# Identificação automática de relações multidocumento

***Erick Galani Maziero***

# Identificação automática de relações multidocumento

**Erick Galani Maziero**

**Orientador: Prof. Dr. Thiago Alexandre Salgueiro Pardo**

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

**USP – São Carlos**

**Março de 2012**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

Gi Galani Maziero, Erick
Identificação Automática de Relações Multidocumento / Erick Galani Maziero; orientador Thiago Alexandre Salgueiro Pardo. -- São Carlos, 2012.
106 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2012.

1. Processamento de Língua Natural. 2. Linguística Computacional. 3. Cross-document Structure Theory. I. Alexandre Salgueiro Pardo, Thiago, orient. II. Título.

# Agradecimentos

Ao único Deus. Meu Senhor e Salvador.

À minha amada família.

Ao meu orientador, Thiago Pardo.

Aos meus companheiros do NILC.

À FAPESP.

# Resumo

O tratamento multidocumento mostra-se indispensável no cenário atual das mídias eletrônicas, em que são produzidos diversos documentos sobre um mesmo tópico, principalmente quando se considera a explosão de informação permitida pela *web*. Tanto leitores quanto aplicações computacionais se beneficiam da análise discursiva multidocumento por meio da qual são explicitadas relações entre as porções dos documentos, por exemplo, relações de equivalência, contradição ou de contextualização de alguma informação. A fim de realizar o tratamento automático multidocumento, adota-se neste trabalho a teoria linguístico-computacional CST (*Cross-document Structure Theory*, Radev, 2000). Esse tipo de conhecimento multidocumento permite que (i) se tratem mais apropriadamente fenômenos como redundância, complementariedade e contradição de informações e, consequentemente, (ii) produzam-se sistemas melhores de processamento textual, como buscadores *web* mais inteligentes e sumarizadores automáticos. Neste trabalho é apresentada uma metodologia de identificação dessas relações explorando-se técnicas de aprendizado automático do paradigma tradicional e hierárquico. Para relações que não são passíveis de identificação por aprendizado automático foram desenvolvidas regras para sua identificação. Por fim, um *parser* é gerado contendo classificadores e regras.

# Abstract

The multi-document treatment is essential in the current scenario of electronic media, in which many documents are produced about a same topic, mainly when considering the explosion of information allowed by the web. Both readers and computational applications are benefited by the discursive multi-document analysis, through which the relations (for example, equivalence, contradiction or background relations) among the portions of text are showed. In order to achieve the automatic multi-document treatment, the CST (*Cross-document Structure Theory*, Radev, 2000) is adopted in this work. This kind of knowledge allow (i) the appropriated treatment of phenomena like redundancy, complementarity and contradiction of information and, consequently, (ii) the production of better systems of text processing, as more intelligent web searchers and automatic summarizers. In this work, a methodology to identify these relations is presented exploring techniques of machine learning of the traditional and hierarchical paradigm. For relations with low frequency in the corpus, hand-crafted rules were developed. Finally, a parser is generated containing classifiers and rules.

# Índice Geral

# Índice de Figuras

# Índice de Tabelas

# 1. Introdução

Nos meios eletrônicos, existem muitas fontes que relatam o mesmo assunto com as mesmas ou diferentes perspectivas. Jornais online são um exemplo. Um mesmo evento é relatado por diversos portais de notícias, o que gera diversos documentos sobre um mesmo assunto. O documento da *International Data Corporation* (IDC) com o título "*Extracting value from caos*" (Gantz e Reisel, 2011) diz que em 2011 a *web* produzirá 1.8 zettabytes[1], uma quantidade nove vezes maior do que a produzida cinco anos atrás. Segundo outros relatos, o Google chegaria a processar 24 petabytes[2] de informação por dia.

Um leitor que queira se informar sobre um evento atual poderá recorrer a notícias de um ou mais jornais. Nessa busca, o leitor terá como retorno uma infinidade de textos, sendo imprescindível selecionar apenas alguns para leitura. Tal cenário leva a um grande esforço por parte do leitor, que muitas vezes não obtém informação plena sobre o assunto sobre o qual está interessado. Imagine, por exemplo, uma busca por informação sobre algum conflito regional. Há muitos fatos e eventos envolvidos, assim como perspectivas e opiniões variadas sobre o assunto. O esforço desprendido pelo leitor seria demasiado para recuperar, organizar e ler a informação relevante.

Alguns desses documentos são produzidos, geralmente, logo após o acontecimento de um evento, e outros documentos são gerados a fim de atualizar as informações divulgadas. Por serem textos gerados por diversas fontes e em diversos momentos, estes podem conter trechos que se contradizem, ou que se sobreponham. Por exemplo, as duas sentenças abaixo, S1 e S2, apresentam uma contradição entre si, com relação à quantidade de bombas em um atentado.

(S1) *O prédio da secretaria da Fazenda, no centro, foi atingido por três bombas caseiras.*

(S2) *A Secretaria da Fazenda também foi atingida por uma bomba.*

[1] Unidade de medida de informação que corresponde a $2^{70}$ Bytes

[2] Unidade de medida de informação que corresponde a $2^{50}$ Bytes

Há diversas aplicações na Internet que tentam organizar a informação, como é o caso do GoogleNews[3], que agrupa notícias de jornais online por assuntos, como "Mundo", "Negócios", "Tecnologia", etc. Além disso, o GoogleNews permite a busca por palavras-chave. Esse tipo de aplicação apenas agrupa as informações sobre determinados temas, mas não retira do leitor a necessidade de selecionar as notícias para obtenção das informações, fazendo-se necessário, ainda, a leitura de diversos textos sobre o assunto desejado.

Assim, sente-se a necessidade de uma ferramenta que, além de realizar a busca e agrupamento de textos sobre um tema de interesse, processe todos os documentos encontrados a fim de apresentar os fenômenos multidocumento (explicitados pelos relacionamentos entre as partes dos textos) contidos no grupo de documentos. Essa informação pode ser utilizada por aplicações por meio das quais se busque manipular os documentos ou utilizada para guiar a leitura por um usuário, permitindo-lhe obter uma visão mais panorâmica, facilitando-lhe encontrar o que lhe satisfaça sobre o assunto.

Frente aos desafios de aplicações de PLN mais avançadas, como sumarização multidocumento, recuperação inteligente de informação e sistemas de perguntas e respostas automáticos, identificou-se, com o passar do tempo, a necessidade de uma teoria de estruturação semântico-discursiva multidocumento (por tratar do relacionamento entre os conteúdos das porções textuais). A fim de identificar as relações existentes entre diversos textos que tratam de um mesmo assunto, como no cenário introduzido anteriormente, diversos modelos foram propostos (e serão tratados no Capítulo 2), e dentre esses está a CST (*Cross-document Structure Theory*, Radev, 2000), baseada na RST (*Rhetorical Structure Theory*) (Mann e Thompson, 1987). A CST relaciona as partes de vários textos correlatos via um conjunto previsto de aproximadamente 20 relações. Tal teoria, referenciada como teoria semântico-discursiva multidocumento (ou apenas "discursiva", como na obra original), é apropriada para o tipo de problema apresentado.

[3] News.google.com

Considere os textos apresentados nas figuras 1 e 2 retirados de dois jornais online (Folha[4] e Estadão[5]). Os documentos tratam de um mesmo fato, no caso, as causas e consequências da queda de um avião na República Democrática do Congo. Os textos estão segmentados e cada segmento (nesse caso, sentença) numerado.

**1-** Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.
**2-** A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto.
**3-** Acidentes aéreos são freqüentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.
**4-** O avião acidentado, operado pela Air Traset, levava 14 passageiros e três tripulantes.
**5-** Ele havia saído da cidade mineira de Lugushwa em direção a Bukavu, numa distância de 130 quilômetros.
**6-** Aviões são usados extensivamente para transporte na República Democrática do Congo, um vasto país no qual há poucas estradas pavimentadas.
**7-** Em março, a União Européia proibiu quase todas as companhias aéreas do Congo de operar na Europa. Apenas uma manteve a permissão.
**8-** Em junho, a Associação Internacional de Transporte Aéreo incluiu o Congo num grupo de vários países africanos que classificou como "uma vergonha" para o setor.

**TEXTO 1**

**Figura 1 - Texto 1 com texto a ser relacionado**

**1-** Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas.
**2-** As vítimas do acidente foram 14 passageiros e três membros da tripulação.
**3-** Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 quilômetros do aeroporto de Bukavu.
**4-** Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.
**5-** O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala.
**6-** "Não houve sobreviventes", disse Okala.
**7-** O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais.

**TEXTO 2**

**Figura 2 - Texto 2 com texto a ser relacionado**

Por exemplo, a sentença 1, do documento 1, contém toda a informação presente na sentença 1, do documento 2, e informação adicional. Essa sentença 1, do documento 1, é

[4] http://www.folha.com.br

[5] http://www.estadao.com.br

elaborada pela sentença 3, do documento 2. Desse modo, segundo a CST, procura-se identificar pares de sentenças que contenham alguma relação entre si.

Diversas ferramentas e aplicações podem se beneficiar do conhecimento do relacionamento das partes de um grupo de textos. Ferramentas de análise e visualização de informações textuais podem explicitar essas relações presentes nos textos, auxiliando um usuário na busca por alguma informação. Aplicações de sumarização multidocumento terão as porções de texto cujas informações são total ou parcialmente redundantes (Otterbacher et al., 2002; Jorge et al., 2011; Zahri e Fukumoto, 2011), auxiliando na obtenção da taxa de compressão desejada, evitando informações duplicadas. Por exemplo, ao se gerar o sumário de vários textos que descrevem um mesmo evento, deve-se identificar informação redundante para exclusão, informação inconsistente para evitar a formação de sumários incoerentes, informações complementares, etc. Isso pode, potencialmente, ajudar na produção de resumos melhores. Com a anotação multidocumento já realizada, esses sistemas podem organizar as informações de acordo com a evolução temporal das mesmas, o que seria mais complexo se não houvesse esse tipo de conhecimento na anotação, pois as partes que compõem o resumo são provenientes de diversos documentos, escritos, possivelmente, por diversos autores. Sistemas de perguntas e respostas e de recuperação e extração de informação beneficiam-se também da análise multidocumento, pois podem focar nas porções textuais que versam sobre um determinado elemento, sobre o qual se deseja informação. Essas aplicações podem, inclusive, trazer informações adicionais, relacionadas ao tema desejado; essas informações podem ser de acordo com as preferências do usuário, tal como informações históricas, elaborações, contradições, etc., enriquecendo as respostas. Buscadores *web* disporão de mais informações sobre o relacionamento entre os documentos e suas partes produzindo melhores resultados, atendendo às necessidades dos usuários. Essas informações podem ser utilizadas como elementos a considerar no ranking das páginas a fim de retornar ao usuário documentos que permitam maior diversidade de informações, podendo, inclusive, apresentar os relacionamentos entre as páginas retornadas em uma busca.

Além do benefício para as aplicações, uma teoria semântico-discursiva multidocumento também pode auxiliar diretamente em tarefas de base, como fusão de informação, ordenação temporal de sentenças que narram algum evento e agrupamento de

documentos, entre outras. Sabendo-se o relacionamento entre trechos de texto provenientes de diferentes fontes, mas que tratam do mesmo assunto, pode-se determinar melhor como fundir as informações. Pela exploração das relações discursivas intertextuais, é possível organizar informações temporais sobre um mesmo evento, pois há mais chances de textos similares (de um mesmo assunto, por exemplo) terem relacionamentos entre suas partes do que textos não similares, oferecendo-se, assim, mais informação a um processo de agrupamento de textos (tradicionalmente conhecido como *clustering*).

Sistemas de agrupamentos de texto visam facilitar a tarefa de busca de informações em grandes conjuntos de textos. Por exemplo, o sistema NewsHead[6] atua na Internet buscando textos (notícias de jornais *online*) que versam sobre um mesmo assunto, dado um tópico de interesse, e faz o agrupamento desses textos em subtópicos com vistas a facilitar a leitura por parte do usuário do sistema, que poderá escolher um assunto específico para leitura. A classificação dos textos fica a cargo de um portal de notícias que, recebendo o tópico de interesse, retorna os textos para o sistema, que faz o agrupamento em subtopicos. Embora auxilie o leitor por dividir um tópico em subtópicos, o sistema não reduz a quantidade de informações que o leitor terá que processar manualmente. Os grupos de textos gerados não são relacionados entre si, ficando a cargo do leitor a escolha dos textos para leitura.

As análises multidocumento, atualmente, são realizadas levando-se em consideração principalmente o que está explicitado na superfície textual, isto é, por meio das palavras dos documentos. O processamento textual multidocumento é, então, muito superficial, resultando em sistemas de resultado insatisfatório, além de não permitir que outros sistemas possam explorar melhor o resultado de uma análise multidocumento. Para a língua portuguesa, temos a ausência de um analisador multidocumento que encontre as relações existentes entre as porções de textos de diversos documentos que versam sobre um mesmo assunto, dados os poucos estudos sobre os fenômenos multidocumentos, não se conhecendo a adequação de modelos de relacionamento multidocumento para a língua.

Neste trabalho, objetivam-se a investigação e a automatização de modelos e métodos para identificação de relações multidocumento em textos escritos em Português,

[6] http://www.nilc.icmc.usp.br/nilc/tools/newshead

via a teoria CST. Como resultado dessa investigação, objetiva-se, também, a definição e caracterização das relações propostas pela teoria em questão (ver Apêndice A e tipologia definida na Figura 10) permitindo maior exploração destas características e posterior aproveitamento por outros sistemas.

Uma das hipóteses que guiaram este trabalho é que é possível estabelecer uma tipologia genérica de relações para uma análise multidocumento. Assim, pode-se estabelecer um conjunto de relações, com suas restrições de aplicação, e essas relações serão aplicáveis a qualquer conjunto de textos, resultando em uma estrutura com as porções textuais dos diversos documentos relacionadas entre si.

Outra hipótese é que a teoria multidocumento em foco é aplicável à língua portuguesa e suas relações são passíveis de serem detectadas automaticamente com bons resultados, possibilitando o desenvolvimento e automatização de metodologias para a identificação das relações. Acredita-se, também, que estratégias híbridas sejam necessárias na identificação das relações da teoria, dadas as frequências das mesmas. Algumas são passíveis de um tratamento estatístico, outras de tratamento simbólico.

Uma análise discursiva multidocumento automática para língua portuguesa é algo inédito e, tendo em vista pesquisas que necessitam desse tipo de informação, o trabalho será de grande valia, automatizando o que antes era feito manualmente, proporcionando, também, a possibilidade de análise de grandes conjuntos de textos.

No desenvolvimento do trabalho, foram utilizados técnicas de aprendizado de máquina, especificamente classificadores tradicionais e hierárquicos. Essa exploração foi possível devido à criação, neste trabalho, do córpus CSTNews (Cardoso et al., 2011a) anotado segundo a teoria CST, permitindo a aprendizagem e teste dos modelos aprendidos. Para a identificação de relações cuja frequência no córpus é insuficiente para um bom aprendizado, regras foram desenvolvidas manualmente. O resultado foi o desenvolvimento de um *parser* multidocumento (uma ferramenta que faz a análise automaticamente).

Os textos do córpus utilizado são do gênero jornalístico e será esse gênero textual utilizado neste trabalho. Essa escolha se dá devido ao texto jornalístico ter uma linguagem usual, cotidiana, devido às facilidades de obter notícias de jornais online e à grande quantidade disponível produzida constantemente na *web*; inclusive, diversas fontes

produzem notícias sobre um mesmo assunto, possibilitando a aplicação do relacionamento multidocumento, auxiliando o leitor no tratamento de diversas notícias sobre um assunto de interesse.

