguação de redes colaborativas:** neste estudo mostramos que a habilidade de distinção entre autores homônimos é melhorada quando conexões de longa distância são consideradas. Resultados otimizados foram obtidos fazendo uso da terceira hierarquia das conexões. Além disso, uma distinção razoável entre autores pode ser alcançada através do uso de técnicas de reconhecimento de padrões para os dados gerados da topologia das redes colaborativas. A adequabilidade do uso de hierarquias profundas foi também averiguada através da aplicação da metodologia

em redes reais.

14.2 Trabalhos publicados em periódicos internacionais

O desenvolvimento desta tese culminou nas seguintes publicações em periódicos internacionais:

1. AMANCIO, D. R.; ALTMANN, E. G.; RYBSKI, D.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. *Probing the statistical properties of unknown texts: application to the Voynich manuscript*. **PLoS ONE**, v. 8, p. e67310, 2013.
2. SILVA, T. C.; AMANCIO, D. R. *Discriminating word senses with tourist walks in complex networks*. **The European Physical Journal B**, v. 86, p. 297, 2013.
3. SILVA, T. C.; AMANCIO, D. R. *Network-based stochastic competitive learning approach to disambiguation in collaborative networks*. **Chaos** (Woodbury, N.Y.), v. 23, p. 013139, 2013.
4. VIANA, M. P.; AMANCIO, D. R.; COSTA, L. da F. *On time-varying collaboration networks*. **Journal of Informetrics**, v. 7, p. 371-378, 2013.
5. AMANCIO, D. R.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. *Identification of literary movements using complex networks to represent texts*. **New Journal of Physics**, v. 14, p. 043029, 2012.
6. AMANCIO, D. R.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. *Unveiling the relationship between complex networks metrics and word senses*. **Europhysics Letters**, v. 98, p. 18002, 2012.
7. AMANCIO, D. R.; ALUISIO, S. M.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. *Complex networks analysis of language complexity*. **Europhysics Letters**, v. 100, p. 58002, 2012.
8. AMANCIO, D. R.; OLIVEIRA JUNIOR, OSVALDO N.; COSTA, L. da F. *On the use of topological features and hierarchical characterization for disambiguating names in collaborative networks*. **Europhysics Letters**, v. 99, p. 48002, 2012.
9. SILVA, T. C.; AMANCIO, D. R. *Word sense disambiguation via high order of learning in complex networks*. **Europhysics Letters**, v. 98, p. 58001, 2012.

10. AMANCIO, D. R.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. *Using complex networks to quantify consistency in the use of words*. **Journal of Statistical Mechanics**, p. P01004, 2012.
11. AMANCIO, D. R.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. *A decaying factor accounts for contained activity in neuronal networks with no need of hierarchical or modular organization*. **Journal of Statistical Mechanics**, v. 2012, p. P11018, 2012.
12. AMANCIO, D. R.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. *Structure semantics interplay in complex networks and its effects on the predictability of similarity in texts*. **Physica A**, v. 391, p. 4406-4419, 2012.
13. AMANCIO, D. R.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. *Three-feature model to reproduce the topology of citation networks and the effects from authors visibility on their h-index*. **Journal of Informetrics**, v. 6, p. 427-434, 2012.
14. AMANCIO, D. R.; NUNES, M. G. V.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. *Using complex networks concepts to assess approaches for citations in scientific papers*. **Scientometrics**, v. 90, p. 2, 2012.
15. AMANCIO, D. R.; NUNES, M. G. V.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. *Extractive summarization using complex networks and syntactic dependency*. **Physica A**, v. 391, p. 1855-1864, 2012.
16. AMANCIO, D. R.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. *On the concepts of complex networks to quantify the difficulty in finding the way out of labyrinths*. **Physica A**, v. 390, p. 4673-4683, 2011.
17. AMANCIO, D. R.; ALTMANN, E. G.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. *Comparing intermittency and network measurements of words and their dependence on authorship*. **New Journal of Physics**, v. 13, p. 123024, 2011.
18. AMANCIO, D. R.; NUNES, M. G. V.; OLIVEIRA JUNIOR, O. N.; PARDO, T. A. S.; ANTIQUEIRA, L.; COSTA, L. da F. *Using metrics from complex networks to evaluate machine translation*. **Physica A**, v. 390, p. 131-142, 2011.

Atualmente, existem ainda os seguintes trabalhos submetidos para publicação:

19. AMANCIO, D. R.; OLIVEIRA JUNIOR, D. R.; COSTA, L. da F. *Efficient community detection via sampling processes in complex networks*. Disponível em: <<http://arxiv.org/abs/1308.6295>>.

20. AMANCIO, D. R. et al.
A systematic comparison of supervised classifiers.
Disponível em: <<http://arxiv.org/abs/1311.0202>>.
21. AMANCIO, D. R.; OLIVEIRA JUNIOR, D. R.; COSTA, L. da F.
Topological-collaborative approach for disambiguating authors' names.
Disponível em: <<http://arxiv.org/abs/1311.1266>>.

14.3 Publicações em congressos internacionais

O desenvolvimento desta tese culminou nas seguintes publicações em congressos internacionais:

1. AMANCIO, D. R.; FABBRI, R.; OLIVEIRA JUNIOR., O. N.; NUNES, M. G. V.; COSTA, L. da F.
Opinion discrimination using complex network features.
In: **2nd Workshop on Complex Networks**, 2011.
2. AMANCIO, D. R. ; FABBRI, R.; NUNES, M. G. V.; OLIVEIRA JUNIOR., O. N.; COSTA, L. da F.
Distinguishing between positive and negative opinions with complex network features.
In: **TextGraphs-5 - 2010 Workshop on Graph-based Methods for Natural Language Processing**, 2010.

14.4 Limitações das abordagens e trabalhos futuros

Ao longo da tese e na literatura por outros grupos de pesquisa, diversas aplicações foram demonstradas para o uso de propriedades topológicas de redes representando textos. É necessário, todavia, ressaltar algumas limitações dessa abordagem que privilegia a estrutura topológica, sem análise profunda do texto e empregando estatísticas. A primeira limitação é que em praticamente todos os trabalhos estabelecemos correlações entre propriedades do texto e métricas da rede, sem poder demonstrar formalmente que as conclusões são irrefutáveis. Segundo, como são usados métodos estatísticos e de aprendizado de máquina (reconhecimento de padrões), as inferências são feitas a partir de médias e tendências. Sendo impossível tratar corpora excessivamente grandes e sem poder fazer testes exaustivos, não é possível provar que as mesmas conclusões seriam obtidas em corpora distintos, para as mesmas aplicações. Por último, em muitos casos observamos que determinadas métricas

correlacionam com propriedades do texto, mas não temos uma explicação unívoca para a correlação. Ou seja, em algumas aplicações o uso de métricas funciona para uma tarefa de PLN, mas não se sabe exatamente o porquê.

Levando-se em conta as limitações acima, depreende-se que há muito a pesquisar no tópico do uso de redes complexas para modelar textos. É necessário verificar a adequação das métricas, e tentar obter maior grau de formalismo na sua aplicação. Do ponto de vista prático, as investigações conduzidas nesta tese também podem ser continuadas da seguinte maneira. Na tarefa de reconhecimento de autoria/estilo, pode-se analisar as transições de estilo em mais baixo nível. Em outras palavras, pode-se analisar como o estilo de autor evolui como uma série temporal dentro do livro. Esta evolução também será avaliada em termos de sua informatividade durante a discriminação de autoria, já que diferentes autores podem fornecer diferentes características de evolução temporal.