No próximo capítulo, é apresentada uma revisão bibliográfica sobre os diversos modelos multidocumento, ferramentas, recursos e *parsing* multidocumento. No Capítulo 3, conceitos de aprendizado de máquina são apresentados e servem para tornar mais clara a apresentação dos experimentos com técnicas de aprendizado automático. No Capítulo 4, o córpus criado e utilizado neste trabalho é apresentado, descrevendo-se sua anotação a suas características numéricas. No Capítulo 5, a metodologia completa de *parsing* multidocumento é apresentada. Especificamente, descrevem-se os experimentos com aprendizado de máquina, que deram origem aos classificadores e as regras que identificam as relações que não foram passíveis de serem identificadas por classificadores. Conclusões e trabalhos futuros são apresentados no último capítulo.

# 2. Revisão Bibliográfica

Nesta seção, são apresentados modelos de análise multidocumento, em especial, a CST (*Cross-document Structure Theory*), foco deste trabalho. São apresentados trabalhos que descrevem a automatização da análise multidocumento e, também, recursos e ferramentas para a análise multidocumento, tanto para a língua inglesa quanto para a portuguesa.

## 2.1. Modelos de análise multidocumento

### 2.1.1. Modelo de Trigg

A modelagem de múltiplos documentos científicos, proposta por Trigg (1983) e Trigg e Weiser (1986), é um dos primeiros esforços empreendidos no tratamento de múltiplos textos como uma estrutura em forma de grafo.

Nesses trabalhos, utiliza-se o formalismo de redes semânticas para estruturar os nós e suas ligações. Esses nós podem ser as porções textuais (*chunks*) ou elementos que indicam a estrutura dos textos (*tocs – table of contents*), a exemplo de índices de documentos, formando uma estrutura hierárquica. Assim, cada *toc* representa uma entrada em um índice de um documento e aponta para um ou mais *chunks*. Os *links* entre os nós indicam a relação entre os nós da rede. É possível a definição de caminhos (*paths*, como listas ordenadas de nós) na estrutura gerada, auxiliando um leitor na leitura sequencial de textos representados pela rede.

Cada *chunk* corresponde a uma porção textual e contém outros campos como, por exemplo, autor e data, e contém um ponteiro para uma porção textual (armazenada em um arquivo). Essa porção textual pode corresponder a um ou mais parágrafos ou até a um texto inteiro. Os *chunks* contêm a indicação dos links que apontam para si e dos links que saem de si. Nos *links*, podem-se conter indicações que orientem o leitor na ordem de leitura dos nós. Por exemplo, sendo dois *chunks*, A e B, ligados por um link L que sai de A para B, indica-se que o nó A é pré-requisito para a leitura de B ou que o nó A deve ser lido depois de B. Assim, as relações que são estabelecidas entre os *chunks* podem conter direcionalidade.

Por exemplo, embora a direcionalidade na estrutura textual (direcionalidade física) seja de um nó A para um nó B, por exemplo, deve-se, primeiramente, ler o nó B e depois o nó A. Isso diferencia a direcionalidade física de uma direcionalidade semântica. Por exemplo, se o nó B é elaborado pelo nó A, a direcionalidade aponta de A para B (A elabora B), mas o leitor deve primeiro ler B e depois ler sua elaboração, presente em A.

Os *links* tornam explícito o relacionamento entre os nós da rede. Eles podem ser divididos em *links* que relacionam o conteúdo do texto e *links* que comentam o conteúdo. Na primeira categoria estão *Summary*, *Argument-by-Analogy*, *Example* e *Continuation*. Na segunda, *Criticism*, *Enviroment-Vacuum*, *Argument-Immaterial* e *Style-Incoherent*. As ligações entre os *chunks* ou *tocs* são realizadas manualmente pelos usuários, que podem dividir as porções textuais em porções menores, facilitando a relação com outros *chunks*. Os usuários podem, também, adicionar comentários ou críticas aos nós da estrutura.

Baseado na modelagem acima, o sistema Textnet, desenvolvido ainda quando as interfaces gráficas não estavam muito desenvolvidas, utiliza algumas estruturas de dados a fim de armazenar porções textuais que são ligadas entre si, formando um grafo que pode ser modificado pelos usuários do sistema. A interface foi baseada em um sistema de janelas e menus para gerenciamento do conteúdo textual.

Um usuário pode navegar na estrutura gerada no Textnet considerando as relações (*links*) entre os *chunks* ou *tocs*. Nessa navegação, o leitor pode ler um texto como em uma leitura sequencial de um assunto, recorrer a porções textuais que elaborem mais um tema ou escolher entre duas ou mais versões de um mesmo assunto.

Quando do desenvolvimento do sistema, os autores justificaram a não automatização da análise textual pelo então cenário de desenvolvimento da área de PLN. A estrutura gerada entre as diversas porções textuais de diversos documentos reflete a estrutura entendida pelos usuários do sistema. Esse processo pode ser útil na estruturação de assuntos complexos e no desenvolvimento de conteúdo colaborativo, mas não satisfaz a necessidade de um sistema que trate automaticamente grandes volumes de informações presentes em diversos documentos, principalmente dada a velocidade com que essas informações são produzidas atualmente

### 2.1.2. Modelo de Allan

Allan (1996) apresenta uma metodologia para identificação automática de ligações entre documentos de acordo com seu conteúdo. Essas ligações são agrupadas em três tipos: *Pattern-matching, Manual* e *Automatic*, levando em consideração a possibilidade de identificação automática das ligações. As ligações do grupo *Pattern-matching* são identificadas por simples ou elaboradas técnicas de casamento de padrões. Por exemplo, as palavras de um texto podem ser buscadas em um dicionário, a fim de estabelecer uma relação *Definition*. No grupo *Manual*, são identificadas ligações que não são possíveis identificar sem a intervenção humana, por necessitarem de um nível de interpretação textual ainda não atingida pelo Processamento de Linguagem Natural da época. Atualmente, com a criação de ferramentas e recursos mais linguisticamente informados (como analisadores sintáticos, semânticos e discursivos, WordNets, reconhecedores de entidades nomeadas, dentre outros) a identificação dessas ligações de forma automática torna-se viável.

As ligações do grupo *Automatic* são: *Tangencial, Revision, Summary, Expansion, Comparison, Constrast, Equivalence* e *Aggregate*. Essas ligações possuem direcionalidade, a exemplo das relações CST, e podem ser estabelecidas entre diversas porções textuais, como uma citação bibliográfica e o documento citado.

A metodologia foi inspirada em técnicas de visualização de estruturas de documentos, em que conceitos ficam nas bordas de um círculo e arestas ligam esses conceitos quando há certa similaridade entre os mesmos.

A metodologia de identificar as ligações é realizada em cinco passos: i) identificar as ligações candidatas em um conjunto de documentos; ii) identificação da ligação *Tangencial*, que ocorre pela desconexão (não relacionamento) com outros documentos; iii) ligações *Aggregate* são construídas das ligações que não sejam *Tangencial*; iv) técnicas de simplificação de grafos são utilizadas para reduzir a complexidade e número das ligações entre as subpartes dos documentos; v) definição do tipo da ligação: *Convolutional*, *Expansion*, *Relative size* ou *Absolute size*. Essa definição do tipo da ligação é feita utilizando um conjunto de regras que atua, principalmente, sobre a meta informação dos *links*.

### 2.1.3. Modelo de Radev e Mckeown

Radev e Mckeown (1998) propõem uma metodologia de sumarização multidocumento que utiliza um conjunto de relações, a saber: *change of perspective*, *contradiction*, *addition*, *refinement*, *agreement*, *superset/generalization*, *trend* e *no information*. Cada uma dessas relações é codificada em estruturas chamadas de operadores, que serão utilizados durante o processo de sumarização multidocumento, inclusive auxiliando no processo de fusão de *templates*.

Um exemplo de operador da relação de contradição (*contradiction*) é exibido na Figura 3. O operador é um conjunto de regras que atuam sobre *templates* que contêm informações extraídas das notícias e buscam identificar relações entre as partes dos textos. No caso específico da figura, caso o local em que ocorreu um incidente for o mesmo, mas o horário do acontecimento do incidente for diferente, para textos de fontes diferentes, ocorre uma contradição. Vale salientar que a metodologia apresentada foi aplicada a notícias que tratam sobre terrorismo.

```
((#TEMPLATES == 2) &&
 (T[1].INCIDENT.LOCATION == T[2].INCIDENT.LOCATION) &&
 (T[1].INCIDENT.TIME < T[2].INCIDENT.TIME) && ...
 (T[1].SECSOURCE.SOURCE != T[2].SECSOURCE.SOURCE)) ==>
(apply (``contradiction'',``with-new-account'', T[1], T[2]))
```

**Figura 3 - Exemplo de operador de sumarização, Radev e Mckeown (1998), pág. 13**

Esses operadores são importantes para que o sumário gerado leve em consideração interesses do usuário, como similaridades nos documentos, contradições, evolução de fatos no tempo, etc.

Neste trabalho (Radev e Mckeown, 1998), surgiu a necessidade do entendimento dos fenômenos multidocumento, a fim de permitir que um sumário contenha o que é de interesse do usuário, o que só é possível com a identificação das relações existentes entre os diversos documentos.

Por fenômenos multidocumento, entenda-se os relacionamentos (contradição, elaboração, dentre outras relações) que ocorrem entre partes de documentos distintos, sobre um mesmo tópico.

Para atender aos interesses do usuário, necessita-se encontrar as relações que ocorrem entre as partes dos diversos textos a sumarizar. Para isso, foram desenvolvidas técnicas para identificar as relações entre as informações dos textos através dos operadores de sumarização, permitindo a combinação dos conteúdos textuais, a fim de gerar o sumário final.

### 2.1.4. Cross-document Structure Theory (CST)

Motivado pelo cenário apresentado e com base em trabalhos anteriores que investigaram o relacionamento entre textos que versam sobre um mesmo assunto (Allan., 1996; Radev e McKeown, 1998; Trigg, 1983; Trigg e Weiser, 1986), Radev (2000) propôs a CST. Fortemente baseada na RST (*Rhetorical Structure Theory*, (Mann e Thompson, 1987) para estruturação discursiva monodocumento), a CST logo se destacou e mostrou seu potencial para diversas pesquisas. Essa teoria surge com a motivação de ser a base para a sumarização multidocumento, permitindo que a preferência do leitor seja considerada durante a sumarização, assim como a ordenação cronológica dos fatos.

Originalmente, propôs-se um conjunto de 24 relações discursivas para relacionamento intertextual. Na Figura 4, listam-se todas as relações.

| | | |
|---|---|---|
| *Identity* | *Modality* | *Judgment* |
| *Equivalence* | *Attribution* | *Fulfillment* |
| *Translation* | *Summary* | *Description* |
| *Subsumption* | *Follow-up* | *Reader profile* |
| *Contradiction* | *Elaboration* | *Contrast* |
| *Historical background* | *Indirect speech* | *Parallel* |
| *Cross-reference* | *Refinement* | *Generalization* |
| *Citation* | *Agreement* | *Change of perspective* |

*Figura 4 – Conjunto original de relações da CST*

É interessante observar que os autores da teoria referem-se à estrutura multidocumento como sendo retórica ou discursiva. No entanto, para múltiplos documentos, a noção de discurso (sequência das frases do texto seguindo determinada ordem de acordo com o que o autor deseja comunicar) pode se perder, dado que os múltiplos textos são produzidos por diversos autores e com distintos objetivos e, para um único documento, tem-se a premissa de que o texto seja coerente. Assim, é mais conveniente referir-se à CST como uma teoria "semântico-discursiva", pois expressa o relacionamento entre os conteúdos presentes nos diversos textos e constrói uma estrutura desse 'discurso' multidocumento.

Reproduz-se, abaixo, um exemplo da relação *Equivalence* (correspondente à paráfrase) entre os segmentos (S1) e (S2) provenientes de textos diferentes (Radev, 2000, p. 6):

(S1) *Ford's program will be launched* In the *United States in April and globally within 12 months.*

(S2) *Ford plans to introduce the program first for its employees* In the *United States, then expand it for workers abroad.*

O exemplo a seguir reproduz uma relação *Contrast* (Radev, 2000, p. 7):

(S1) *Agriculture Minister Loyola de Palacio estimated the loss at dlrs 10 million.*

(S2) *Agriculture Minister Loyola de Palacio has estimated losses from ruined produce at 1.5 billion pesetas (dlrs 10 million), although farmers groups earlier claimed total damages of nearly eight times that amount.*

Nem todas as partes dos textos envolvidos no processo de análise CST são relacionadas. Apenas alguns trechos de cada texto são relacionados, pois, em geral, há partes nos textos que não se referem a um mesmo assunto ou apresentam informações muito diferentes entre si.

As relações da CST atribuem diferentes status, ou direcionalidade, aos segmentos que relacionam. As relações podem ser simétricas ou assimétricas. A relação *Equivalence* é um exemplo de relação simétrica, enquanto a relação *Historical-background* é assimétrica, pois um segmento fornece o cenário histórico para outro e, portanto, tem que ser entendido antes do outro.

Segundo a CST, qualquer unidade discursiva pode ser considerada na análise. Podem-se relacionar palavras, expressões multipalavras, sintagmas, orações, sentenças, parágrafos ou blocos de textos maiores. Apesar de orações e sentenças serem tradicionalmente consideradas as unidades discursivas por excelência, tarefas particulares podem exigir um relacionamento entre unidades menores. Por exemplo, para a fusão de informações, o relacionamento de sintagmas pode ser mais adequado do que orações ou sentenças.

A Figura 5 mostra um esquema genérico de relacionamento entre as sentenças de textos sobre um mesmo tópico. Nesta ilustração, pode-se identificar a ocorrência de relações com direção ou sem direcionalidade. Como na proposta de trabalho, os segmentos a serem considerados serão as sentenças dos textos.

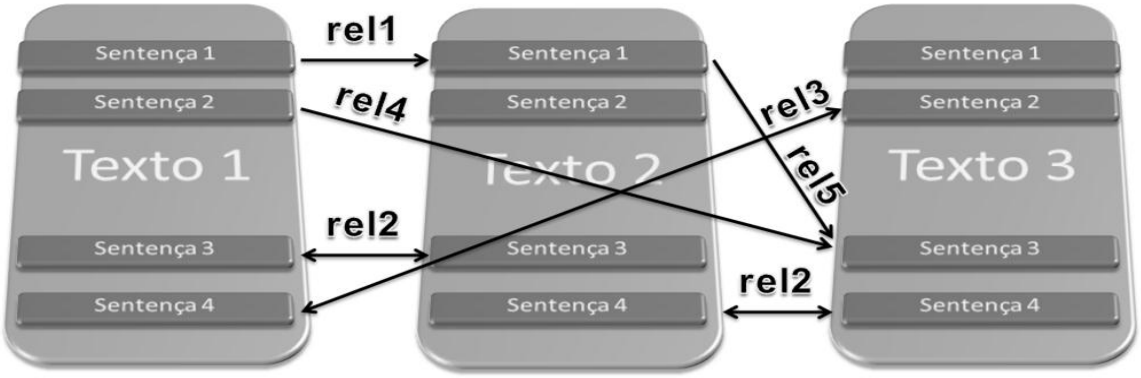


*Figura 5 - Esquema Genérico de Análise Multidocumento*

É possível ter, também, mais de uma relação por par de sentenças. Inclusive, pode-se ter, para um mesmo par, relações que referem-se ao conteúdo em si (por exemplo, *Overlap*) e

relações que referem-se à forma como o conteúdo é apresentado (por exemplo, *Indirect Speech*).

É importante dizer que existe ambiguidade na análise CST. Analisadores humanos diferentes podem identificar diferentes relações entre os mesmos segmentos ou, ainda, podem selecionar segmentos diferentes para relacionar. Essa subjetividade é inerente à maioria das tarefas que envolvem a semântica dos conteúdos textuais. Como será visto na descrição da criação de córpus anotado segundo a CST, a definição detalhada de todas as relações utilizadas permitiu uma concordância entre os anotadores em um nível aceitável para a automatização da análise CST.

O resultado da análise CST é um grafo, em que as porções textuais (segmentos) são os vértices e as arestas, rotuladas, representam as relações entre os segmentos.

#### 2.1.3.1 Usos da CST

Várias pesquisas têm utilizado CST. Radev (2000) mostra uma aplicação da CST para a fusão de informações e sumarização multidocumento.

Otterbacher et al. (2002) fazem um estudo em que examinam os problemas que afetam a coesão de sumários multidocumento e propõem uma arquitetura de sistemas de sumarização multidocumento numa abordagem de revisão do sumário por operadores que buscam tratar os problemas examinados. Nessa arquitetura, a CST é utilizada como informação útil para a revisão do sumário, corrigindo os possíveis problemas de coerência. O sistema proposto tem como entrada um conjunto de textos e como saída uma lista de sentenças escolhidas, que compõem o sumário final.

Jorge et al. (2011) apresentam um sistema de sumarização multidocumento que utiliza seleção de conteúdo valendo-se das informações obtidas pela análise CST dos documentos a serem sumarizados. As preferências dos usuários são consideradas buscando tratar os fenômenos de redundância, complementariedade e contradição. O sistema recebe um conjunto de textos que tratam de um mesmo assunto e que foram anotados segundo a CST. Dada a segmentação dos textos em sentenças, um ranking dessas sentenças é montado de acordo com o número de relações de cada uma das sentenças: quanto mais

relações uma sentença tiver, mais alto no ranking ficará. Após a construção do ranking, diversas estratégias de seleção são aplicadas a fim de obter as sentenças que comporão o sumário final. Essas estratégias são codificadas como operadores. Cada operador reflete os interesses do usuário na obtenção do resumo. São apresentados 4 operadores, a saber: Contexto, Contradição, Autoria e Evolução de Eventos. Um operador de Redundância é sempre aplicado a fim de retirar as informações repetidas.

Por exemplo, caso um leitor queira que as informações contextuais tenham preferência, o operador de Contexto é utilizado e atua sobre o ranking inicial de sentenças. As relações *Historical-background* e *Elaboration* são consideradas e as sentenças relacionadas por essas relações são melhores posicionadas no ranking. Os resultados obtidos indicam que o uso de informações CST enriquece e aperfeiçoa o sumário gerado.

Zahri e Fukumoto (2011) apresentam uma metodologia de sumarização multidocumento que faz uso de diversas técnicas, dentre essas, uma que se utiliza da identificação das relações CST entre as sentenças dos documentos. Na identificação das relações entre as sentenças, os autores utilizam uma abordagem baseada em aprendizado de máquina (utilizam a técnica *SVM*) para identificar as relações *Identity*, *Paraphrase*, *Subsumption*, *Overlap* e *Elaboration*, mas não relatam a avaliação dessa abordagem.

Cardoso et al. (2011b) propõem uma metodologia de sumarização multidocumento que usa diversos modelos semânticos-discursivos, com informações de cada texto (RST) e do relacionamento entre os textos (CST). Os autores propõem dois métodos. No primeiro, utilizam-se, para compor o sumário, as sentenças com mais relações CST (o que indica as sentenças mais relevantes dos textos) e realiza, com base nas informação da análise RST, a poda dos satélites (considerando a análise retórica intra-sentencial, em que numa relação pode-se ter um segmento mais importante, nuclear, e outro menos importante, satélite). No segundo método, primeiro utiliza-se a informação da análise RST para podar os satélites das sentenças. Feito isso, faz-se a análise CST dos textos e selecionam-se as sentenças com mais relações CST. Em ambos os métodos, a fusão de sentenças é realizada e os segmentos são ordenados para compor o sumário final.

### 2.1.4. Modelo de Afantenos et al.

Afantenos et al. (2004) propõem uma metodologia de definição e identificação de relações multidocumento, criticando o modo como foi definida a *Cross-document Struture Theory* (CST).

Os autores justificam que a *Rhetorical Structure Theory* (RST), proposta por Mann e Thompsom (1987), não pode ser simplesmente expandida de uma análise monodocumento para multidocumento. Essa crítica foi motivada pela baixa concordância obtida na aplicação da CST na anotação do córpus CST Bank (Radev et al. (2004), e dado que, das diversas relações propostas pela CST, apenas algumas foram utilizadas na anotação e nenhuma outra relação foi proposta durante a anotação, como é permitido pela teoria.

Os autores propõem, então, que as relações entre as partes de diversos documentos não são genéricas, como na CST, mas dependem do assunto tratado pelos diversos documentos em análise, pois nem todas as porções textuais dos documentos contêm relações entre si.

Desta forma, é necessária a definição de uma ontologia sobre o assunto tratado nos textos (que foram agrupados) a fim de identificar os tipos das entidades presentes, seus atributos e seus papeis nos eventos descritos. Por exemplo, em textos que descrevem partidas de futebol, as entidades são "time", "jogador", etc. Um evento pode ser "penalidade". Os papeis das entidades são, por exemplo, "time vencedor" e "time que sofreu falta". Essa ontologia é criada manualmente.

Feita a ontologia, deve-se definir um conjunto de mensagens que são aplicáveis ao assunto. Por exemplo, considerando as partidas de futebol, poderiam ser definidas mensagens que indiquem a performance do time, cancelamento de um gol, a vitória de um time, etc. O sistema de sumarização empregado pelos autores realiza geração textual baseada no conjunto de mensagens definido.

Tendo a ontologia e as mensagens, definem-se as relações, que são específicas ao tópico tratado e que poderiam ser, inclusive, o mesmo conjunto da CST. Para a definição do conjunto de relações, são analisadas as mensagens e seus possíveis valores. Nessa definição, podem-se considerar, inclusive, os objetivos na análise multidocumento como

interesse na evolução temporal do assunto. As relações podem explicitar essa evolução temporal de um evento ou podem simplesmente relacionar segmentos textuais, mostrando elaborações, contradições, etc. No exemplo de partidas de futebol, as relações utilizadas podem ser *Identity*, *Equivalence*, *Elaboration*, *Contradiction* e *Preciseness*, categorizadas como relações "sincrônicas", por tratarem do mesmo evento (mesma partida) relatado por fontes diferentes. Outras relações poderiam ser *Stability*, *Antithesis*, *Positive Graduation*, *Negative Graduation*, *Variation*, *Identity*, *Analogy*, que são categorizadas como relações "Diacrônicas", por tratarem de evolução no tempo de um evento, relatado por uma mesma fonte (por exemplo, diferentes partidas de futebol de um mesmo time).

Essa abordagem proposta para análise multidocumento é muito custosa e impraticável automaticamente, principalmente na definição da ontologia, que automaticamente é um grande desafio. A definição das mensagens depende não somente de conhecimentos de mundo, como da manipulação desses conhecimentos de forma a definir as mensagens sobre um assunto.

### 2.1.5. Modelos da área de RTE (*Recognizing Textual Entailment*)

*Recognizing Textual Entailment* (RTE) (Dagan et al., 2005) é uma tarefa genérica em que dados dois fragmentos textuais, visa-se reconhecer se o significado de um pode ser inferido a partir do significado do outro fragmento. A tarefa maior da RTE é classificar relações semânticas entre um Texto (T) e uma Hipótese (H) em *Entailment*, *Contradiction* ou *Unknown*. Na identificação dessas relações, muitas técnicas têm sido empregadas, baseando-se em medidas de sobreposição lexical entre *bag-of-words* (Jijkoun e de Rijke, 2005) ou alinhamento de grafos criado a partir de dependências sintáticas e semânticas (Marsi e Krahmer, 2005 e MacCartney et al., 2006). Esses trabalhos serão descritos na próxima subseção.

Murakami et al. (2010) apresentam um sistema que identifica um conjunto de relações semânticas entre fatos e opiniões em textos da *web* escritos em japonês. Esse conjunto de relações foi baseado na CST e na RTE. No entanto, como justificam os autores, essas duas abordagens não têm as relações suficientes para a tarefa proposta. Para essa tarefa, os autores definiram uma unidade de informação semelhante a uma sentença,

chamada de *statement*. Essa unidade engloba tanto fatos como opiniões nos textos. As relações tratadas são: *Agreement, Conflict, Confinement, Evidence* e as relações CST. Para identificar os pares de *statements* a serem relacionados, é utilizada uma técnica chamada alinhamento estrutural, em que primeiro se faz o alinhamento lexical dos pares utilizando o nível de similaridade superficial e a similaridade das estruturas de predicado-argumento.

Esses pares são classificados em uma das relações utilizando-se os seguintes atributos: distância (em arestas) no grafo de dependência, distância em *chunks*, atributos binários que indicam se cada *chunk* é predicado ou argumento, a etiqueta de *part-of-speech* da primeira e última palavra de um *chunk* e o ranking do alinhamento lexical para cada par de *chunk*. Um *chunk* é uma porção textual semelhante a uma oração e as dependências sintáticas entre os *chunks* são representadas como arestas em um grafo. A técnica de aprendizado automático utilizada é a SVM e a metodologia proposta obtém o valor 0.52 para precisão, cobertura e medida-F, na identificação das relações consideradas. O córpus utilizado foi o CST Bank (Radev et al., 2004).

Nas avaliações das metodologias e ferramentas apresentadas nessa dissertação, considere precisão como o número de exemplos corretamente tratados com relação ao número total de exemplos tratados automaticamente pela ferramenta. Já cobertura é a quantidade de exemplos tratados corretamente com relação ao número total de exemplos considerados como ideais, na avaliação em questão. A medida-F é uma média harmônica desses dois valores.

## 2.2. Técnicas de parsing multidocumento

Os trabalhos de Zhang et al. (2003) e Zhang e Radev (2005) consistem na única tentativa conhecida de automatizar o processo de análise CST para a língua inglesa. Os autores procedem à análise CST em duas etapas: inicialmente, criam um classificador para determinar se dois segmentos quaisquer (sentenças, no caso) provenientes de textos diferentes são relacionados por alguma relação CST, independentemente de qual seja; em seguida, usam outro classificador para determinar qual a relação CST existente entre os segmentos. Para o primeiro classificador, os atributos utilizados se baseiam em medidas de similaridade lexical, como a tradicional medida do cosseno (Salton e Lesk, 1968) e a

BLEU (Papineni et al., 2002). Para o segundo classificador, atributos de três níveis foram utilizados: atributos léxicos (por exemplo, número de palavras de cada segmento e número de palavras em comum entre eles), sintáticos (por exemplo, número de palavras pertencentes a algumas classes morfossintáticas de cada segmento e número de palavras com classes em comum entre os segmentos) e semânticos (por exemplo, similaridade semântica entre os principais conceitos de cada segmento, obtidos pela seleção dos substantivos e verbos mais importantes dos segmentos, utilizando-se a WordNet de Princeton[7]).

Utilizam-se, nesse método, tanto dados rotulados quanto não rotulados, por meio de uma técnica de *boosting*, especificamente AdaBoost (Freund e Scharpire, 1997). Verifica-se que o uso de dados não rotulados aumenta o desempenho dos classificadores. Os dados rotulados utilizados vieram do CST Bank (Radev et al. (2004), um córpus anotado segundo a CST, para o Inglês).

Os resultados do experimento realizado pelos autores são exibidos na Tabela 1. São listadas pelos autores apenas as relações que tiveram uma frequência maior que 20 nos dados de teste.

**Tabela 1 - Resultados obtidos por Zhang e Radev, 2005**

| Relação CST | Precisão | Cobertura | Medida-F |
|---|---|---|---|
| *No relation* | 0.8875 | 0.9605 | 0.9226 |
| *Equivalence* | 0.5000 | 0.3200 | 0.3902 |
| *Subsumption* | 0.1000 | 0.0417 | 0.0588 |
| *Follow-up* | 0.4727 | 0.2889 | 0.3586 |
| *Elaboration* | 0.3125 | 0.1282 | 0.1818 |
| *Description* | 0.3333 | 0.1071 | 0.1622 |
| *Overlap* | 0.5263 | 0.2941 | 0.3773 |

Algumas relações tiveram resultados bem baixos (abaixo de 0.2), como *Subsumption*, *Elaboration* e *Description*. Os autores justificam esses resultados pela esparsidade dos dados de treinamento iniciais para essas relações, considerando a grande quantidade de classes do problema.

[7] http://wordnet.princeton.edu/

Jijkoun e de Rijke (2005) propõem um método baseado no cálculo direto da similaridade lexical entre o texto e a hipótese. Para isso, baseiam-se em uma técnica de peso do termo mais frequente em combinação com duas outras medidas de similaridade: *Dekang Lin's dependency-based word similarity* (Lin, 1998) e uma medida baseada nas cadeias lexicais em WordNet. Na aplicação dessas duas últimas medidas, as palavras são lematizadas. Para realizar o peso das palavras, utilizam uma medida chamada *normalized inverse collection frequency*. Os autores obtiveram 0.55 de desempenho nos dados de teste (*PASCAL-2005 Recognizing Textual Entailment Challenge*).

Marsi e Krahmer (2005) abordam a classificação de relações semânticas entre pares de sentenças de um córpus em Holandês. Os autores utilizam cinco relações mutuamente exclusivas: *equals*, *generalizes*, *specifies*, *restates* e *intersects*. No trabalho, avalia-se também o desempenho humano (utilizando medida-F), que atinge 0.98 no alinhamento das sentenças e 0.95 na identificação das relações. A metodologia proposta atinge 0.85 no alinhamento e 0.80 na identificação das relações semânticas. O alinhamento é feito entre as árvores de dependência das sentenças em análise. Os autores, utilizam técnicas de aprendizado de máquina na tarefa de identificar as relações automaticamente, com os seguintes atributos: i) se as sentenças são idênticas, ii) atributos binários para cada uma das cinco relações indicando se elas acontecem em pelo menos um nó filho, iii) se pelo menos um dos nós filhos não está alinhado e iv) a relação semântico-lexical entre os nós conforme encontrado em uma WordNet.

MacCartney et al. (2006) representam o texto em um grafo de tipo de dependência, em que os nós representam as palavras e as arestas rotuladas com as relações gramaticais entre as palavras. O grafo de cada sentença em um par é alinhado de forma que cada palavra de um grafo é mapeada a apenas uma palavra do outro grafo ou a nenhuma palavra. Para identificar o relacionamento entre as sentenças, utiliza-se um classificador estatístico (para regressão). Os atributos utilizados são: i) atributos de modalidade, que consistem na resposta a aplicação de padrões que verificam a ocorrência de marcadores de modalidade, ii) atributos de factualidade, que verifica a qual classe de verbos de factualidade a sentença pertence, iii) atributos de quantificação, que captura as relações de quantificação entre as sentenças, iv) atributos de tempo, data e número, que reconhecem o casamento ou não desses itens nas sentenças e a v) atributos de alinhamento, que consiste

em três valores, a saber, o *score*, retornado durante a fase de alinhamento, *goodscore* e *badscore*, utilizados para verificar a qualidade do alinhamento. Os autores testaram a abordagem no *fisrt PASCAL-2005 Recognizing Textual Entailment Challenge* e obtiveram acurácia máxima de 0.65. Acurácia é um valor que resume o desempenho da metodologia ou sistema em avaliação, indicando a porcentagem geral de acerto.

Miyabe et al. (2008) propuseram uma metodologia para identificar as relações *Equivalence* e *Transition* entre pares de sentenças que foram agrupadas de acordo com suas similaridades. Foi construído um classificador binário para cada grupo de sentenças para identificar a relação *Equivalence*, que é estabelecida entre sentenças que contêm a mesma informação, porém não com as mesmas palavras. Essa abordagem para a relação *Equivalence* foi utilizada na proposta de identificar relações *Transition*, que é estabelecida entre duas sentenças que contêm a mesma informação diferindo apenas por valores numéricos. O classificador utilizado foi SVM (*Support Vector Machines*, Vapnik (1995)) e, nessa abordagem, a medida de similaridade entre sentenças foi a medida cosseno.