Nesta tese, sempre enfatizamos que a análise estrutural pode ser útil não apenas para identificar padrões significativos no texto, mas também ser combinada com técnicas tradicionais. Pode-se, então, introduzir um método de classificação *fuzzy* (68) para combinar ambas as estratégias num único classificador através de uma combinação linear convexa. Isso geraria uma abordagem híbrida. Também é possível definir modelagens e caracterizações híbridas de modo a captar ambos os aspectos estruturais e semânticos da rede nas aplicações. Como comprovado na análise da natureza das medidas, a maioria das medidas das redes de co-ocorrência capta fatores sintáticos de textos. Dessa forma, sugerimos a introdução de uma modelagem generalizada com tamanho de janela de co-ocorrência variável para capturar também fatores semânticos. Neste caso, será importante detectar comunidades, já que tais redes provavelmente apresentam uma estrutura organizada em campos semânticos.

REFERÊNCIAS

- 1 PAGE, S. E. *Diversity and complexity*. New Jersey: Princeton University Press, 2010. 296 p. ISBN 978-0691137674.
- 2 NEWMAN, M. E. J. *Networks*. Oxford: Oxford University Press, 2010. 720 p. ISBN 978-0199206650.
- 3 COSTA, L. da F.; OLIVEIRA JUNIOR, O. N.; TRAVIESO, G.; RODRIGUES, F. A.; BOAS, P. R. V.; ANTIQUEIRA, L.; VIANA, M. P.; ROCHA, L. E. C. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, v. 60, n. 3, p. 329–412, 2011.
- 4 WATTS, D. J.; STROGATZ, S. H. Collective dynamics of small-world networks. *Nature*, v. 393, p. 440–442, 1998. DOI: 10.1038/30918.
- 5 BARABÁSI, A. L. Scale-free networks: a decade and beyond. *Science*, v. 325, n. 5939, p. 412–413, 2009.
- 6 WATTS, D. J. *Six degrees*. New York: W. W. Norton and Company, 2010. 384 p. ISBN 978-0393325423.
- 7 NEWMAN, M. E. J. Modularity and community structure in networks. *Proceedings of the National Academy Society of the United States of America*, v. 103, n. 23, p. 8577–8582, 2006.
- 8 CLAUSET, A.; MOORE, C.; NEWMAN, M. E. J. Hierarchical structure and the prediction of missing links in networks. *Nature*, v. 453, p. 98–101, 2008. DOI:10.1038/nature06830
- 9 ZIPF, G. K. *Human behavior and the principle of least effort*. Eastford: Martino Fine Books, 2012. 588 p. ISBN 978-1614273127.

-
- 10 COSTA, L. da F.; RODRIGUES, F. A.; TRAVIESO, G.; BOAS, P. R. V. Characterization of complex networks: a survey of measurements. *Advances in Physics*, v. 56, n. 1, p. 167–242, 2007.
- 11 NEWMAN, M. E. J. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, v. 46, n. 5, p. 323–351, 2005.
- 12 SIMON, H. A. On a class of skew distribution functions. *Biometrika*, v. 42, n. 3/4, p. 425–440, 1955.
- 13 CANCHO, R. F.; SOLÉ, R. V. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America*, v. 100, n. 3, p. 788–791, 2003.
- 14 MIEGHEM, P. V. *Graph spectra for complex networks*. Cambridge: Cambridge University Press, 2011. 362 p. ISBN 978-0521194587.
- 15 CANCHO, R. F.; SOLÉ, R. V. The small world of human language. *Proceedings of The Royal Society of London B*, v. 268, n. 1482, p. 2261–2265, 2001.
- 16 FELLBAUM, C.; MILLER, G. *WordNet*. Cambridge: Bradford Book, 1998. 423 p. ISBN 978-0262061971.
- 17 SIGMAN, M.; CECCHI, G. A. Global organization of the wordnet lexicon. *Proceedings of the National Academy of Sciences of the United States of America*, v. 99, n. 3, p. 1742–1747, 2002.
- 18 MOTTER, A. E.; MOURA, A. P. S.; LAI, Y. C.; DASGUPTA, P. Topology of the conceptual network of language. *Physical Review E*, v. 65, p. 065102, 2002. DOI: 10.1103/PhysRevE.65.065102
- 19 PANG, B.; LEE, L. *A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts*. 2004. Disponível em: <dx.doi.org/10.3115/1218955.1218990>. Acesso em: 3 de maio de 2013.
- 20 AMANCIO, D. R.; FABBRI, R.; OLIVEIRA JUNIOR, O. N.; NUNES, M. G. V.; COSTA, L. da F. *Opinion discrimination using complex network features*. 2011. Disponível em: <dx.doi.org/10.1007/978-3-642-25501-4_16>. Acesso em: 3 de maio de 2013.

- 21 AMANCIO, D. R.; FABBRI, R.; OLIVEIRA JUNIOR, O. N.; NUNES, M. G. V.; COSTA, L. da F. *Distinguishing between positive and negative opinions with complex network features*. 2010. Disponível em: <dl.acm.org/citation.cfm?id=1870503>. Acesso em: 3 de maio de 2013.
- 22 AMANCIO, D. R.; ANTIQUEIRA, L.; PARDO, T. A. S.; COSTA, L. da F.; OLIVEIRA JUNIOR, O. N.; NUNES, M. G. V. Complex network analysis of manual and machine translations. *International Journal of Modern Physics C*, v. 19, n. 4, p. 583–598, 2008.
- 23 AMANCIO, D. R.; NUNES, M. G. V.; OLIVEIRA JUNIOR, O. N.; PARDO, T. A. S.; ANTIQUEIRA, L.; COSTA, L. da F. Using metrics from complex networks to evaluate machine translation. *Physica A*, v. 390, n. 1, p. 131–142, 2011.
- 24 AMANCIO, D. R.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. Structure semantics interplay in complex networks and its effects on the predictability of similarity in texts. *Physica A*, v. 391, n. 18, p. 4406–4419, 2012.
- 25 AMANCIO, D. R.; NUNES, M. G. V.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. Extractive summarization using complex networks and syntactic dependency. *Physica A*, v. 391, n. 4, p. 1855–1864, 2012.
- 26 ANTIQUEIRA, L.; NUNES, M. G. V.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. Strong correlations between text quality and complex networks features. *Physica A*, v. 373, p. 811–820, 2007. DOI: 10.1016/j.physa.2006.06.002.
- 27 AMANCIO, D. R.; ALTMANN, E. G.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. Comparing intermittency and network measurements of words and their dependence on authorship. *New Journal of Physics*, v. 13, p. 123024, 2011. DOI: 10.1088/1367-2630/13/12/123024.
- 28 AMANCIO, D. R.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. Unveiling the relationship between complex networks metrics and word senses. *Europhysics Letters*, v. 98, p. 18002, 2012. DOI: 10.1209/0295-5075/98/18002.
- 29 SILVA, T. C.; AMANCIO, D. R. Word sense disambiguation via high order of learning in complex networks. *Europhysics Letters*, v. 98, n. 5, p. 58001, 2012.
- 30 JACKSON, P. *Natural language processing for online applications*. Amsterdam: John Benjamins Pub Co, 2007. 231 p. ISBN 978-9027249937.