Para criar o classificador, os seguintes atributos foram utilizados: i) similaridade de cosseno entre as sentenças, ii) comprimento (em caracteres) normalizado das sentenças, iii) diferenças nas datas de publicação dos textos de onde vieram as sentenças, iv) posição das sentenças nos documentos, v) similaridades semânticas, vi) conjunções, vii) expressões ao final das sentenças, viii) entidades nomeadas nas sentenças, e ix) tipos de entidades nomeadas seguidas de algum marcador de caso (foram definidas 11 possíveis marcadores de caso).

Na identificação da relação *Equivalence*, foi utilizado um método de dois estágios, de acordo com o grau de similaridade entre as relações. Os pares de sentenças foram agrupados nos seguintes grupos: alta similaridade, média similaridade e baixa similaridade. Verifica-se a relação *Equivalence* em todos os grupos e, para isso, são utilizados dois classificadores: um para os pares de sentenças no grupo de alta similaridade e outro para os pares presentes nos grupos de média ou baixa similaridade. Essa abordagem é vantajosa por considerar as especificidades dos pares de sentença de acordo com o grau de similaridade apresentado.

Foram realizados experimentos na identificação da relação *Equivalence* e posteriormente, na identificação da relação *Transition*. Os resultados foram obtidos

utilizando uma avaliação cruzada de dez pastas, e as medidas precisão, cobertura e medida-F foram calculadas. Os resultados obtidos são sumarizados na Tabela 2, em que precisão indica a quantidade de relações corretamente identificadas, considerando todas as relações identificadas automaticamente, cobertura indica a quantidade de relações identificadas corretamente com relação ao total que a automatização deveria identificar e medida-F é uma combinação das medidas precisão e cobertura.

**Tabela 2 - Resultados obtidos, resumidos de Miyabe et al. 2005**

| Relação | Precisão | Cobertura | Medida-F |
|---|---|---|---|
| *Equivalence* | 0.9499 | 0.6265 | 0.7550 |
| *Transition* | 0.4306 | 0.4855 | 0.4564 |

Aleixo e Pardo (2008c) desenvolveram a primeira etapa da análise multidocumento para a língua portuguesa ao realizar a detecção de pares de sentenças a relacionar. Essa identificação é necessária devido ao alto número de pares de sentenças que são gerados ao se relacionar todas as sentenças de todos os documentos de um agrupamento de textos. Além disso, nem todas as sentenças têm alguma relação entre si. A identificação dos pares de segmentos candidatos a serem relacionados foi implementada na ferramenta CSTTool (Aleixo e Pardo, 2008b), apresentada posteriormente, nesta seção.

A etapa de detecção de pares de sentenças a relacionar consiste na aplicação da medida *word overlap* na escolha dos pares que apresentarem valor acima de determinado limiar. Um experimento foi realizado a fim de escolher essa medida: dois grupos de textos do córpus CSTNews, apresentado neste trabalho, foram selecionados aleatoriamente. Nesses dois grupos de texto, havia 2658 possíveis pares de sentenças a relacionar e, destes, 91 continham alguma relação CST. As medidas verificadas foram *word overlap* e medida cosseno. Variações das duas medidas foram realizadas utilizando lematização, remoção de *stopwords* e thesaurus. A medida que se mostrou mais adequada foi a *word overlap* que apresentou cobertura de 0.93 a 1 com valor de 0.1.

Ohki et al. (2011), na realização da RTE, fazem a identificação das três relações entre os pares de sentenças: *Entailment*, *Contradiction* e *Unknown* e *Confinement*. Os autores também reconhecem a ocorrência das relações *Entailment* e *Contradiction* entre

apenas partes das sentenças. Essa relação *Confinement* é reconhecida com medida-F de 61% para textos em Japonês. Nessa tarefa, são utilizados *templates* semânticos, que utilizam um conjunto de características do par de sentenças como entrada e as categorizam em uma das relações. As características utilizadas foram três: o tipo de restrição, o tipo de premissa e o tipo de consequência, pois foi assumido que cada sentença consiste de uma premissa e uma consequência e o tipo de restrição indica a relação lógica entre esses elementos (“Se” ou “Somente Se”).

## 2.3. Recursos e ferramentas

Radev et al. (2004) apresentam a construção de um córpus da língua inglesa anotado segundo a CST, o CST Bank. Segundo os autores, tal córpus deverá subsidiar investigações em relacionamentos multidocumento, estudos sobre a CST e, eventualmente, o desenvolvimento de sistemas automáticos de análise CST.

O CST Bank é composto de seis grupos de textos, extraídos de fontes diferentes, totalizando 41 textos, e uma média de mais de 28 sentenças por texto. Desses seis grupos, cinco foram agrupados manualmente e um deles automaticamente. Os grupos são distintos com relação ao assunto que tratam, tamanho e fontes da informação.

Um dos grupos foi utilizado para treinamento da equipe de anotação, composta por dois dos autores. Durante o treinamento, esquemas de marcação e um guia de anotação foi desenvolvido. Após o treinamento, uma equipe de oito anotadores recebeu o guia de anotação e anotou, os outros cinco grupos de textos. Os autores reportam uma concordância de 0.53 na medida Kappa (Carletta, 1996) para a concordância de haver alguma relação CST, independente de qual seja.

Como as relações são aplicadas entre sentenças dos documentos, a combinação de pares de sentenças de todos os documentos que compõem um grupo pode ser muito grande e inviável para uma anotação humana. Assim, utilizaram-se medidas de similaridade lexical para selecionar os pares de sentenças que serão julgados quanto a serem ou não relacionados por alguma relação CST, pois foi verificado que não é comum se ter relações CST entre pares de sentenças muito dissimilares lexicalmente entre si.

Os autores realizaram um experimento em que utilizaram medidas de similaridade e verificaram sua correlação com a ocorrência de relações CST. As medidas de similaridade utilizadas foram: medida cosseno, *word overlap*, *longest commom subsequence* e BLEU. A medida que se mostrou mais adequada foi a *word overlap*, cujo valor vai de 0 a 1, indicando que, quanto maior o valor, maior a similaridade entre as sentenças. O melhor valor encontrado para a medida foi de 0.12. Interessante notar que a exclusão de *stopwords* não modificou os resultados obtidos.

Alguns trabalhos foram iniciados recentemente para o Português. Foi produzido um córpus, chamado CSTNews, com 50 grupos de 2 a 4 textos jornalísticos cada, sobre um mesmo assunto, os quais foram anotados segundo a CST (Aleixo e Pardo, 2008a). Esse córpus foi reanotado durante o refinamento da teoria; esse procedimento será descrito no Capítulo 4. Desenvolveu-se um editor para anotação CST (usado para a anotação do córpus anterior), chamado CSTTool (Aleixo e Pardo, 2008b). Investigou-se a questão da identificação automática de segmentos (sentenças, no caso) relacionados e das melhores medidas automáticas para o Português (Aleixo e Pardo, 2008c), sendo que a detecção das relações entre os pares de segmentos identificados ficou como um tópico futuro de pesquisa, que é, agora, o foco deste trabalho.

A ferramenta CSTTool foi desenvolvida por Aleixo e Pardo (2008b) para a anotação do córpus CSTNews com vistas a posterior extração de conhecimentos para a realização automática da análise multidocumento.

A ferramenta realiza a segmentação textual em sentenças, que pode ser automática ou manual (Figura 6). Feita a detecção dos segmentos textuais dos textos que serão analisados, procede-se à detecção dos pares de segmentos candidatos a serem relacionados por alguma relação CST.

Na detecção dos pares de segmentos a relacionar, utiliza-se a medida *word overlap* (**Equação 1**). Essa medida é aplicada a toda combinação de pares de sentenças dos documentos que serão relacionados, criando uma lista com os pares cujo valor da medida esteja acima de um valor estipulado previamente (0.12, o mesmo do CST Bank).

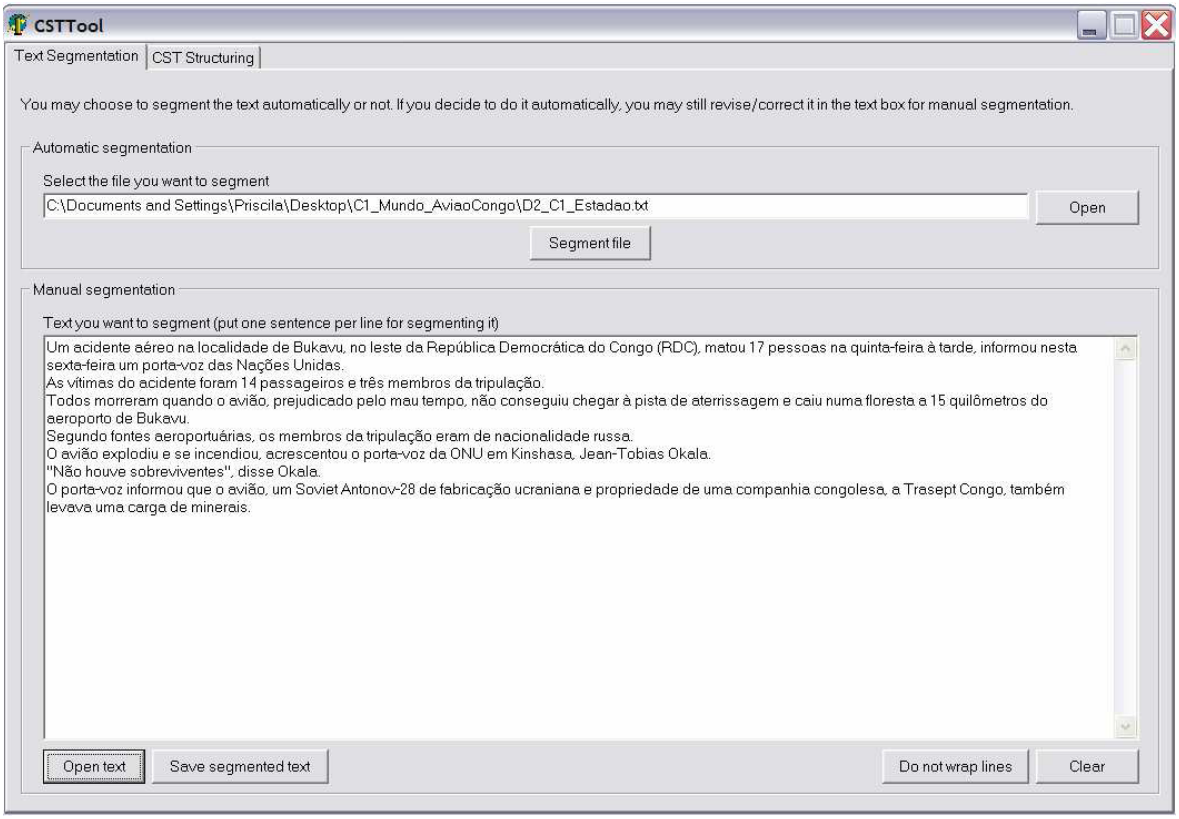


**Figura 6 - Segmentação textual na CSTTool**

***Equação 1 - Medida word overlap***

$$Word\ Overlap\ (S1,S2) = \#Palavras\ em\ Comum\ /\ (\#Palavras(S1) + \#Palavras(S2))$$

Sendo os textos segmentados e os pares candidatos selecionados, procede-se à identificação manual das relações CST entre dois documentos. Na Figura 7, veem-se duas caixas de texto, cada uma contendo um dos textos em análise. Os segmentos são rotulados por números sequenciais em cada texto. Esse número é utilizado na anotação da relação identificada entre dois segmentos textuais.

A cada relação identificada para um par de segmentos, deve-se escolher a direcionalidade da relação. Caso a relação não tenha direcionalidade, deixa-se o valor "*None*". Caso desejado, alguma nova relação pode ser adicionar à lista de relações originais presentes na ferramenta.

A anotação é feita, a cada vez, para um grupo de dois textos. Caso se tenha um grupo de mais de dois textos a serem relacionados (cenário mais realista), os textos são relacionados dois a dois, até que todos os possíveis pares de textos tenham sido relacionados. Isso não afeta o resultado, pois cada relação CST aplica-se sempre a um par de sentenças.

A ferramenta CSTTool foi o primeiro esforço empreendido no desenvolvimento de um analisador discursivo multidocumento para a língua portuguesa, segundo a teoria CST, pois faz a seleção automática de pares de sentenças a serem analisadas.

Verificou-se a complexidade de identificar as relações multidocumento, inclusive durante a anotação do córpus CSTNews. Mais detalhes sobre o córpus CSTNews, inclusive numéricos, são dados no Capítulo 3.

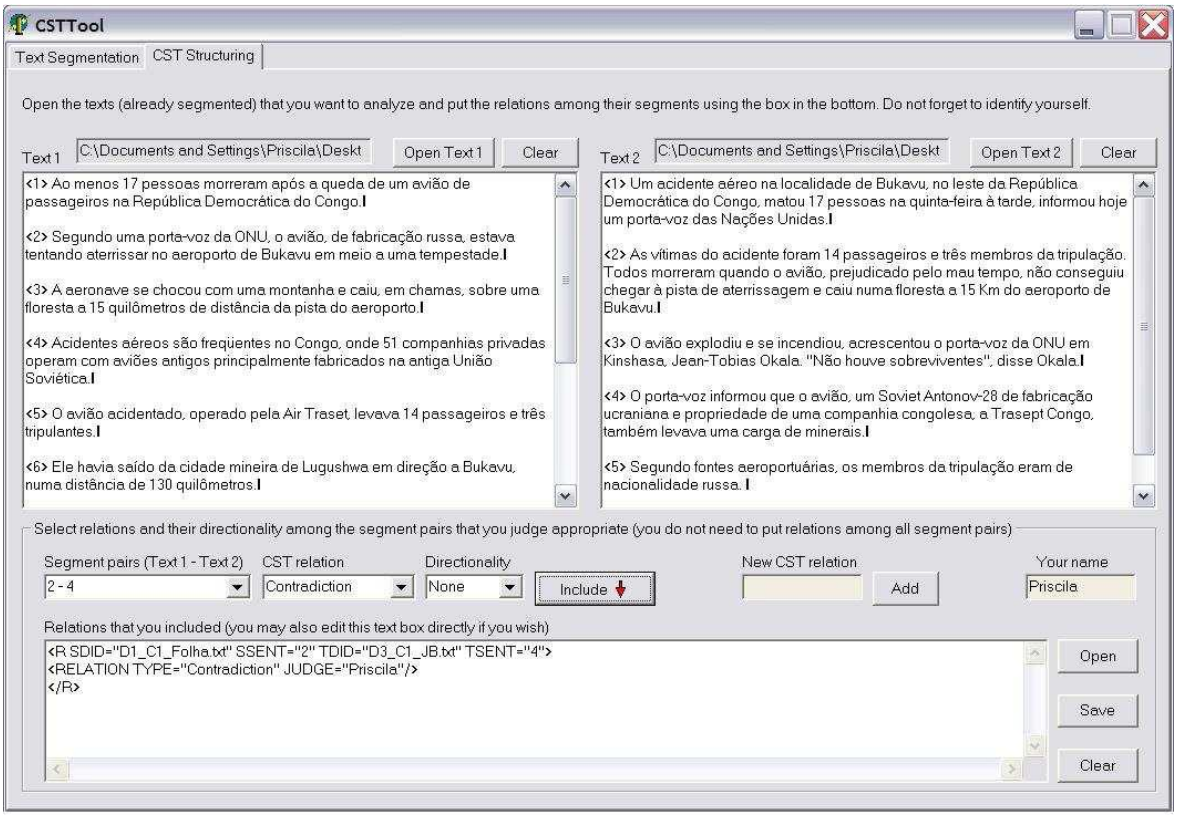


**Figura 7 - Ambiente para realização da análise multidocumento manualmente**

## 2.4. Considerações Finais

Tanto para a língua portuguesa, como para muitas outras línguas, há a ausência de um analisador discursivo multidocumento automático tornando necessária a anotação dos múltiplos documentos por algum anotador humano, quando desejada uma estruturação entre as partes desses diversos documentos.

Neste trabalho, optou-se pela CST, ao invés de outro modelo de análise e estruturação multidocumento, pois é uma teoria possível de ser aplicada automaticamente, não dependendo da definição de ontologias automaticamente, ou outras tarefas ainda com desempenho não suficiente para automatização, como a criação automática de ontologias.

O desenvolvimento do analisador CST será beneficiado pelos esforços já empregados nesse objetivo, inclusive pela existência de um córpus anotado segundo a teoria e de uma ferramenta de anotação multidocumento (CSTTool), que realiza a seleção de pares de segmentos a serem relacionados.

# 3. Métodos de Aprendizado de Máquina

Sempre foi interessante a possibilidade de que os computadores pudessem aprender tarefas automaticamente e melhorar seu desempenho com a experiência na execução dessas tarefas. Esse aprendizado se dá mediante a análise de exemplos feitos da tarefa em questão.

Por exemplo, para um aprendizado automático de como realizar uma análise sintática, diversas análises são apresentadas para a máquina que, utilizando algum algoritmo, gera um modelo que representa o conhecimento aprendido e que será utilizado na tarefa a ser desempenhada automaticamente. Segundo Mitchell (1997), um programa de computador aprende a partir de uma experiência E com respeito a alguma classe de tarefas T e medida de desempenho P, se seu desempenho na tarefa T, medido por P, melhora com a experiência E.

Neste trabalho, por exemplo, a tarefa T a ser aprendida e executada automaticamente é a identificação de relações CST. Como exemplos da tarefa (experiência E) são utilizados os pares de sentença relacionados segundo a teoria em questão e diversos modelos representando o conhecimento aprendido são gerados e avaliados (medida de desempenho P) para uso no *parser* multidocumento.

Para apresentar os exemplos da tarefa para algum algoritmo de aprendizado, os mesmos devem ser pré-processados e atributos, ou características, são extraídos. Esses atributos podem ser na forma numérica ou simbólica. Todo o processamento durante o aprendizado é feito baseado nos valores dos atributos. Além de extrair os atributos, para algumas tarefas, é possível identificar qual a classe desejada para cada exemplo, ou seja, qual o valor que se espera que o algoritmo encontre, ao processar os atributos dos exemplos. Assim, para um par de sentenças, extraem-se os valores dos atributos em forma numérica (esses atributos serão tratados posteriormente neste documento) e identifica-se a relação CST (classe). Essa classe é utilizada pelo algoritmo de aprendizado a fim de verificar o desempenho que se obtem na tarefa.

## 3.1. Paradigmas de treinamento

Para um aprendizado automático, pode-se recorrer a alguns paradigmas de treinamento, que variam, por exemplo, de acordo com a disponibilidade e quantidade dos exemplos da tarefa a ser aprendida. São três os principais paradigmas, a saber: supervisionado, semissupervisionado e não supervisionado.

No supervisionado, os exemplos da tarefa são apresentados para um algoritmo de aprendizado na forma de um conjunto de pares (*X*,*C*), em que *X* indica os atributos do exemplo e *C*, a classe ideal do exemplo. Esses exemplos são chamados de dados rotulados, por apresentarem a classe alvo. Nesse paradigma, o valor encontrado pelo algoritmo é comparado com a classe ideal e o modelo de conhecimento é ajustado a fim de melhorar o desempenho do modelo aprendido.

No semissupervisionado, utiliza-se tanto dados rotulados quanto não rotulados. Essa abordagem é utilizada quando o esforço em se obter dados rotulados é muito grande e custoso. Pode-se, assim, treinar modelos de aprendizado sobre esse conjunto inicial de dados rotulados e utilizar o modelo para rotular novos dados. Esses novos dados são, então, utilizados no treinamento de novos modelos, contando agora com mais dados de treinamento e um melhor desempenho na tarefa é alcançado.

Já no aprendizado não supervisionado, não é apresentada a classe dos exemplos, ou seja, os dados são não rotulados. Desta forma, o algoritmo de aprendizado deve ser capaz de encontrar alguma relação entre os exemplos e gerar grupos, que constituirão as classes da tarefa a ser aprendida. Novos exemplos apresentados serão, então, designados para alguma dessas classes.

Neste trabalho utilizaram-se algoritmos de aprendizado do paradigma supervisionado, pois o córpus utilizado é constituído de exemplos e suas classes. Além disso, se um algoritmo não supervisionado fosse utilizado, ele poderia gerar novas classes que não corresponderiam às classes CST.

Uma abordagem semissupervisionada pode ser futuramente explorada, utilizando os modelos aprendidos para classificar novos pares de sentenças, que, por sua vez, seriam utilizados no aprimoramento dos modelos já aprendidos.

## 3.2. Técnicas de aprendizado

As técnicas de aprendizado geram os modelos de representação do conhecimento aprendido. Esses modelos podem ser classificados em estatísticos, simbólicos ou probabilísticos, dentre outros. Para cada classe de modelo, algumas técnicas, ou algoritmos, são responsáveis por gerar o modelo aprendido.

As técnicas simbólicas geram modelos que são "humanamente compreensíveis", geralmente expressos na forma de regras ou árvores de decisão. Essa característica é desejável para muitas tarefas, o que inviabiliza o uso de técnicas estatísticas e probabilísticas, cujo conhecimento codificado é ininteligível humanamente. Algumas técnicas da classe simbólica geram modelos que dependem explicitamente dos valores dos atributos, por exemplo, considere a árvore criada pelo algoritmo J48 na Figura 8 (uma implementação do algoritmo C4.5 de Quinlan (1993)), que é facilmente convertida em um conjunto de regras lógicas para a identificação de uma relação entre um par de sentenças. Cada nó da árvore é um atributo do problema e as arestas são caminhos para se chegar a um nó folha (uma solução para o problema de classificação). Nas arestas, há limites de valores dos atributos, que definem qual o caminho a tomar na decisão por uma classe.

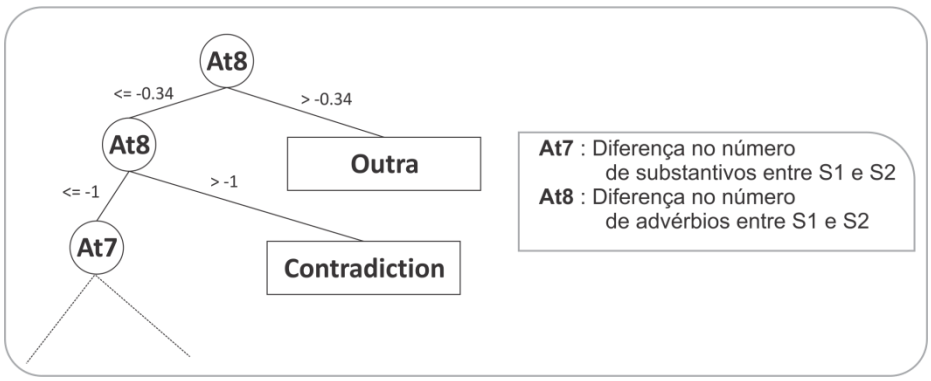


**Figura 8 - Esboço de parte da árvore de decisão gerada pelo algoritmo**

Como uma das formas de se chegar à classe *Contradiction* (Figura 8), temos que a diferença no número de advérbios entre S1 e S2 deve ser menor ou igual a -0.34 (número

normalizado) e maior que -1 (número normalizado). Essa facilidade de interpretar uma solução é uma das maiores características de técnicas simbólicas.

Outro exemplo de classificador da classe simbólica, gerado segundo o algoritmo OneR (Holte, 1993), é exibido na Tabela 3, em que o atributo 3 foi escolhido para, de acordo com seu valor, identificar uma relação no classificador multiclasse para as relações de "conteúdo" (com uma acurácia de 34.5%). Por exemplo, se o valor do atributo 3 for 0.030, pelas regras "< 0.033 ► *Follow-up*" e "< 0.019 ► *Elaboration*", temos que a relação escolhida será *Follow-up*.

**Tabela 3 - Exemplo de regras gerada segundo o algoritmo OneR**

| **Porcentagem de palavras em comum em S2 (Atributo 3):** | |
|---|---|
| < 0.004 ► *Elaboration* | < 0.210 ► *Subsumption* |
| < 0.016 ► *Follow-up* | < 0.224 ► *Overlap* |
| < 0.019 ► *Elaboration* | < 0.231 ► *Subsumption* |
| < 0.033 ► *Follow-up* | < 0.266 ► *Overlap* |
| < 0.040 ► *Elaboration* | < 0.353 ► *Subsumption* |
| < 0.044 ► *Overlap* | < 0.391 ► *Overlap* |
| < 0.078 ► *Elaboration* | >= 0.391 ► *Subsumption* |
| < 0.127 ► *Overlap* | |
| < 0.134 ► *Elaboration* | |
| < 0.203 ► *Overlap* | |

Um exemplo de algoritmo da classe probabilística é o NaiveBayes (John e Langley, 1995). É assim chamado (*naive* = ingênuo) por considerar condicionalmente independentes os atributos dos exemplos utilizados no aprendizado. Essa técnica utiliza o teorema de Bayes das probabilidades condicionais. Essa técnica estima a probabilidade de cada possível valor da classe da tarefa em questão, utilizando os valores dos atributos. A classe com maior probabilidade será a escolhida. Apesar de sua simplicidade, essa técnica tem apresentado bom desempenho em inúmeras tarefas, com a vantagem de, dado que assume a independência condicional dos atributos, precisar de um número menor de exemplos para realizar um bom aprendizado.

Como exemplo de técnica estatística de aprendizado, tem-se o *Support Vector Machines* (SVM, Vapnik, 1995) ou Máquinas de Vetores de Suporte. Nessa técnica, os exemplos são representados como pontos em um espaço multidimensional. Assim, esse algoritmo gera um hiperplano nesse espaço que separa os pontos de acordo com as classes

desejadas. Diferentemente das técnicas simbólicas, como o algoritmo J48 (que gera uma árvore de decisão), a hipótese aprendida pelo SVM não permite interpretação.

## 3.3. A tarefa de classificação

Em aprendizado automático, objetiva-se classificar novas entradas em classes. Quando essas classes são discretas, tem-se a tarefa de classificação. Outra tarefa muito comum em aprendizado de máquina é o agrupamento, em que se têm, principalmente, dados não rotulados e objetiva-se agrupar esses dados segundo alguma medida de similaridade ou dissimilaridade. Para este trabalho, temos a criação de classificadores, dado que as classes já estão definidas a priori: relações CST, e estas são finitas.

Um classificador tem como objetivo, portanto, receber dados não rotulados e associá-los a uma das classes.

### 3.3.1 Tipos de Classificadores

Os classificadores podem se diferenciar de acordo com a quantidade de classes consideradas e o arranjo das classes. Algumas tarefas podem ser definidas como uma classificação binária, em que cada novo exemplo é classificado entre duas classes.

Quando há mais de duas classes na tarefa, diz-se que o classificador é multiclasse. Muitos algoritmos são originalmente projetados para apenas duas classes, como o SVM, e sua utilização na criação de classificadores multiclasse demanda adaptação. Na tarefa de identificar relações CST, criaram-se classificadores multiclasse, para, em apenas um passo, atribuir a um par de sentenças uma relação CST.

Se as classes estão organizadas em uma forma hierárquica, como as relações CST, que estão em uma hierarquia, algoritmos de aprendizado podem utilizar essa informação a fim de obter melhor desempenha na classificação. Esses classificadores hierárquicos podem seguir as abordagens *top-down* e *big-bang*. Para uma comparação entre as abordagens, ver Freitas e Carvalho (2007). Na abordagem *top-down*, a cada nível da tipologia, um classificador é utilizado até que se chegue a uma folha da tipologia, que conterá a classe a ser atribuída ao exemplo em questão. Outra possível abordagem à

classificação *top-down* é a identificação da relação em apenas um passo, levando em consideração a hierarquia das relações como um todo. Essa abordagem é conhecida como *big-bang* e consiste de uma alteração do algoritmo C4.5 (Quinlan, 1993). Em Clare (2003), essa alteração é descrita.

A sobreposição das classes é algo que não é tratado pelos algoritmos convencionais de aprendizado listados acima. Nesse caso, os algoritmos têm de ser adaptados ou novas técnicas devem ser implementadas a fim de realizar a classificação multirótulo (Cherman e Monard, 2009), em que, para um mesmo exemplo, atribui-se mais de um rótulo.

## 3.4. Questões relacionadas ao AM

As duas principais questões relacionadas a um aprendizado automático, em geral, são o desbalanceamento dos dados e a sobreposição das classes da tarefa, questões que podem diminuir o desempenho dos classificadores. Como dito acima, a sobreposição pode ser tratada com a criação de classificadores multirótulo. Para o desbalanceamento, há diversas técnicas que visam balancear os dados de treinamento (Batista et al., 2004).

O balanceamento das classes pode ser realizado segundo duas principais abordagens ou uma combinação destas. A primeira é fazer a replicação dos exemplos das classes menos frequentes. Assim, os exemplos das classes menos frequentes teriam de ser replicados até atingir número próximo ao dos exemplos da classe mais frequente. Dependendo do grau de desbalanceamento, essa replicação pode levar a uma acurácia de 100% nos testes. Esse resultado claramente seria provocado pelo fenômeno chamado *overfitting* (pois diversos exemplos iguais seriam utilizados no treinamento e teste), algo indesejado na criação de classificadores.

Outra forma de balancear os dados é através da exclusão de exemplos das classes mais frequentes. Essa abordagem se mostra inviável quando algumas classes têm poucos exemplos, pois retirar exemplos das classes mais numerosas resultaria em poucos exemplos para a realização do aprendizado automático, que tem maior aprendizado quanto maior o número de exemplos (sendo esses exemplos representativos do problema em questão).

Uma junção das duas abordagens anteriores seria retirar exemplos das classes mais frequentes e replicar exemplos das menos frequentes. Essa estratégia não foi testada, pois o tamanho do córpus não é grande o suficiente, o que provocaria um decréscimo no desempenho dos classificadores gerados.

Outra abordagem, semelhante à replicação, tem como exemplo a técnica de balanceamento SMOTE (*Synthetic Minority Over-sampling Technique*, Chalwa, 2002). Nessa técnica, inserem-se exemplos realizando interpolação dos exemplos que estejam próximos uns dos outros. Pode-se especificar uma semente para a randomização dos exemplos a serem criados no balanceamento. Diferentemente de fazer apenas uma replicação dos exemplos já presentes no conjunto de dados, o que poderia provocar um ajuste do modelo de aprendizado aos exemplos utilizados.

A sobreposição de classes também é uma característica indesejável na criação de classificadores, que são treinados para escolher entre apenas uma das classes do problema. Soluções como classificadores multirótulo são conhecidas e uma foi testada, mas com péssimo desempenho. A estratégia utilizada nesse teste foi criar novas classes baseadas na sobreposição das classes. Assim, um exemplo (par de sentenças) que seja da relação *Overlap* e da relação *Attribution*, será alocado para uma nova classe então chamada "*Overlap-Attribution*". Isso aumentou o desbalanceamento das classes no córpus, pois essas combinações geraram classes com pouca frequência.