-
- 31 KHALIL, H. K. *Nonlinear systems*. Upper Saddler River: Prentice Hall, 2001. ISBN 978-0130673893.
- 32 JANSON, S.; LUCZAK, T.; RUCINSKI, A. *Random graphs*. New York: Wiley-Interscience, 2000. 333 p.
- 33 ERDOS, P.; RÉNYI, A. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, v. 5, p. 17–61, 1960.
- 34 NICOLIS, J. S. *Chaos and information processing*. Singapore: World Scientific Publishing Company, 1991. 284 p.
- 35 LI, W. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, v. 38, n. 6, p. 1842–1845, 1992.
- 36 MASUCCI, A. P.; RODGERS, G. J. Differences between normal and shuffled texts: structural properties of weighted networks. *Advances in Complex Systems*, v. 12, p. 113–129, 2009. DOI: 10.1142/S0219525909002039.
- 37 TRAVENCOLO, B. A. N.; VIANA, M. P.; COSTA, L. da F. Border detection in complex networks. *New Journal of Physics*, v. 11, n. 6, p. 063019, 2009.
- 38 LIMA, G. F.; MARTINEZ, A. S.; KINOUCHI, O. Deterministic walks in random media. *Physical Review Letters*, v. 87, n. 1, p. 010603, 2001.
- 39 MANNING, C. D.; SCHUETZE, H. *Foundations of statistical natural language processing*. Massachusetts: The MIT Press, 1999. 620 p. ISBN 978-0262133609.
- 40 CORMEN, T. H.; LEISERSIN, C. E.; RIVEST, R. L.; STEIN, C. *Introduction to algorithms*. Massachusetts: The MIT Press, 2009. 1312 p. ISBN 978-0262033848.
- 41 FORNEY, G. D. The Viterbi algorithm. *Proceedings of the IEEE*, v. 61, p. 268–278, 1973.
- 42 CANCHO, R. F.; SOLÉ, R. V.; KOHLER, R. Patterns in syntactic dependency networks. *Physical Review E*, v. 69, n. 5, p. 051915, 2004.

- 43 SOLÉ, R. V.; COROMINAS-MURTRA, B.; VALVERDE, S.; STEELS, L. Language networks: their structure, function, and evolution. *Complexity*, v. 15, n. 16, p. 20–26, 2010.
- 44 ALVAREZ-LACALLE, E.; DOROW, B.; ECKMANN, J. P.; MOSES, E. Hierarchical structures induce long-range dynamical correlations in written texts. *Proceedings of the National Academy Society of the United States of America*, v. 103, n. 21, p. 7956–7961, 2006.
- 45 CANCHO, R. F. Euclidean distance between syntactically linked words. *Physical Review E*, v. 70, n. 5, pt. 2, p. 056135, 2004.
- 46 CANCHO, R. F. Why do syntactic links not cross? *Europhysics Letters*, v. 76, n. 6, p. 1228, 2006.
- 47 STEYVERS, M.; TENENBAUM, J. B. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Sciences*, v. 29, n. 1, p. 41–78, 2005.
- 48 LUHN, H. P. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, v. 2, n. 2, p. 159–165, 1958.
- 49 HERRERA, J. P.; PURY, P. A. Statistical keyword detection in literary corpora. *European Physical Journal B*, v. 63, n. 1, p. 135–146, 2008.
- 50 MONTEMURRO, M. A.; ZANETTE, D. H. Entropic analysis of the role of words in literary texts. *Advances in Complex Systems*, v. 5, n. 7, p. 7–17, 2002.
- 51 MOSTELLER, F.; WALLACE, D. L. *Applied Bayesian and classical inference*. New York: Springer, 1984. 303 p. ISBN 978-0387909912.
- 52 ALTMANN, E. G.; PIERREHUMBERT, J. B.; MOTTER, A. E. Beyond word frequency: bursts, lulls, and scaling in the temporal distributions of words. *PLoS ONE*, v. 4, n. 11, p. e7678, 2009.
- 53 CARPENA, P.; BERNAOLA-GALVÁN, P.; HACKENBERG, M.; CORONADO, A. V.; OLIVER, J. L. Level statistics of words: finding keywords in literary texts and symbolic sequences. *Physical Review E*, v. 79, n. 3, pt. 2, p. 035102, 2009.

-
- 54 KATZ, S. M. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, v. 2, n. 1, p. 15-59, 1996.
- 55 ORTUNO, M.; CARPENA, P.; BERNAOLA-GALVAN, P.; MUNOZ, E.; SOMOZA, A. M. Keyword detection in natural languages and DNA. *Europhysics Letters*, v. 57, n. 5, p. 759, 2002.
- 56 ALTMANN, E. G.; CRISTADORO, G.; ESPOSTI, M. D. On the origin of long-range correlations in texts. *Proceedings of the National Academy of Sciences of the United States of America*, v. 109, n. 29, p. 11582-11587, 2012.
- 57 RYBSKI, D.; BULDYREV, S. V.; HAVLIN, S.; LILJEROS, F.; MAKSE, H. A. Scaling laws of human interaction activity. *Proceedings of the National Academy of Sciences of the United States of America*, v. 106, n. 31, p. 12640-12645, 2009.
- 58 SCHINNER, A. The voynich manuscript: evidence of the hoax hypothesis. *Cryptologia*, v. 31, n. 2, p. 95-107, 2007.
- 59 COSTA, L. da F.; SPORNS, O.; ANTIQUEIRA, L.; OLIVEIRA JUNIOR, O. N. Correlations between structure and random walk dynamics in directed complex networks. *Applied Physics Letters*, v. 91, n. 5, p. 054107, 2007.
- 60 BAUKE, H. Parameter estimation for power-law distributions by maximum likelihood methods. *European Physical Journal B*, v. 58, n. 2, p. 167-173, 2007.
- 61 BISHOP, C. M. *Pattern recognition and machine learning*. New York: Springer, 2007. 738 p. ISBN 978-0387310732.
- 62 SHAKHNAROVICH, G.; DARRELL, T.; INDYK, P. *Nearest-neighbor methods in learning and vision*. Massachussets: The MIT Press, 2006. 280 p. ISBN 978-0262195478.
- 63 QUINLAN, J. R. *C4.5*. California: Morgan Kaufmann, 1992. 302 p. ISBN 978-1558602380.
- 64 SILVA, T. C.; LIANG, Z. Network-based high level data classification. *IEEE Transactions on Neural Networks and Learning Systems*, v. 23, n. 6, p. 954-970, 2012.

- 65 DUDA, R. O. *Pattern classification*. New York: Wiley-Interscience, 2000. 654 p. ISBN 978-0471056690.
- 66 KULLBACK, S.; LEIBLER, R. On information and sufficiency. *Annals of Mathematical Statistics*, v. 22, n. 1, p. 79-86, 1951.
- 67 DOMINGOS, P.; PAZZANI, M. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, v. 29, n. 2-3, p. 103-130, 1997.
- 68 YUAN, B.; KLIR, G. J. *Fuzzy sets and fuzzy logic*. Upper Saddle River: Prentice Hall, 1995. 592 p. ISBN 978-0131011717
- 69 VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. Relative clustering validity criteria: a comparative overview. *Statistical Analysis and Data Mining*, v. 3, n. 4, p. 209-235, 2010.
- 70 MASUCCI, A. P.; RODGERS, G. J. Network properties of written human language. *Physical Review E*, v. 74, n. 2, p. 026102, 2006.
- 71 BELFIELD, R. *The six unsolved ciphers*. Berkeley: Ulysses Press, 1992. 312 p. ISBN 978-1569756287.
- 72 AMANCIO, D. R.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. Using complex networks to quantify consistency in the use of words. *Journal of Statistical Mechanics*, v. 2012, p. P01004, 2012. DOI: 10.1088/1742-5468/2012/01/P01004
- 73 MCKAY, A. T. Distribution of the coefficient of variation and the extended “t” distribution. *Journal of the Royal Statistical Society*, v. 95, n. 4, p. 695-698, 1932.
- 74 PARZEN, E. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, v. 33, n. 3, p. 1065-1076, 1962.
- 75 CARRETERO-CAMPOS, C.; BERNAOLA-GALVAN, P.; CORONADO, A. V.; CARPENA, P. Improving statistical keyword detection in short texts: entropic and clustering approaches. *Physica A*, v. 392, n. 6, p. 1481-1492, 2013.
- 76 LARSEN, R. J.; MARX, M. L. *Introduction to mathematical statistics and its applications*. Upper Saddle River: Pearson, 2011. 768 p. ISBN 978-0321693945.