## 3.5. Considerações Finais

Este capítulo apresentou os conceitos básicos sobre aprendizado de máquina, úteis no entendimento do Capítulo 5 deste documento. Nos experimentos que serão apresentados, utilizou-se o paradigma de treinamento supervisionado e técnicas estatísticas, simbólicas e probabilísticas foram aplicadas aos exemplos do córpus utilizado na tarefa.

Como resultado, diversos classificadores foram gerados e avaliados a fim de encontrar a melhor configuração para compor o *parser* multidocumento. Geraram-se classificadores multiclasse, multirótulo, binários e hierárquicos, para todas as relações CST e para apenas um grupo, a ser discutido posteriormente.

# 4. O Córpus CSTNews

## 4.1. Criação

Construído a partir de textos jornalísticos e anotado segundo a teoria CST, o córpus possui 50 grupos de textos e cada grupo trata de um assunto diferente. Esses documentos foram coletados manualmente de jornais online entre Agosto e Setembro de 2007. Os jornais online utilizados foram: Folha de São Paulo, Estadão, O Globo, Jornal do Brasil e Gazeta do Povo. A escolha desses textos reflete bem um dos objetivos da CST, que é trabalhar com textos, geralmente de fontes diversas, que tratem de um mesmo assunto.

Devido à baixa concordância do córpus (Kappa igual a 0,258), este foi reanotado no âmbito deste trabalho e algumas de suas características iniciais foram alteradas, como pode ser entendido em Cardoso et al. (2011a). Neste Capítulo, consideramos a nova versão do córpus CSTNews e apenas a análise CST dos grupos de textos, dado que o córpus tem sido acrescido de outras anotações, como a análise monodocumento RST dos textos.

A existência desse córpus possibilita o aprendizado automático para a identificação das relações entre as sentenças de textos diversos. Essa é, inclusive, umas das motivações para a criação do córpus. Outra razão é o entendimento dos fenômenos multidocumento, o que possibilita a criação de regras para a estruturação CST, uma outra abordagem da pesquisa em questão.

## 4.2. Anotação

A anotação foi realizada por quatro anotadores, que passaram por um treinamento com diversos grupos de textos por três mêses. Ao final de cada treinamento (anotação) de um conjunto de textos, as relações identificadas por cada anotador eram discutidas a fim de possibilitar maior entendimento e concordância entre os anotadores. Esse treinamento possibilitou, também, o refinamento das descrições das relações. O rol final das relações e suas descrições são apresentados no Anexo A.

Cada grupo de textos contem, no máximo, três textos e como cada anotação foi feita entre um par de textos, os anotadores dividiram-se em três (um dos grupos continha

sempre 2 anotadores) a fim de anotar os três possíveis pares (combinações) de textos formados pelos três textos, conforme ilustrado na Figura 9. O processo de anotação teve a duração de aproximadamente dois meses, sendo diariamente anotado um grupo no período de uma hora.

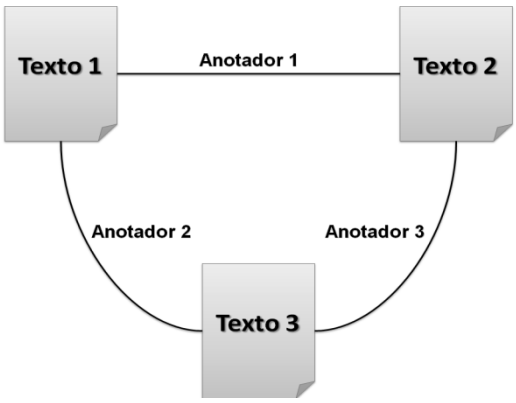


**Figura 9 - Esquema de combinação de textos na anotação do córpus**

No processo de anotação, foi utilizada a ferramenta CSTTool (Aleixo e Pardo, 2008b), que possibilita a anotação CST de forma manual. A ferramenta realiza automaticamente alguns passos prévios à identificação das relações, como a segmentação sentencial e a seleção, por meio da medida *word overlap*, de pares de sentenças (S1 e S2).

## 4.3. As relações e sua tipologia

Percebeu-se na anotação do córpus que as relações podem ser organizadas em uma tipologia que leva em consideração algumas características comuns (Maziero et al., 2010). A tipologia é ilustrada na Figura 10, em que as relações estão no nível mais inferior da hierarquia.

Algumas relações se ocupam principalmente do "conteúdo" das sentenças e, nessa ocupação, analisa-se a "contradição" entre os conteúdos, a "redundância", que pode ser "total" ou "parcial", ou o "complemento" entre os conteúdos, levando em consideração o aspecto "temporal" ou "atemporal" dessa característica.

Outras relações se ocupam da "forma" como as sentenças trazem as informações, considerando a "fonte/autoria" e o "estilo". As relações desse grupo podem ocorrer

conjuntamente com alguma outra relação do grupo de “conteúdo”, pois o par de sentenças em análise sempre apresentará algum conteúdo similar.

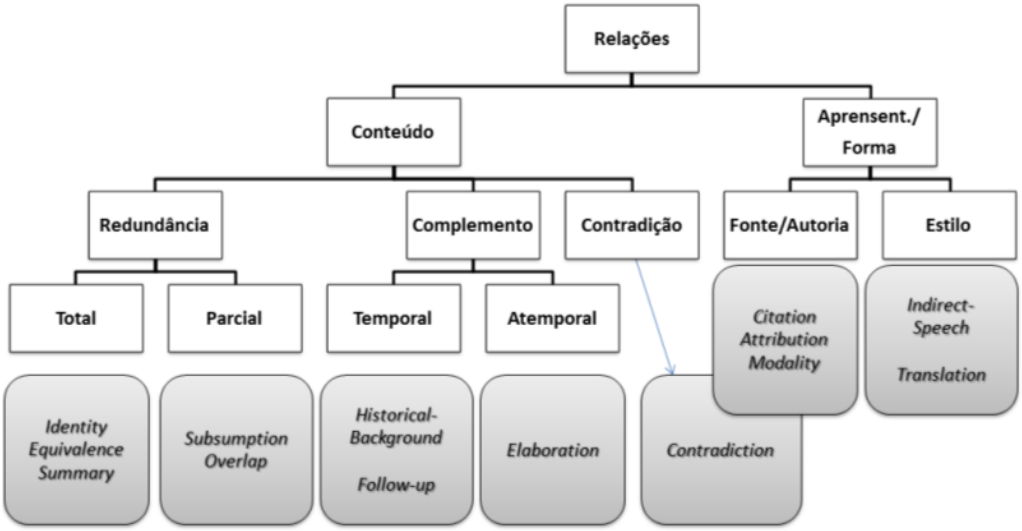


**Figura 10 - Tipologia das relações CST**

Por exemplo, uma sentença S1, de determinado documento, pode conter toda a informação de outra sentença, S2, de outro documento, mas apresenta um autor para essas informações; assim, dizemos que a sentença S1 é relacionada à sentença S2 pela relação *Subsumption*. A relação *Subsumption* indica redundância parcial de informações entre as sentenças relacionadas. No entanto, esse mesmo par de sentenças, S1 e S2, pode conter a relação *Attribution*, que diz ter uma sentença atribuição de autoria da informação contida nas sentenças relacionadas. A relação *Attribution* pertence ao grupo de relações de “forma” e a relação *Subsumption* ao grupo “conteúdo”.

Alguns experimentos foram conduzidos considerando esse agrupamento das relações, como um aprendizado hierárquico. Inclusive, a ocorrência de mais de uma relação entre um mesmo par de sentenças constitui-se em uma importante questão durante o aprendizado automático.

## 4.4. Características numéricas

Na Tabela 4, apresentam-se algumas estatísticas sobre o córpus, dentre elas o número de textos por grupo, quantidade de sentenças do grupo e quantidade de palavras. No total, são 140 documentos, 2088 sentenças e 47240 palavras. A Figura 11 apresenta a quantidade de textos por seção dos jornais utilizados.

Durante a anotação do córpus, os anotadores calcularam periodicamente a concordância entre eles para um grupo de textos. Os grupos que tiveram dois textos possibilitaram a anotação de apenas uma combinação de textos. Esse par foi anotado pelos três grupos de anotadores, possibilitando o cálculo da concordância na tarefa. Após o cálculo da concordância, outra anotação para o par de textos foi gerada de acordo com o consenso de todos os anotadores.

A Tabela 5 mostra a média das seguintes medidas Kappa: i) concordância na identificação de relações (não considerando a direcionalidade), ii) concordância da sua direcionalidade (as opções foram da primeira para a segunda sentença, da segunda para a primeira sentença e nenhuma), iii) concordância para as categorias das relações da tipologia, principalmente as categorias de redundância, complemento, contradição, autoria e estilo. Como pode-se observar, o valor da medida Kappa para esse córpus reanotado foi significantemente melhor que a versão original do córpus CSTNews (de fato, 96% acima). Como esperado, quando as relações são agrupadas em suas categorias, os resultados são ainda melhores. A boa concordância atingida indica que a tarefa em foco é bem definida e, portanto, passível de sistematização e posterior implementação.

A medida Kappa é muito útil para tarefas de anotação realizada por diversos anotadores, indicando se o material anotado é confiável. Além disso, ela possibilita avaliar um sistema automático em comparação com a tarefa humana. A medida varia dos valores -1 a 1. O valor 1 indica concordância máxima e o valor 0 indica concordância aleatória. Segundo Landis e Koch (1977), valores entre 0.41 a 0.60 indicam uma concordância moderada e valores acima de 0.60 que a anotação é substancial. A reanotação do córpus mostrou-se eficiente, pois o valor Kappa para as relações passou de 0,258 para 0,509.

**Tabela 4 - Estatísticas do córpus CSTNews**

| Grupo | Assunto | N° de textos | N° sentenças | N° palavras |
|---|---|---|---|---|
| C1 | Mundo | 3 | 24 | 432 |
| C2 | Política | 3 | 51 | 996 |
| C3 | Cotidiano | 3 | 50 | 1243 |
| C4 | Cotidiano | 3 | 39 | 832 |
| C5 | Cotidiano | 2 | 23 | 572 |
| C6 | Cotidiano | 3 | 36 | 925 |
| C7 | Ciência | 2 | 23 | 585 |
| C8 | Esportes | 3 | 25 | 593 |
| C9 | Política | 3 | 36 | 965 |
| C10 | Mundo | 3 | 39 | 962 |
| C11 | Cotidiano | 3 | 56 | 987 |
| C12 | Mundo | 3 | 37 | 960 |
| C13 | Mundo | 3 | 37 | 962 |
| C14 | Mundo | 3 | 25 | 739 |
| C15 | Mundo | 3 | 26 | 565 |
| C16 | Política | 3 | 47 | 1031 |
| C17 | Política | 2 | 41 | 963 |
| C18 | Mundo | 3 | 70 | 1301 |
| C19 | Esportes | 2 | 13 | 298 |
| C20 | Política | 3 | 42 | 949 |
| C21 | Política | 3 | 41 | 870 |
| C22 | Cotidiano | 3 | 50 | 964 |
| C23 | Mundo | 2 | 25 | 572 |
| C24 | Esportes | 3 | 24 | 541 |
| C25 | Esportes | 3 | 88 | 1558 |
| C26 | Mundo | 3 | 58 | 1406 |
| C27 | Esportes | 3 | 89 | 1542 |
| C28 | Esportes | 3 | 35 | 717 |
| C29 | Mundo | 3 | 48 | 1167 |
| C30 | Dinheiro | 3 | 46 | 1131 |
| C31 | Esportes | 2 | 10 | 217 |
| C32 | Mundo | 3 | 66 | 1328 |
| C33 | Cotidiano | 3 | 68 | 1638 |
| C34 | Cotidiano | 3 | 59 | 1139 |
| C35 | Mundo | 3 | 36 | 876 |
| C36 | Cotidiano | 3 | 74 | 1357 |
| C37 | Cotidiano | 2 | 26 | 475 |
| C38 | Esportes | 3 | 26 | 535 |
| C39 | Cotidiano | 3 | 35 | 914 |
| C40 | Política | 3 | 28 | 745 |
| C41 | Esportes | 3 | 45 | 958 |
| C42 | Política | 2 | 39 | 1061 |
| C43 | Política | 3 | 49 | 1267 |
| C44 | Política | 2 | 26 | 719 |
| C45 | Cotidiano | 3 | 47 | 1223 |
| C46 | Mundo | 3 | 38 | 740 |
| C47 | Mundo | 3 | 43 | 1373 |
| C48 | Esportes | 2 | 43 | 800 |
| C49 | Cotidiano | 3 | 23 | 1001 |
| C50 | Política | 3 | 63 | 1546 |
| Total de Documentos | | 140 | | |
| Total de Sentenças | | | 2088 | |
| Total de Palavras | | | | 47240 |

**Tabela 5 - Concordância Kapa**

| | Média do valor Kappa |
|---|---|
| Kappa de relações | 0.5094 |
| Kappa de direcionalidade | 0.4459 |
| Kappa de relações agrupadas | 0.6141 |

A Kappa da direcionalidade teve um valor baixo possivelmente devido à direção ser algo arbitrário, diferentemente da RST. A direcionalidade das relações foi definida por Radev (2000) sem critérios muito claros.

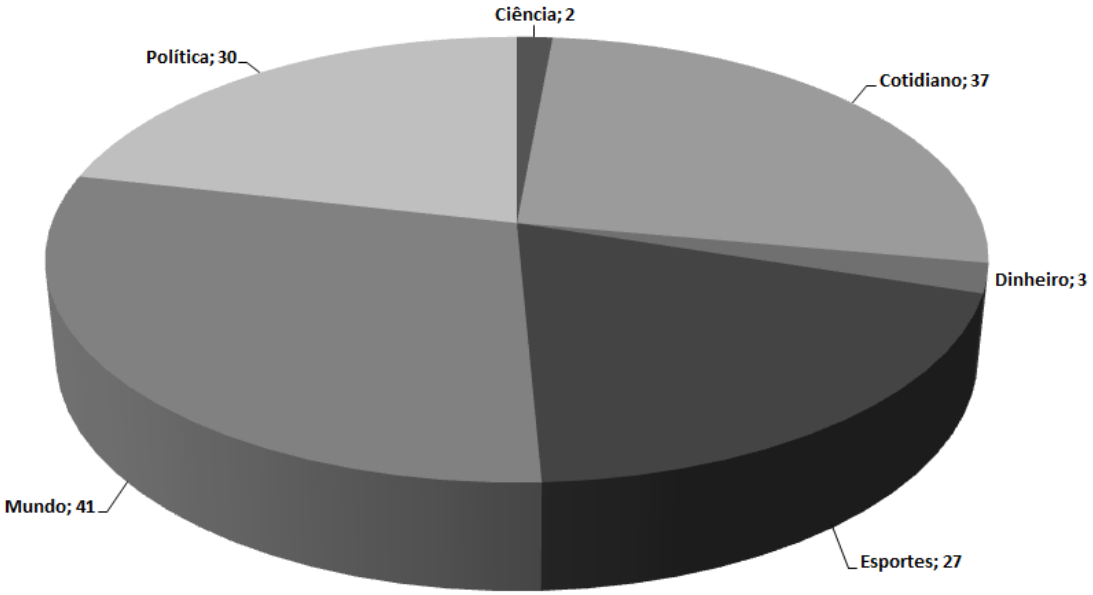


**Figura 11 - Quantidade de textos por seção**

Computou-se também a quantidade de vezes em que houve concordância entre os juízes. Foi calculada a concordância total, a concordância parcial (quando a maioria dos juízes indicou a mesma relação) e a concordância nula (quando cada juiz indicou uma relação diferente dentre a dos demais). Essa medida permitiu um melhor entendimento dos resultados. O percentual médio alcançado nesses resultados é exibido nas tabelas 6, 7 e 8. Como se pode notar, o percentual de concordância parcial ou completa é muito bom, verificado para mais de 80% das relações (contra 58% para a língua inglesa, como contam Zhang et al., 2002) e suas direcionalidades. Os resultados são ainda melhores quando consideradas as categorias das relações.