- 77 EISENSTEIN, J.; SMITH, N. A.; XING, E. P. *Discovering sociolinguistic associations with structured sparsity*. 2011. Disponível em: <www.cc.gatech.edu/~jeisenst/papers/acl2011.pdf>. Acesso em: 3 de maio de 2013.
- 78 RAO, D.; PAUL, M.; FINK, C.; YAROWSKY, D.; OATES, T.; COPPERSMITH, G. *Hierarchical Bayesian models for latent attribute detection in social media*. 2011. Disponível em: <www.cs.jhu.edu/~delip/icwsm.pdf>. Acesso em: 3 de maio de 2013.
- 79 OTT, M.; CHOI, Y.; CARDIE, C.; HANCOCK, J. T. *Finding deceptive opinion spam by any stretch of the imagination*. 2011. Disponível em: <www.aclweb.org/anthology-new/P/P11/P11-1032.pdf>. Acesso em: 3 de maio de 2013.
- 80 MENDENHALL, T. C. The characteristic curves of composition. *Science*, v. 9, n. 214S, p. 237-249, 1887. DOI: 10.1126/science.ns-9.214S.237
- 81 YULE, G. On sentence length as a statistical characteristic of style in prose with application to two cases of disputed authorship. *Biometrika*, v. 30, n. 3/4, p. 363-390, 1938.
- 82 GOLCHER, F. *A new text statistical measure and its application to stylometry*. 2007. Disponível em: <korpling.german.hu-berlin.de/~felix/cl07.pdf>. Acesso em: 3 maio 2013.
- 83 BARONI, M.; BERNARDINI, S. A new approach to the study of translationese: machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, v. 21, n. 3, p. 259-274, 2006.
- 84 WEST, M. D. *Applications of computer content analysis*. Westport: Praeger, 2001. 208 p. ISBN B000QCS6TO.
- 85 HAVLIN, S. The distance between zipf plots. *Physica A*, v. 216, n. 1, p. 148-150, 1995.
- 86 VILENSKY, B. Can analysis of word frequency distinguish between writings of different authors? *Physica A*, v. 231, n. 4, p. 705-711, 1996.
- 87 YANG, A. C. C.; PENG, C. K.; YIEN, H. W.; GOLDBERGER, A. L. Information categorization approach to literary authorship disputes. *Physica A*, v. 329, n. 3/4, p. 473-483, 2003.

- 88 UZUNER, O.; KATZ, B. *Style vs. expression in literary narratives*. 2005. Disponível em: <people.csail.mit.edu/ozlem/sigir-05-cc-UzunerO-cr.pdf>. Acesso em: 3 maio 2013.
- 89 GRIEVE, J. Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing*, v. 22, n. 3, p. 251-270, 2007.
- 90 MICHEL, J.-B.; SHEN, Y. K.; AIDEN, A. P.; VERES, A.; GRAY, M. K.; TEAM, T. G. B.; PICKETT, J. P.; HOIBERG, D.; CLANCY, D.; NORVIG, P.; ORWANT, J.; PINKER, S.; NOWAK, M. A.; AIDEN, E. L. Quantitative analysis of culture using millions of digitized books. *Science*, v. 331, n. 6014, p. 176-182, 2011.
- 91 GOLDER, S. A.; MACY, M. W. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, v. 333, n. 6051, p. 1878-1881, 2011.
- 92 MOCANU, D.; BARONCHELLI, A.; PERRA, N.; B., G.; ZHANG, Q.; VESPIGNANI, A. The twitter of babel: mapping world languages through microblogging platforms. *PLoS ONE*, v. 8, n. 4, p. e61981, 2013.
- 93 LEE, J.; WONG, D. W. S. *Statistical analysis with arcView GIS*. New York: Wiley, 1998. 192 p. ISBN 978-0471348740.
- 94 EVERITT, B. S.; LANDAU, S.; LEESE, M.; STAHL, D. *Cluster analysis*. West Sussex: Wiley, 2011. 346 p. ISBN 978-0470749913.
- 95 ELIZABETHAN era. Disponível em: <en.wikipedia.org/wiki/Elizabethan_era>. Acesso em: 3 de maio de 2013.
- 96 COMPENDIUM of literary movements and respective periods. Disponível em: <sparkcharts.sparknotes.com/lit/literaryterms/section5.php>. Acesso em: 3 maio 2013.
- 97 NEOCLASSICISM. Disponível em: <en.wikipedia.org/wiki/Neoclassicism>. Acesso em: 3 maio 2013.
- 98 ENLIGHTENMENT. Disponível em: <en.wikipedia.org/wiki/Age_of_Enlightenment>. Acesso em: 3 maio 2013.
- 99 GOTHIC fiction. Disponível em: <en.wikipedia.org/wiki/Gothic_fiction>. Acesso em: 3 maio 2013.

- 100 NATURALISM. Disponível em: <[en.wikipedia.org/wiki/Naturalism_\(literature\)](http://en.wikipedia.org/wiki/Naturalism_(literature))>. Acesso em: 3 de maio de 2013.
- 101 MODERNISM. Disponível em: <en.wikipedia.org/wiki/Modernism>. Acesso em: 3 de maio de 2013.
- 102 VIEIRA, V.; FABBRI, R.; TRAVIESO, G.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. A quantitative approach to evolution of music and philosophy. *Journal of Statistical Mechanics*, v. 2012, n. 8, p. P08010, 2012.
- 103 SHADBOLT, N.; BERNERS-LEE, T.; HALL, W. The semantic web revisited. *IEEE Intelligent Systems and Their Applications*, v. 21, n. 3, p. 96-101, 2006.
- 104 KOEHN, P. *Statistical machine translation*. New York: Cambridge University Press, 2010. 446 p. ISBN 978-0521874151.
- 105 NAVIGLI, R. Word sense disambiguation: a survey. *ACM Computing Surveys*, v. 41, n. 2, p. 1-69, 2009. ISSN 0360-0300.
- 106 NG, T. H. *Getting serious about word sense disambiguation*. 1997. Disponível em: <acl.ldc.upenn.edu/W/W97/W97-0201.pdf>. Acesso em: 3 de maio de 2013.
- 107 LESK, M. *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone*. 1986. Disponível em: <dl.acm.org/citation.cfm?id=318728>. Acesso em: 3 de maio de 2013.
- 108 GAMALLO, P.; GONZALEZ, M.; AGUSTINI, A.; LOPES, G.; LIMA, V. S. *Mapping syntactic dependencies onto semantic relations*. 2002. Disponível em: <www.sop.inria.fr/acacia/WORKSHOPS/ECAI2002-OLT/Proceedings/Gamallo.pdf>. Acesso em: 3 de maio de 2013.
- 109 VERONIS, J. Hyperlex: lexical cartography for information retrieval. *Computer Speech and Language*, v. 18, n. 3, p. 223-252, 2004.
- 110 ARNOLD, J.; WASOW T, LOSONGCO, A.; GINSTROM, R. Heaviness versus newness: the effects of structural complexity and discourse status on constituent ordering. *Language*, v. 17, n. 1, 2000.