**Tabela 6. Concordância das relações**

| | Média de Concordância (%) |
|---|---|
| Concordância Total de relações | 54 |
| Concordância parcial de relações | 28 |
| Concordância nula de relações | 18 |

**Tabela 7. Concordância da direcionalidade**

| | Média de Concordância (%) |
|---|---|
| Concordância total de direcionalidade | 59 |
| Concordância parcial de direcionalidade | 27 |
| Concordância nula de direcionalidade | 14 |

**Tabela 8. Concordância de relações agrupadas**

| | Média de Concordância (%) |
|---|---|
| Concordância total de relações agrupadas | 70 |
| Concordância parcial de relações | 21 |
| Concordância nula de relações | 9 |

Apesar de se ter melhorado na concordância, ainda se observaram valores consideráveis para as concordâncias nulas, devidas principalmente à subjetividade da análise. No caso das relações agrupadas, foi observado 91% de concordância total ou parcial, o que mostra que a nova classificação feita ajuda a ter uma ideia mais clara e uniforme das relações.

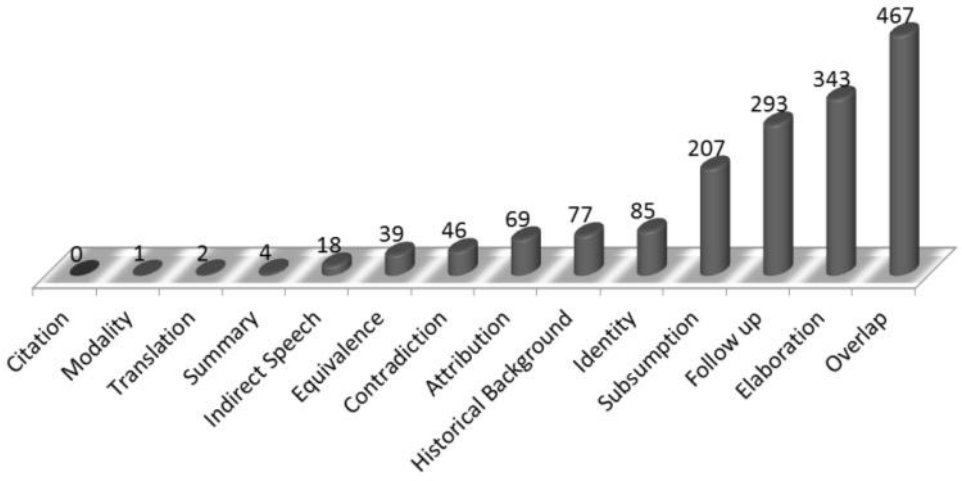


Figura 12 - Frequência das relações no córpus

A Figura 12 apresentada a frequência de cada relação CST no córpus anotado. A relação *Citation* não aparece no córpus, dada a raridade de um jornal atribuir uma notícia relatada a outro jornal. Relações como *Modality*, *Translation* e *Summary* têm baixa frequência, dadas suas características pouco encontradas em textos jornalísticos. Outras relações, principalmente as relações que tratam do conteúdo das sentenças, aparecem com alta frequência, como *Overlap*, *Elaboration*, *Follow-up* e *Subsumption*, como será discutido posteriormente. Para a identificação automática das relações CST pelo aprendizado automático, as relações com baixa frequência não foram satisfatoriamente tratadas, mostrando que outras abordagens para a identificação automática devem ser consideradas.

## 4.5. Disponibilidade do córpus

Para proporcionar um ambiente de navegação e visualização das relações e sentenças presentes no córpus, foi criado um sistema *web*. O objetivo era permitir, em um único lugar, navegar no córpus de modo a compreender melhor a CST, suas relações e o relacionamento entre segmentos. O sistema fornece ainda uma pesquisa completa por coleções de texto e também a possibilidade de *download* do córpus em um arquivo compactado.

Dentre as funcionalidades previstas pelo sistema, é possível: (1) navegar por meio de relações (Figura 13); (2) visualizar um grupo específico do córpus (Figura 14); e (3) baixar o CSTNews em um arquivo comprimido. O córpus CST pode ser obtido através da página do NILC[8] e está disponível de forma livre.

[8] http://nilc.icmc.usp.br/nilc/tools/CSTNews

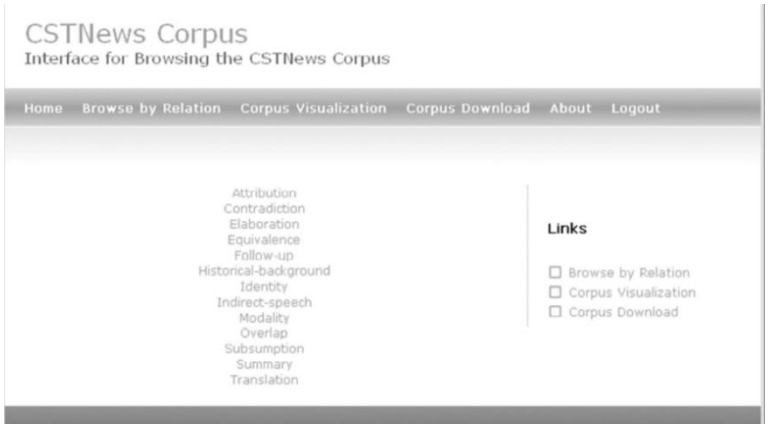


**Figura 13 - Navegação por meio das relaões do córpus**


**Figura 14 - Visualização do primeiro grupo do córpus**

Os arquivos no córpus CSTNews estão organizados em pastas, sendo cada pasta relativa a um grupo de documentos, variando de 1 a 50. Na Tabela 9, exemplifica-se a função de cada arquivo contido nas pastas, tomando como exemplo o grupo 1.

Os itens de documento 1 ao 3 são os arquivos em texto puro extraídos de suas fontes. O texto apresentado possui apenas o corpo da notícia, não apresentando, por exemplo, o título, autor e subtítulo.

**Tabela 9. Descrição dos arquivos contidos dentro do grupo 1 do córpus CSTNews (Diretório C1_Mundo_AviaoCongo)**

| Item | Arquivo | Descrição |
|---|---|---|
| 1 | D1_C1_Folha.txt | Documento 1 (D1) do grupo 1 (C1) pertencente a fonte Jornal Folha de São Paulo. |
| 2 | D2_C1_Estadao.txt | Documento 2 (D2) do grupo 1 (C1) pertencente a fonte Jornal O Estado de São Paulo. |
| 3 | D3_C1_JB.txt | Documento 3 (D3) do grupo 1 (C1) pertencente a fonte Jornal do Brasil. |
| 4 | D1_C1_Folha.txt.seg | Arquivo resultante da segmentação do arquivo D1_C1_Folha.txt |
| 5 | D2_C1_Estadao.txt.seg | Arquivo resultante da segmentação do arquivo D2_C1_Estadao.txt |
| 6 | D3_C1_JB.txt.seg | Arquivo resultante da segmentação do arquivo D3_C1_JB.txt |
| 7 | D1_C1_Folha.txt.xml | Arquivo com as marcações xml do arquivo D1_C1_Folha.txt |
| 8 | D2_C1_Estadao.txt.xml | Arquivo com as marcações xml do arquivo D2_C1_Estadao.txt |
| 9 | D3_C1_JB.txt.xml | Arquivo com as marcações xml do arquivo D3_C1_JB.txt |
| 10 | D1_D2_C1.cst | Arquivo contendo as relações pertencentes entre os pares de documentos D1 e D2. |
| 11 | D1_D3_C1.cst | Arquivo contendo as relações pertencentes entre os pares de documentos D1 e D3. |
| 12 | D2_D3_C1.cst | Arquivo contendo as relações pertencentes entre os pares de documentos D2 e D3. |
| 13 | C1_resumo | Resumo da grupo de textos do tipo *abstract* feito por humanos. |

Os itens 4 a 6 são resultados da execução dos itens 1 a 3 no software de segmentação textual sentencial SENTER, (em que a segmentação foi corrigida manualmente durante a anotação). Os itens 7 a 9 são documentos que apresentam os textos extraídos de suas fontes (1 ao 3) em uma marcação XML própria para seu processamento.

Na pasta raiz do córpus, além das pastas de cada grupo de documentos, também pode ser encontrado o arquivo CSTNews.sentrel, que contém todas as relações entre todos os pares de segmentos presentes no córpus. Esse arquivo foi obtido através da mesclagem dos arquivos de relações CST presentes nas pastas dos grupos. Também são encontradas as gramáticas docsent.dtd e sentrel.dtd, relativas aos arquivos XML de sentenças e de relações respectivamente.

# 5. A Análise Multidocumento

Neste capítulo, trata-se da metodologia de *parsing* multidocumento. Esta seção está dividida em três principais partes. Na primeira seção, apresenta-se a arquitetura do *parser* e cada passo é descrito. Essa arquitetura serviu como base para o desenvolvimento do *parser*. Na segunda parte, trata-se de toda a exploração de cenários e técnicas na criação de classificadores utilizando aprendizado de máquina. Na última parte, as regras, que tratam algumas relações CST não abordadas pelos classificadores, são apresentadas.

O principal objetivo deste trabalho é a apresentação de uma metodologia de identificação automática que resultou em um *parser* a ser utilizado por diversas aplicações que necessitem de uma estruturação multidocumento.

## 5.1. O *parser* multidocumento

### 5.1.1. Arquitetura da ferramenta

A Figura 15 apresenta a arquitetura do *parser* multidocumento desenvolvido. O processo inicia-se na obtenção de um grupo de textos que tratam de um mesmo tópico, passando pela segmentação em sentenças e identificação dos pares potenciais a serem relacionados por alguma relação CST. A identificação das relações é feita primeiramente pela utilização de classificadores que identificam as relações de "conteúdo" e, posteriormente, um conjunto de regras faz a identificação das relações de "forma" e de uma relação de "conteúdo", a relação *Contradiction*. O resultado do *parsing* é apresentado na forma de um grafo, representado em um arquivo XML.

A seguir, cada etapa da análise multidocumento é apresentada através da arquitetura do *parser* multidocumento. Essas etapas guiaram todo o desenvolvimento da análise automática.

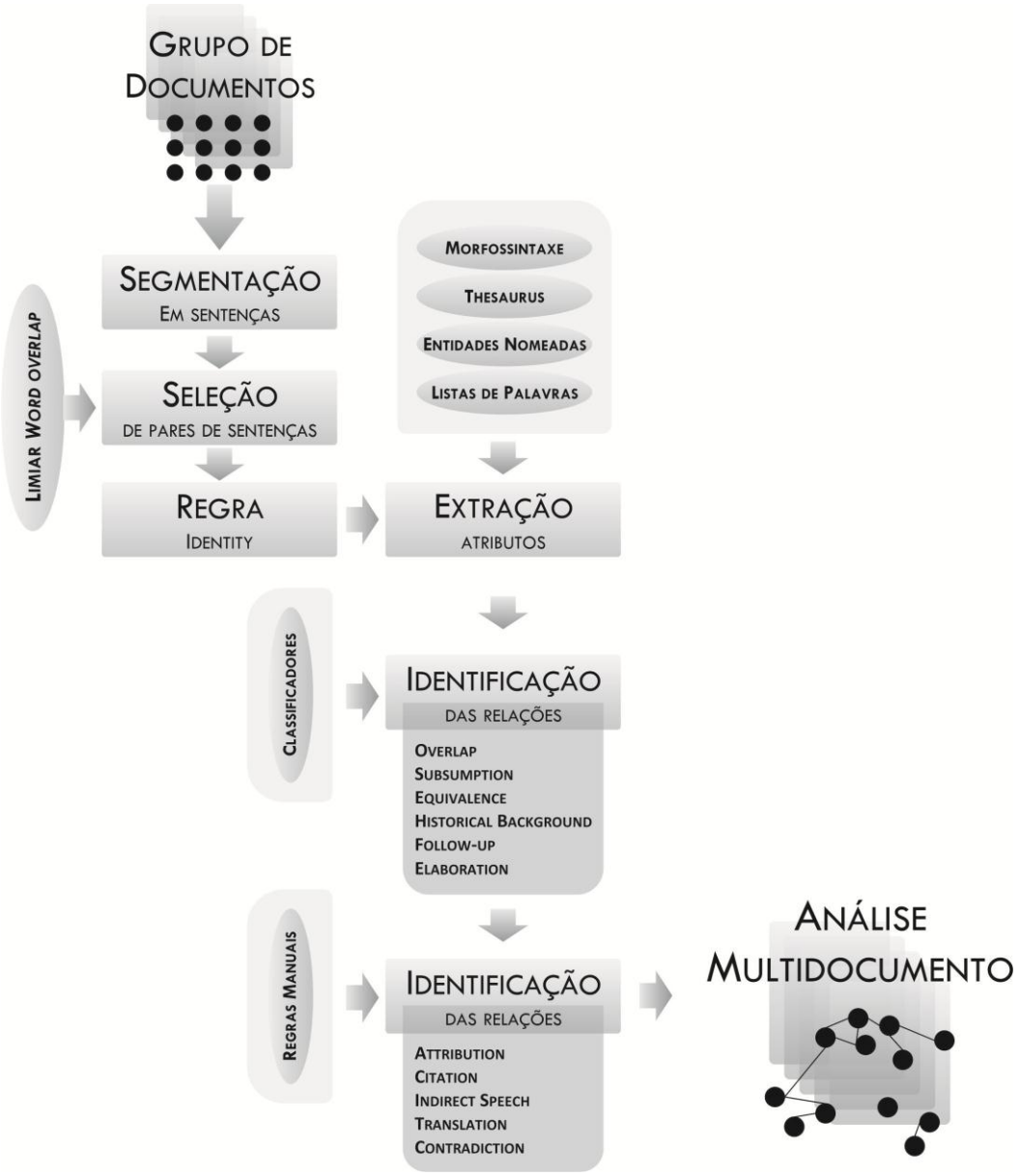


Figura 15 - Arquitetura do *parser* multidocumento

### 5.1.1.1. Entrada: grupo de documentos

A entrada para a análise multidocumento consiste de um conjunto de documentos que tratam de um mesmo assunto. Por exemplo, um conjunto de notícias sobre um mesmo acontecimento, publicadas por diversos jornais. O *parser* pode obter esses conjuntos de documentos de duas maneiras. Uma delas é através do fornecimento de cada documento pelo usuário. Dessa maneira, o responsável por selecionar os documentos correlatos é o

usuário do *parser*. Outra forma é obter o conjunto através de algum buscador *web*, que recebe como entrada palavras-chave e retorna os endereços das páginas que contêm os documentos que tratam do assunto fornecido. Nesse último caso, é utilizada uma API (*Application Programming Interface*) de um sistema de busca de notícias. Uma grande dificuldade nessa segunda opção, em que os documentos são as páginas *web*, é a filtragem do corpo da notícia, visto que cada portal de notícia utiliza esquemas de *tags* diferentes para identificar o corpo da notícia, de onde se extrai o texto.