- 111 KOVACS, L.; REPASI, T.; BAKSA-VARGA, E.; BARABAS, P. *Clustering based on context similarity*. 2008. Disponível em: <<http://dx.doi.org/10.1109/CANS.2008.26>>. Acesso em: 28 de novembro de 2013.
- 112 RAPP, R. *Discovering the senses of an ambiguous word by clustering its local contexts*. 2005. Disponível em: <http://dx.doi.org/10.1007/3-540-28084-7_61>. Acesso em: 28 de novembro de 2013.
- 113 WIEBE, J.; MIHALCEA, R. *Word sense and subjectivity*. 2006. Disponível em: <<http://dx.doi.org/10.3115/1220175.1220309>>. Acesso em: 28 de novembro de 2013.
- 114 LIMPET, E.; STAHEL, W. A.; ABBT, M. Log-normal distributions across the sciences: keys and clues. *BioScience*, v. 51, n. 5, p. 341–352, 2001.
- 115 JOHNSON, N. L.; KOTZ, S.; BALKRISHAN, N. *Continuous univariate distributions*. New York: Wiley-Interscience, 1995. 752 p. ISBN 978-0471584940.
- 116 SACHS, L. *Angewandte statistik: anwendung statistischer methoden*. Berlin: Springer-Verlag, 2002. ISBN 978-3540424482.
- 117 GROTH, B. H. A. The golden mean in the inheritance of size. *Science*, v. 39, n. 8, p. 581–584, 1914.
- 118 POWERS, L. The nature of the interaction of genes affecting four quantitative characters in a cross between *Hordeum deficiens* and *H. vulgare*. *Genetics*, v. 21, p. 398–420, 1936.
- 119 SINNOT, E. W. The relation of gene to character in quantitative inheritance. *Proceedings of the National Academy of Sciences*, v. 23, n. 4, p. 224–227, 1937.
- 120 WILLIAMS, C. B. A note on the statistical analysis of sentence length as a criterion of literary style. *Biometrika*, v. 31, p. 356–331.
- 121 HERDAN, G. The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics. *Biometrika*, v. 45, p. 222–228, 1958.

-
- 122 GIBRAT, R. Une loi des reparations économiques: l'effet proportionnel. *Bulletin de la Statistique Générale de la France*, v. 19, p. 469, 1930.
- 123 GIBRAT, R. *Les inégalités économiques*. Paris: Libraire du Recueil Sirey, 1931.
- 124 MITZENMACHER, M. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, v. 1, n. 2, p. 226–251, 2004.
- 125 COLTHEART, M. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, v. 33, p. 497–505, 1981.
- 126 MILLER, G. A. WordNet: a lexical database for english. *Communications of the ACM*, v. 38, n. 11, p. 39–41, 1995.
- 127 KISS, G. R.; ARMSTRONG, C.; MILROY, R.; PIPER, J. *An associative thesaurus of English and its computer analysis*. Edinburgh: University Press, 1973.
- 128 JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data* New York: Prentice Hall. 304 p. 1988.
- 129 KATZ, L. A new status index derived from sociometric analysis *Psychometrika* v. 18, n. 39, 1953
- 130 ARNOLD, D. J.; BALKAN, L.; MEIJER, S.; HUMPHREYS, R. L.; SADLER, L. *Machine translation: an introductory guide*. London: Blackwells-NCC. 200 p. 1993. ISBN: 1855542464.
- 131 DORR, B. J.; JORDAN, P. W.; BENOIT, J. W. A survey of current paradigms in machine translation. *Advances in Computers*, v. 49, n. 2, p. 1–68, 1999.
- 132 BROWN, P. F.; COCKE, J.; DELLA PIETRA, S.; DELLA PIETRA, V. J.; JELINEK, F. J. D.; LAFFERTY, R. L.; P. S., ROOSSIN. A statistical approach to machine translation. *Computational Linguistics*, v. 16, n. 2., p. 79–85, 1990.
- 133 BROWN, P.F.; DELLA PIETRA, S.; DELLA PIETRA, V. J.; MERCER, R. L. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, v. 19, p. 263–312, 1993.

- 134 OCH, F. J.; TILLMANN, C.; NEY, H. *Improved alignment models for statistical machine translation*. 1999. Disponível em: <<http://acl.ldc.upenn.edu/W/W99/W99-0604.pdf>>. Acesso em: 28 de novembro de 2013.
- 135 OCH, F.; GILDEA, D.; KHUDANPUR, S.; SARKAR, A.; YAMADA, K.; FRASER, A.; KUMAR, S.; SHEN, L.; SMITH, D.; ENG, K.; JAIN, V.; JIN, Z.; RADEV, D. *A smorgasbord of features for statistical machine translation*. 2004. Disponível em: <<http://acl.ldc.upenn.edu/N/N04/N04-1021.pdf>>. Acesso em: 28 de novembro de 2013.
- 136 YAMADA, K.; KNIGHT, K. *A syntax-based statistical translation model*. 2001. Disponível em: <<http://dx.doi.org/10.3115/1073012.1073079>>. Acesso em: 28 de novembro de 2013.
- 137 KOEHN, P.; OCH, F. J.; MARCU, D. *Statistical phrase-based translation*. 2003. Disponível em: <<http://www.isi.edu/marcu/papers/phrases-hlt2003.pdf>>. Acesso em: 28 de novembro de 2013.
- 138 PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W. J. *BLEU: a method for automatic evaluation of machine translation*. 2002. Disponível em: <<http://dx.doi.org/10.3115/1073083.1073135>>. Acesso em: 28 de novembro de 2013.
- 139 SORICUT, R.; BRILL, E. *A unified framework for automatic evaluation using N-gram co-occurrence statistics*. 2004. Disponível em: <<http://dx.doi.org/10.3115/1218955.1219033>>. Acesso em: 28 de novembro de 2013.
- 140 NIST open evaluation of machine translation. Disponível em: <<http://www.itl.nist.gov/iad/894.01/tests/mt>>. Acesso em: 28 de novembro de 2013.
- 141 COSTA, L. da F.; SILVA, F. N. Hierarchical characterization of complex networks. *Journal of Statistical Physics*, v. 125, 841–872, 2006.
- 142 REVISTA pesquisa fapesp. Disponível em: <<http://revistapesquisa.fapesp.br>>. Acesso em: 28 de novembro de 2013.
- 143 APERTIUM. Disponível em: <<http://www.apertium.org>>. Acesso em: 28 de novembro de 2013.

- 144 INTERTRAN. Disponível em: <<http://www.tranexp.com:2000/Translate/result.shtml>>. Acesso em: 28 de novembro de 2013.
- 145 FREE translation. Disponível em: <<http://www.freetranslation.com>>. Acesso em: 28 de novembro de 2013.
- 146 CASELI, H. M.; NUNES, M. G. V.; FORCADA, M. L. *LIHLA: A lexical aligner based on language-independent heuristics*. 2005. Disponível em: <<http://www.nilc.icmc.usp.br/nilc/download/CaNuFoENIA05.pdf>>. Acesso em: 28 de novembro de 2013.
- 147 TELLES, G. P.; MINGHIM, R.; PAULOVICH, F. V. Normalized compression distance for visual analysis of document collections. *Computers and Graphics*, v. 31, n. 3, p. 327–337. 2007.
- 148 WARD, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, v. 58, n. 30, p. 236–244, 1963.
- 149 HASTIE, T.; TIBSHIRANI, T.; FRIEDMAN, J. *The elements of statistical learning*. New York: Springer. 745 p. 2011. ISBN 978-0387848570.
- 150 COSTA, L. da F.; ROCHA, L. E. C. A generalized approach to complex networks. *European Physical Journal B*, v. 50, n. 1-2, p. 237–242, 2006.
- 151 SPRINZAK, E.; MARGALIT, H. Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology*, v. 311, n. 4, p. 681–692, 2001.
- 152 PATEL, A.; RADEV, D. *Lexical similarity can distinguish between automatic and manual translations*. 2006. Disponível em: <<http://wenku.baidu.com/view/46acc4c78bd63186bcebbc69.html>>. Acesso em: 28 de novembro de 2013.
- 153 CHAUDRON, C. Vocabulary elaboration in teachers' speech to L2 learners. *Studies in second language acquisition*, v. 4, p. 170–180, 1982.
- 154 YANO, Y.; LONG, M. H.; ROSS, S. The effects of simplified and elaborated texts on foreign language reading comprehension. *Language Learning*, v. 44, p. 189–219, 1994.

- 155 BLAU, E. K. The effect of syntax on readability for ESL students in Puerto Rico. *TESOL Quarterly*, v. 16, p. 517–528, 1982.
- 156 HAWKINS, J. A parsing theory of word order universals. *Linguistic Inquiry*, v. 21, n. 2, p. 223–261, 1990.
- 157 HAWKINS, J. *A performance theory of order and constituency*. 1994. Cambridge: Cambridge University Press. ISBN 978-0521378673.
- 158 SZMRECSNYI, B. On operationalizing syntactic complexity. *Journes internationals d'Analyse statistique des Donnes Textualles*, v. 7, p. 1031–1038, 2004.
- 159 DALE, E.; CHALL, J. S. A formula for predicting readability. *Educational Research Bulletin*, v. 27, p. 11–20, 1948.
- 160 FLESCH, R. F. A new readability yardstick. *Journal of Applied Psychology*, v. 32, 221–233, 1948.
- 161 COLEMAN, E. B. On understanding prose: some determiners of its complexity. *National Science Foundation*, Technical Report GB-2604, 1965.
- 162 HOLQUIST, J. B. *A determination of whether the Dale-Chall readability formula may be revised to evaluate more validly the readability of high school science materials*. PhD thesis. Colorado: Colorado State University. 1968.
- 163 GUNNING, R. *The technique of clear writing*. 1952. New York: McGraw-Hill. ISBN 978-0070252066.
- 164 MCLAUGHLIN, G. H. SMOG grading: a new readability formula. *Journal of Reading*, v. 12, p. 639–646, 1969.
- 165 ALUISIO, S. M.; SPECIA, L.; GASPERIN, C.; SCARTON, C. *Readability assessment for text simplification*. 2010. Disponível em: <<http://dl.acm.org/citation.cfm?id=1866796>>. Acesso em: 29 de novembro de 2013.
- 166 LIU, H.; HU, F. What role does syntax play in a language network? *Europhysics Letters*, v. 83, n. 18002, 2008.

- 167 SPARCK, K. J. *Automatic summarising: factors and directions*. 1999. Disponível em: <<http://www.cl.cam.ac.uk/archive/ksj21/ksjdigipapers/summbook99.pdf>>. Acesso em: 29 de novembro de 2013.
- 168 ANTIQUEIRA, L.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F.; NUNES, M. G. V. A complex network approach to text summarization. *Information Sciences*, v. 179, n. 5, p. 584–599, 2009.
- 169 SEDGEWICK, R. *Algorithms in C*. Boston: Addison-Wesley. 1998. 720 p. ISBN 978-0201314526.
- 170 LIN, C. Y.; HOVY, E. *Automatic evaluation of summaries using n-gram co-occurrence statistics*. 2003. Disponível em: <<http://dx.doi.org/10.3115/1073445.1073465>>. Acesso em: 29 de novembro de 2013.
- 171 TEXT analysis conference. Disponível em: <<http://www.nist.gov/tac/>>. Acesso em: 29 de novembro de 2013.
- 172 LEITE, D.; RINO, L.; PARDO, T. A. S.; NUNES, M. G. V. *Does more linguistic knowledge make a difference?* 2007. Disponível em: <<http://eprints.eemcs.utwente.nl/11548/01/W07-0203.pdf>>. Acesso em: 29 de novembro de 2013.
- 173 SPEARMAN, C. The proof and measurement of association between two things. *American Journal of Psychology*, v. 15, p. 72–101, 1904.
- 174 YNGVE, V. H. *A model and a hypothesis for language structure*. 1960. Disponível em: <<http://idiom.ucsd.edu/~rlevy/teaching/esslli2009/yingve-1960.pdf>>. Acesso em: 29 de novembro de 2013.
- 175 BALL, P. Index aims for fair ranking of scientists. *Nature*, v. 436, p. 900, 2005.
- 176 BORNMANN, L.; SCHIER, H.; MARX, W.; DANIEL, H-D. What factors determine citation counts of publications in chemistry besides their quality? *Journal of Informetrics*, v. 6, p. 11–18, 2012.
- 177 GARFIELD, E. Citation analysis as a tool in journal evaluation. *Science*, v. 178, p. 471–479, 1972.

- 178 WIMMER, G.; KOHLER, R.; GROTHJAHN, R.; ALTMANN, G. Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, v. 1, n. 1, p. 98-106, 1994.
- 179 YULE, G. U. On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. *Biometrika*, v. 30, n. 3/4, p. 363-390, 1939.
- 180 ROTH, C.; WU, J.; LOZANO S. Assessing impact and quality from local dynamics of citation networks. *Journal of Informetrics*, v. 6, n. 1, p. 111-120.
- 181 RESCHER, N. *Epistemetrics*. 126 p. 2011. Cambridge: Cambridge University Press. ISBN 978-0521178501.
- 182 HARNAD, S. Open access scientometrics and the UK research assessment exercise. *Scientometrics*, v. 79, n. 1, p. 147-156, 2009.
- 183 ALBERT, R.; BARABASI, A. L. Statistical mechanics of complex networks. *Review of Modern Physics*, v. 74, p. 47-97, 2002.
- 184 BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, Y.; D.-U., HWANG. Complex networks: structure and dynamics. *Physics Reports*, v. 424, n. 4-5, p. 175-308, 2006.
- 185 PRICE, D. J. DE S. Networks of scientific papers. *Science*, v. 149, p. 510-515, 1965.
- 186 AMANCIO, D. R.; NUNES, M. G. V.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. Using complex networks concepts to assess approaches for citations in scientific papers. *Scientometrics*, v. 90, n. 2,
- 187 GENG, X. M.; WANG, Y. Degree correlations in citation networks model with aging. *Europhysics Letters*, v. 88, n. 3, p. 38002, 2009.
- 188 KAMALIKA, B. H. Aging in citation networks. *Physica A*, v. 346, p. 44-48, 2005.
- 189 MENCZER, F. Evolution of document networks. *Proceedings of the National Academy of Science*, v. 101, n. 1, p. 5261-5265, 2004.

-
- 190 WU, Z. X.; HOLME P. Modeling scientific-citation patterns and other triangle-rich acyclic networks. *Physical Review E*, v. 80, p. 037101, 2009.
- 191 COSTAS, R., BORDONS, M. The h-index: advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics*, v. 1, n. 3, p. 193–203, 2007.
- 192 HIRSCH, J. E. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Science*, v. 102, n. 46, p. 16569–16572, 2005.
- 193 PRESS, W. H.; TEUKOLSKY, S. A.; VETTERLING, W. T.; FLANNERY, B. P. *Numerical recipes: the art of scientific computing*. 2007. Cambridge: Cambridge University Press. ISBN 978-0521880688.
- 194 De GROOTE, S. L.; SHULTZ, M.; DORANSKI, M. Online journals' impact on the citation patterns of medical faculty. *Journal of the Medical Library Association*, v. 93, n. 2, p. 223–228, 2005.
- 195 STREMERSCHE, S.; VERNIERS, I.; VERHOEF, P. C. The quest for citations: drivers of article impact. *Journal of Marketing*, v. 71, n. 3, p. 171–193, 2007.
- 196 HIGHLY cited research. Disponível em: <<http://highlycited.com>>. Acesso em: 29 de novembro de 2013.
- 197 EGGHE, L. *Power laws in the information production process: Lotkaian informetrics*. 2005. Bingley: Emerald Group Publishing Limited. ISBN 978-0120887538.
- 198 ALONSO, S.; CABRERIZO, F. J.; HERRERA-VIEDMA, E.; HERRERA, F. h-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, v. 3, n. 4, p. 273–289, 2009.
- 199 De MEY, M. *The cognitive paradigm*. 1982. Chicago: University Of Chicago Press. ISBN 978-0226142593.
- 200 MCCLELLAN, J. E. Specialist control: the publications committee of the Academie Royal des Sciences. *Transactions of the American Philosophical Society*, v. 93, p. 1700–1793, 2003.

- 201 GROSS, P. L. K.; GROSS, E. M. College libraries and chemical education. *Science*, v. 66, 385–389, n. 1713, 1927.
- 202 BÖRNER, K.; MARUS, J. T.; GOLDSTONE, R. L. The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Science*, v. 101, n. 1, p. 5266–5273, 2004.
- 203 BHARVI, D.; GARG, K. C.; BALI, A. Scientometrics of the international journal Scientometrics. *Scientometrics*, v. 56, n. 1, p. 81–93, 2003.
- 204 SHIRABE, M.; FUJIGAKI, Y. The introduction of economic methods to scientometrics: the citing-cited table and the autopoietic systems of citations. *Scientometrics*, v. 47, n. 1, p. 117–130, 2000.
- 205 VANDERELST, D.; SPEYBROECK, N. Scientometrics reveals funding priorities in medical research policy. *Journal of Informetrics*, v. 7, n. 1, p. 240–247, 2013.
- 206 LEYDESDORFF, L. Does the specification of uncertainty hurt the progress of scientometrics? *Journal of Informetrics*, v. 7, n. 2, p. 292–293, 2013.
- 207 SUGIMOTO, C. R.; LARIVIÈRE, V.; NIA, C.; CRONIN, V. Journal acceptance rates: A cross-disciplinary analysis of variability and relationships with journal measures. *Journal of Informetrics*, v. 7, n. 4, p. 897–906, 2013.
- 208 WRIGHT, M.; ARMSTRONG, J. S. The Ombudsman: verification of citations: faulty towers of knowledge? *Interfaces*, v. 38, p. 125–139, 2008.
- 209 LILIEN, G. L. The Ombudsman: who's at fault at faulty towers? Commentaries on the citation dilemma. *Interfaces*, v. 38, p. 123–124, 2008.
- 210 ARXIV.org e-Print archive. Disponível em: <<http://arXiv.org>>. Acesso em: 7 de dezembro de 2013.
- 211 WEB of Science. Disponível em: <<http://thomsonreuters.com/web-of-science>>. Acesso em: 7 de dezembro de 2013.
- 212 PATRICK, D. A measure of standing of journals in stratified networks. *Journal of the American Society for Information Science*, v. 8, n.5-6, p. 341–363, 1985.

- 213 PASULA, H.; MARTHI, B.; MILCH, B.; RUSSELL, S.; SHPITSER, I. *Advances in Neural Information Processing Systems* 2003. Cambridge: MIT Press.
- 214 COTTA, C.; MERELO, J. J. *The complex network of evolutionary computation authors: an initial study*. 2005. Disponível em: <<http://arxiv.org/abs/physics/0507196>>. Acesso em: 30 de novembro de 2013.
- 215 AMANCIO, D. R.; OLIVEIRA JUNIOR, O. N.; COSTA, L. da F. Three-feature model to reproduce the topology of citation networks and the effects from authors' visibility on their h-index. *Journal of Informetrics*, v. 6, n. 3, p. 427–434, 2012.
- 216 MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. 496 p. 2008. Cambridge: Cambridge University Press. ISBN 978-0521865715.
- 217 LIU, Y.; NICULESCU-MIZIL, A.; GRYC, W. *Topic-link Linear Discriminant Analysis: joint models of topic and author community*. 2009. Disponível em: <<http://dx.doi.org/10.1145/1553374.1553460>>. Acesso em: 30 de novembro de 2013.
- 218 HUANG, S.; YU, Y.; XUE, G-R.; ZHANG, B-Y.; CHEN, Z.; MA, W-Y. TSSP: multi-features based reinforcement algorithm to find related papers. *Web Intelligence and Agent Systems*, v. 4, n. 3, p. 271–287, 2006.
- 219 SAO/NASA astrophysics data system. Disponível em: <<http://adsabs.harvard.edu>>. Acesso em: 30 de novembro de 2013.
- 220 WHITE, M. D.; WANG, P. L. A qualitative study of citing behavior: contributions criteria, and meta-level documentation concerns. *Library Quarterly*, v. 67, p. 122–154, 1997.
- 221 CRONIN, B. Norms and functions in citation - the view of journal editors and referees in psychology. *Social Science Information Studies*, v. 2, p. 65–78, 1982.
- 222 VINKLER, P. A quasi-quantitative citation model. *Scientometrics*, v. 12, p. 47–72, 1987.
- 223 WANG, M.; YU, G.; YU, D. Effect of the age of papers on the preferential attachment in citation networks. *Physica A*, v. 388, n. 19, p. 4273–4276, 2009.

- 224 GINGRAS, Y.; LARIVIÈRE, V.; ARCHAMBAULT, É. Literature citations in the Internet era *Science*, v. 323, n. 5910 p. 36, 2009.
- 225 PETERS, H. P. F.; VAN RAAN, A. F. J. On determinants of citations scores - a case study in chemical engineering *Journal of the American Society for Information Science*, v. 27, p. 292–306, 1994.
- 226 KING, J. A review of bibliometric and other science indicator and their role in research evaluation. *Journal of Information Science*, v. 13, p. 261–276, 1987.
- 227 WHITE, H. D. Authors as citers over time. *Journal of the American Society for Information Science and Technology*, v. 52, p. 87–108, 2001.
- 228 MACROBERTS, M. H.; MACROBERTS, B. R. Citation content analysis of a botany journal. *Journal of American Society for Information Science*, v. 48, p. 274–275, 1997.
- 229 VAN RAAN, A. F. J. For your citations only? *Scientometrics*, v. 59, p. 467–472, 2005.
- 230 LAWRENCE, S. Free online availability substantially increases a paper's impact. *Nature*, v. 411, p. 521, 2001.
- 231 BORNMANN, L.; DANIEL, H-D. What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, v. 64, p. 45–80, 2008.
- 232 VELHO, L. The meaning of citation in the context of a scientifically peripheral country. *Scientometrics*, v. 9, n. 1–2, p. 71–89, 1986.
- 233 LANCASTER, F. W.; LEE, S.-Y. K.; DILUVIO, C. Does the place of publication influence citation behavior? *Scientometrics*, v. 19, n. 3–4, p. 239–244, 1990.
- 234 MARTINS, W. S.; GONÇALVES, M. A.; LAENDER, A. H. F.; ZIVIANI, N. Assessing the quality of scientific conferences based on bibliographic citations. *Scientometrics*, v. 83, n. 1, p. 133–155, 2010.
- 235 THOMAS, J.; HARDEN, A.; OAKLEY, A.; OLIVER, S.; SUTCLIFFE, K.; REES, R.; BRUNTON, G.; KAVANAGH, J. Integrating qualitative research with trials in systematic

reviews: an example from public health. *British Medical Journal*, v. 328, p. 1010–1012, 2004.

236 RADICCHI, F.; FORTUNATO, S.; MARKINES, B.; VESPIGNANI, A. Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, v. 80, p. 056103, 2009.

237 CITESEERX. Disponível em: <<http://citeseerx.ist.psu.edu>>. Acesso em: 7 de dezembro de 2013.

238 DBLP: DBLP Bibliography. Disponível em: <<http://dblp.uni-trier.de>>. Acesso em: 7 de dezembro de 2013.

239 GOOGLE Acadêmico. Disponível em: <<http://scholar.google.com.br>>. Acesso em: 7 de dezembro de 2013.

240 SARAWAGI, S.; BHAMIDIPATY, A. *Interactive deduplication using active learning*. 2002. Disponível em: <<http://dx.doi.org/10.1145/775047.775087>>. Acesso em: 7 de dezembro de 2013.

241 KANG, I-S.; NA, S-H.; LEE, S.; JUNG, H.; KIM, P.; SUNG W-K.; LEE, J-H. *Information Processing and Management*, v. 45, n. 1, p. 84–97, 2009.

242 HAN, J.; ZHA, H. Y.; GILES, C. L. *Name disambiguation in author citations using a K-way spectral clustering method*. 2005 Disponível em: <<http://dx.doi.org/10.1145/1065385.1065462>>. Acesso em: 7 de dezembro de 2013.

243 TORVIK, V. I.; WEEBER, M.; SWANSON, D.R.; SMALHEISER, N. R. A probabilistic similarity metric for Medline records: a model for author name disambiguation. *Journal of the American Society for Information Science and Technology*, v. 56, n. 2, p. 140–158, 2005.

244 MCRAE-SPENCER, D.; SHADBOLT, N. R. *Also by the same author: AKTi-veAuthor, a citation graph approach to name disambiguation*. 2006. Disponível em: <<http://ieeexplore.ieee.org/xpl/abstractAuthors.jsp?arnumber=4119097>>. Acesso em: 7 de dezembro de 2013.

245 ZHAO, D.; STROTSMANN, A. Comparing all-author and first-author co-citation analyses of information science. *Journal of Informetrics*, v. 2, n. 3, p. 229, 2008.

- 246 MALIN, B. *Unsupervised name disambiguation via social network similarity*. 2005. Disponível em: <http://www.scimaps.org/exhibit/docs/self_org_author_disamb.pdf>. Acesso em: 7 de dezembro de 2013.

Apêndice A – Lista de *stopwords* para o Inglês

A lista abaixo ilustra as *stopwords* em Inglês que foram utilizadas no pré-processamento dos textos. Portanto, todas estas palavras foram eliminadas da análise de estilo de escrita.

I A ABOARD ABOUT ABOVE ACROSS AFTER AFTERWARDS AGAINST AGIN
AGO AGREED-UPON AH ALAS ALBEIT ALL ALL-OVER ALMOST ALONG ALONG-
SIDE ALTHO ALTHOUGH AMID AMIDST AMONG AMONGST AN AND ANOTHER
ANY ANYONE ANYTHING AROUND AS ASIDE ASTRIDE AT ATOP AVEC AWAY
BACK BE BECAUSE BEFORE BEFOREHAND BEHIND BEHYNDE BELOW BENE-
ATH BESIDE BESIDES BETWEEN BEWTEEN BEYOND BI BOTH BUT BY CA. DE
DES DESPITE DO DOWN DUE DURIN DURING EACH EH EITHER EN EVERY
EVER EVERYONE EVERYTHING EXCEPT FAR FER FOR FROM GO GODDAMN
GOODY GOSH HALF HAVE HE HELL HER HERSELF HEY HIM HIMSELF HIS HO
HOW HOWEVER IF IN INSIDE INSOFAR INSTEAD INTO IT ITS ITSELF LA LE LES
LEST LIEU LIKE ME MINUS MOREOVER MY MYSELF NEAR NEAR-BY NEARER
NEAREST NEITHER NEVERTHELESS NEXT NO NOR NOT NOTHING NOTWITHS-
TANDING O O'ER OF OFF ON ONCE ONE ONESELF ONLY ONTO OR OTHER
OTHERS OTHERWISE OUR OURS OURSELVES OUT OUTSIDE OUTTA OVER
PER RATHER REGARDLESS ROUND SE SHE SHOULD SINCE SO SOME SOME-
ONE SOMETHING THAN THAT THE THEIR THEM THEMSELVES THEN THERE
THEREFORE THESE THEY THINE THIS THOSE THOU THOUGH THROUGH
THROUGHOUT THRU TILL TO TOGETHER TOWARD TOWARDES TOWARDS
UH UNDER UNDERNEATH UNLESS UNLIKE UNTIL UNTO UP UPON UPPON US
VIA VIS-A-VIS VIS-à-VIS WE WELL WHAT WHATEVER WHATSOEVER WHEN
WHENEVER WHERE WHEREAS WHEREFORE WHEREUPON WHETHER WHICH
WHICHEVER WHILE WHO WHOEVER WHOM WHOSE WHY WITH WITHAL
WITHIN WITHOUT YE YEA YEAH YES YET YONDER YOU YOUR YOURS YOUR-
SELF YOURSELVES

Apêndice B – Lista de *stopwords* para o Português

A lista abaixo ilustra as *stopwords* em Português que foram utilizadas no pré-processamento dos textos. Portanto, todas estas palavras foram eliminadas da análise de estilo de escrita.

A AFORA ALGO ALGUM ALGUMA ALGUMAS ALGUNS ALGURES ALGUÉM ALHURES ALIÁS ANTE AO AONDE AOS APENAS APESAR APÓS AQUELA AQUELAS AQUELE AQUELES AQUILO AS ASSIM ATÉ BASTANTE BASTANTES CADA CASO CERTA CERTAS CERTO CERTOS COM COMIGO COMO CONFORME CONOSCO CONQUANTO CONSEQÜENTEMENTE CONSIGO CONSOANTE CONTANTO CONTIGO CONTRA CONTUDO CONVOSCO CUJA CUJAS CUJO CUJOS DA DALI DAQUELA DAQUELAS DAQUELE DAQUELES DAQUI DAQUILO DAS DAÍ DE DELA DELAS DELE DELES DEMAIS DENTRE DEPOIS DESDE DESSA DESSAS DESSE DESSES DESTA DESTAS DESTE DESTES DIANTE DIFERENTES DISSO DISTO DIVERSOS DO DONDE DOS DURANTE E ELA ELAS ELE ELES EM EMBORA ENFIM ENQUANTO ENTANTO ENTRE ENTRETANTO ENTÃO ESSA ESSAS ESSE ESSES ESTA ESTAS ESTE ESTES EU EXCETO FORA GRAÇAS IDEM ISSO ISTO JÁ LHA LHAS LHE LHES LHO LHOS LOGO MAIS MAL MALGRADO MAS ME MEDIANTE MENOS MESMA MESMAS MESMO MESMOS MEU MEUS MIM MINHA MINHAS MUITA MUITAS MUITO MUITOS NA NADA NAQUELA NAQUELAS NAQUELE NAQUELES NAQUILO NAS NELA NELAS NELE NELES NEM NENHUM NENHUMA NENHUMAS NENHUNS NENHURES NESSA NESSAS NESSE NESSES NESTA NESTAS NESTE NESTES NINGUÉM NISSO NISTO NO NOS NOSSA NOSSAS NOSSO NOSSOS NÓS O ONDE ORA OS OU OUTRA OUTRAS OUTREM OUTRO OUTROS OUTROSSIM PARA PATAVINA PELA PELAS PELO PELOS PERANTE POIS POR PORQUANTO PORQUE PORTANTO PORÉM POSTO POUCA POUCAS POUCO POUÇOS PRA PRO PROS PRÓPRIA PRÓPRIAS PRÓPRIO PRÓPRIOS QUAIS QUAISQUER QUAL QUALQUER QUANDO QUANTA QUANTAS QUANTO QUANTOS QUE QUEM QUER SALVO SE SEGUNDO SEJA SEM SEMELHANTE SEMELHANTES SENHOR SENHORA SENHORAS SENHORES SENHORITA SENHO-

RITAS SENÃO SEU SEUS SOB SOBRE SUA SUAS TAIS TAL TAMBÉM TANTA
TANTAS TANTO TANTOS TAS TE TEU TI TODA TODAS TODAVIA TODO TODOS
TOS TRÁS TU TUA TUAS TUDO UM UMA UMAS UNS VERSUS VIA VISTO VOCÊ
VOCÊS VOS VOSSA VOSSAS VOSSO VOSSOS VÁRIA VÁRIAS VÁRIOS VÓS À
ÀQUELA ÀQUELAS ÀQUELE ÀQUELES ÀQUILO ÀS