#### 5.1.1.2. Seleção de pares de sentenças

Tendo o conjunto de documentos para análise, cada documento é segmentado em sentenças utilizando o segmentador SENTER (Pardo, 2006). As relações CST ocorrem sempre entre dois segmentos (sentenças) de documentos diferentes e, para um mesmo par, podem ocorrer duas relações (uma de "conteúdo" e uma de "forma"). Assim, cada possível combinação de duas sentenças entre as possíveis combinações de documentos teriam de ser verificadas. Dessa forma, se fossem 4 documentos (D1, D2, D3, D4), contendo cada um 10 sentenças (S1..S10), seriam 6 combinações de documentos (D1-D2, D1-D3, D1-D4, D2-D3, D2-D4, D3-D4) e, para cada uma dessas comparações, seriam 100 combinações entre as sentenças dos dois documentos (D1:S1-D2:S1, D1:S1-D2:S2, D1:S1-D2:S3 ... D1:S1-D2:S10, D1:S2-D2:S1 ... D1:S2-D2:S10 ... D1:S10-D2:S10; em que Dx:Sy indica sentença y do documento x). Para esse conjunto de documentos, seria necessário verificar 600 pares de sentenças. Para um conjunto de 10 documentos, são possíveis 45 combinações de documentos (10!/(2!*(10-2)!) = 45). Tendo cada documento 15 sentenças cada, são 225 pares de sentenças (15*15 = 225) para cada par de documentos. Assim, 10125 pares de sentenças (45*225 = 10125) deveriam ser analisados, tornando o processo muito demorado, dado o custo de extração de atributos para cada sentença.

Além do exposto anteriormente, é conhecido que as relações CST ocorrem entre sentenças que têm alguma similaridade lexical entre si (Zhang e Radev, 2005). Assim, para minimizar o número de pares de sentenças a analisar, para cada combinação de sentenças é calculada uma medida chamada de *word overlap* (**Equação 1**, página 26). Cada par de sentenças que tiver o valor da medida acima de determinado valor é selecionado para as

próximas etapas da análise (geralmente 0.12, mesmo valor utilizado para a língua inglesa, seguindo o trabalho de Aleixo e Pardo (2008b)). Os demais pares são desprezados. A medida *word overlap* identifica o relacionamento semântico por meio do léxico das sentenças.

Como resultado dessa etapa, uma lista contendo a indicação dos pares de sentenças passíveis de serem relacionados é criada, servindo como entrada para as próximas etapas. Nessa lista, são indicados os documentos e as sentenças de cada par.

#### 5.1.1.3. Aplicação da regra *Identity*

Essa etapa serve para poupar ainda mais as próximas etapas de analisar pares de sentenças que sejam idênticos, visto que, nesse caso, estabelece-se a relação *Identity*. Essa regra será descrita ainda neste capítulo. Quando identificada essa relação, o par de sentenças que a contém é excluído da lista de pares de sentenças a serem analisadas, pois não precisam ser relacionados por outra relação CST.

Aqui, é criado um arquivo (com extensão ".cst") que conterá todas as relações identificadas em todos os passos da análise. As relações *Identity*, portanto, são as primeiras a serem adicionadas a esse arquivo.

#### 5.1.1.4. Extração de atributos e classificação

Os pares de sentenças têm os valores de seus atributos extraídos utilizando-se diversas ferramentas e recursos. Abaixo, descreve-se a obtenção de cada atributo, assim como a motivação de sua escolha.

Feita a extração dos valores dos atributos, uma lista contendo os pares de sentenças e seus respectivos atributos é gerada e é utilizada como entrada aos classificadores. A escolha do tipo de classificador a ser utilizado é um parâmetro que pode ser definido pelo usuário (binário, hierárquico ou multiclasse).

Nessa etapa de classificação, são identificadas apenas as relações de "conteúdo", dada a frequência das relações desse tipo. Feita a classificação, o arquivo ".cst" é acrescido

das relações que foram identificadas pelos classificadores. Diferentemente da relação *Identity*, os pares de sentença que tiverem alguma relação identificada neste passo não são eliminados da lista de pares de sentenças a analisar, pois podem conter alguma relação de "forma". Assim, procede-se para o próximo passo.

#### 5.1.1.5. Aplicação das regras

Os pares de sentenças que contêm alguma relação de "conteúdo" identificada pelos classificadores do passo anterior são então analisadas através das regras. Na escolha de uma relação, adiciona-se ao arquivo ".cst" a relação identificada.

#### 5.1.1.6. Apresentação da análise CST

Como resultado dos passos anteriores, tem-se o arquivo ".cst" contendo todas as relações identificadas para o conjunto de textos utilizados como entrada da análise. Esse resultado pode ser utilizado por alguma outra ferramenta que precise de conhecimento das relações entre os documentos em análise, como buscadores *web*, sumarizadores multidocumento, etc.

Uma interface *web* foi implementada a fim de que todo o processo anterior possa ser executado por qualquer usuário na *web*. Nessa interface, o resultado é apresentado na forma de pares de sentenças e as relações entre esses pares, permitindo uma experiência de navegação multidocumento pelos usuários.

### 5.1.2. Relações desconsideradas

Nem todas as relações exibidas na tipologia da Figura 10 foram consideradas neste trabalho. A principal razão é sua escassez no córpus utilizado, assim como a complexidade na identificação de algumas relações. As relações *Modality*, *Summary* e um tipo de *Contradiction* foram desconsideradas. A relação *Modality* apresenta apenas 1 exemplo, impossibilitando o aprendizado automático, e sua definição apresenta muita subjetividade para tratamento via regras. Veja a restrição da relação, contida no Anexo A: "*S1 e S2*

*apresentam informação em comum e em S2 a fonte/autoria da informação é indeterminada/relativizada/amenizada*".

A relação *Summary* também não foi tratada devido à baixa frequência no córpus (apenas 4 exemplos) e sua identificação consiste em verificar se uma sentença resume todas as informações de outra sentença.

A baixa frequência no córpus não é o único fator para o não tratamento de algumas relações, haja vista a relação *Translation*, que contém apenas dois exemplos, mas sua descrição leva a um tratamento possível por regras.

Aqui vale diferenciar dois tipos de relacionamento segundo a relação *Contradiction*, as que são evidenciadas pelas informações explícitas (*Contradiction* explícita) no texto das sentenças e as que necessitam de inferência para serem identificadas (*Contradiction* de inferência).

A relação *Contradiction* por inferência tem sua dificuldade na complexidade de fazer inferências gerais em duas sentenças a fim de identificar a contradição daquilo que é inferido.

Como exemplo de relação *Contradiction* explícita, considere o par de sentenças abaixo extraído do córpus. Em S1, diz-se que a Secretaria da Fazenda foi atingida por três bombas e S2 diz que a Secretaria foi atingida por uma bomba. A divergência é numérica e está claramente expressa na superfície textual das duas sentenças.

(S1) *O prédio da secretaria da Fazenda, no centro, foi atingido por três bombas caseiras.*

(S2) *A Secretaria da Fazenda também foi atingida por uma bomba.*

A exemplo de *Contradiction* de inferência, considere o exemplo abaixo. Em S1 diz-se que a aeronave não apresentou problemas no dia 16 de julho, em S2, diz-se que no mesmo dia foram apresentados problemas pela aeronave. Trata-se de informações que não são numéricas e necessitam de certa inferência para ser percebida.

(S1) *Em nota enviada após a exibição da reportagem, a TAM afirma "que não teve registro de qualquer problema mecânico neste avião no dia 16 de julho".*

(S2) *Um dia antes do acidente, na segunda-feira, 16, o avião também teria apresentado problemas ao aterrissar em Congonhas, durante o vôo 3215, procedente de Belo Horizonte (Confins), só conseguindo parar muito próximo do final da pista.*

## 5.2. Identificação com técnicas de aprendizado automático

Nesta seção, são relatados experimentos realizados com diversas técnicas de aprendizado automático (probabilísticas, estatísticas e simbólicas), tanto tradicionais quanto hierárquicas, em diversos cenários (definidos pelo conjunto de relações consideradas) na criação de classificadores para identificação das relações CST. Esses experimentos foram possíveis graças à existência e disponibilidade do córpus CSTNews, já apresentado.

Diversas questões devem ser consideradas nessa abordagem, como i) a sobreposição entre as classes (relações de "conteúdo" e de "forma") e ii) o desbalanceamento das relações no córpus. Por exemplo, relações como *Overlap* e *Subsumption* são muito mais frequentes que relações como *Modality* e *Translation*.

A sobreposição das relações também é algo natural, dada a diferenciação entre relações de "conteúdo" e de "forma", que geralmente co-ocorrem entre pares de sentenças. Um par de sentenças que contém sobreposição de informações e é relacionado, consequentemente, pela relação *Overlap*, pode apresentar também uma relação de "forma", como *Indirect-speech*.

Na consideração dessas questões anteriormente apresentadas, e na realização de diversos experimentos, chegou-se a decisão de considerar, no aprendizado automático, apenas um conjunto de relações: as que apresentaram os melhores resultados nos classificadores. Todos os experimentos, no entanto, são apresentados e discutidos, assim como seus resultados, permitindo visualizar a trajetória até a escolha do melhor cenário.

Nas próximas subseções, serão apresentados dados gerais sobre os experimentos, como as ferramentas e recursos utilizados, os atributos utilizados na criação dos classificadores, as questões de desbalanceamento, sobreposição das classes e seleção de atributos. Por fim, são apresentados os experimentos em duas subseções: uma trata sobre os primeiros experimentos, considerando, na maioria dos casos, todas as relações CST; e, na outra subseção, relatam-se os experimentos considerando apenas as relações de conteúdo. Conclui-se com comentários e comparações dos resultados obtidos.

### 5.2.1. Ferramentas e recursos

Para cada experimento, foram criados três classificadores e esses foram comparados utilizando o recurso Experimenter[9], da ferramenta WEKA (Waikato Environment for Knowledge Analysis; Witten e Frank, 2005). Essa ferramenta é amplamente utilizada no meio acadêmico quando se explora técnicas de aprendizado de máquina, pois contém uma grande quantidade de algoritmos de aprendizado automático implementados.

As técnicas para criação dos classificadores foram: NaiveBayes (John e Langley, 1995), SVM (*Suport Vector Machine;* Vapnik (1995)) e J48 (Quinlan, 1993). NaiveBayes é uma técnica probabilística, SVM é estatística, e J48 é simbólica, criando uma árvore de decisão. O classificador OneR (Holte, 1993) também foi executado para exemplificar outra técnica simbólica, além da J48. A avaliação de cada classificador foi realizada utilizando a validação cruzada de 10 pastas.

O principal recurso que possibilitou a realização desses experimentos foi o córpus CSTNews, já apresentado no Capítulo 4. O córpus foi pré-processado a fim de gerar os dados em formato adequado à realização dos experimentos, obtendo-se os valores dos atributos, considerados na próxima subseção.

[9] Permite realizar diversos experimentos automaticamente e conduzir testes estatísticos nos resultados.

### 5.2.2. Os atributos

A Tabela 10 apresenta os atributos extraídos do córpus utilizado a fim de gerar os classificadores. Após o cômputo (descrito a seguir para cada atributo), todos os atributos são normalizados, evitando-se, assim, possíveis discrepâncias na classificação.

Os primeiros 6 atributos foram obtidos utilizando apenas a superfície textual, trabalhando-se com as palavras das sentenças. O atributo 1 indica a diferença no tamanho das sentenças, para o caso de sentenças que, por exemplo, contenham a mesma informação de outra sentença e informação adicional. Nesse caso a maior sentença engloba a menor e uma relação *Subsumption* pode ocorrer. Os atributo 2 e 3 indicam as porcentagens de palavras em comum entre as sentenças, presentes em S1 e S2, respectivamente. Esses valores podem indicar a ocorrência de alguma relação de "conteúdo", por indicar a sobreposição de informações. Os atributos 4 e 5 indicam a posição das sentenças S1 e S2, respectivamente, no texto. Essa informação é valiosa em textos do gênero jornalístico, que seguem uma ordem de apresentação das informações no texto de uma notícia. O atributo 6 mede a sobreposição da maior *substring* entre as sentenças, indicando o grau de redundância de informações entre as sentenças.

Para a extração dos atributos 7 a 12, foi utilizada a ferramenta de etiquetação morfossintática MXPOST (Ratnaparkhi, 1996) treinada para o Português do Brasil (Aires et al., 2000). Esse etiquetador realiza uma análise morfossintática dos textos, apresentando as classes gramaticais das palavras com uma precisão de mais de 96%. Esses atributos não verificam a palavra em si, mas a quantidade de palavras de mesma classe gramatical presente nas sentenças, como um indício da existência de alguma relação de "conteúdo" entre o par de sentenças.

Os atributos 13 e 14 foram obtidos utilizando o *parser* sintático Palavras (Bick, 2000). Bick reporta um desempenho médio de mais de 98% na análise sintática. O *parser* também realiza a lematização dos verbos, que foi necessária para realizar uma busca em uma lista de verbos de atribuição, a fim de calcular o valor do atributo 13. Esse atributo visa identificar principalmente as relações *Attribution* e *Citation*, em que há a existência de um indicador de atribuição, principalmente de um verbo de atribuição ou outro indicador de atribuição (como "segundo...", "de acordo com...", etc.). Para o atributo 14, a base de

sinônimos TeP 2.0 foi utilizada (Maziero et al., 2008). O TeP (Thesaurus para o Português) armazena determinadas informações provenientes da base da WordNet.Br (doravante, WN.Br; Di Felippo e Dias-da-Silva, 2007; Dias-da-Silva et al., 2007). Essa base contém 19888 conjuntos de sinônimos (*synsets*), com um total de 44678 palavras; conta também com 18163 relações de antonímia entre os *synsets*. Neste atributo 14, para cada palavra, desconsiderando as *stopwords*, foi compilada uma lista de identificadores de todos os conjuntos de sinônimos em que ocorre cada palavra, dado que não foi realizada a desambiguação lexical. Após o cruzamento das listas de cada palavra de cada sentença, o valor do atributo foi gerado. A base de sinônimos é essencial para identificar a sobreposição de palavras que não são idênticas, mas pertencem a um mesmo conjunto de sinônimos, o que auxilia na identificação da relação *Equivalence*, principalmente.

**Tabela 10 - Atributos utilizados nos experimentos**

| |
|---|
| 1- Diferença de tamanho em palavras (S1-S2) |
| 2- Porcentagem de palavras em comum em S1 |
| 3- Porcentagem de palavras em comum em S2 |
| 4- Posição de S1 no texto (0- início, 2- fim, 1- meio) |
| 5- Posição de S2 no texto (0- início, 2- fim, 1- meio) |
| 6- Número de palavras na maior substring entre S1 e S2 |
| 7- Diferença no número de substantivos entre S1 e S2 |
| 8- Diferença no número de advérbios entre S1 e S2 |
| 9- Diferença no número de adjetivos entre S1 e S2 |
| 10- Diferença no número de verbos entre S1 e S2 |
| 11- Diferença no número de nomes próprios entre S1 e S2 |
| 12- Diferença no número de numerais entre S1 e S2 |
| 13- Diferença no número de verbos de atribuição entre S1 e S2 |
| 14- Sobreposição de sinônimos entre S1 e S2 |

### 5.2.3. Desbalanceamento e sobreposição

Como percebido pela Figura 12, o desbalanceamento é algo natural nesse tipo de tarefa. Por se tratar de textos do gênero jornalístico, há mais relações de "conteúdo" do que relações de "forma".

Como será mostrada no decorrer desta seção, a solução para o problema da sobreposição das classes foi considerar apenas as relações de "conteúdo" no aprendizado

automático e tratar as relações de “forma” por meio de regras. Essa estratégia busca contornar as duas características indesejáveis: desbalanceamento e sobreposição das relações. Contorna-se o desbalanceamento, pois as relações de “forma” são as menos frequentes e não serão consideradas nos classificadores, que agora contarão com um conjunto mais balanceado, apesar de não totalmente balanceado. Contorna-se a sobreposição, pois os classificadores trataram classes que não se sobrepõem, visto que tratam apenas das relações de “conteúdo”.

Nas próximas seções, os resultados serão apresentados para os dados desbalanceados e para os dados balanceados segundo a abordagem anterior, de replicação dos exemplos menos frequentes.

O *parser* multidocumento utiliza classificadores construídos sobre um conjunto de dados não balanceados, com as relações mais frequentes do córpus, devido *overfitting* que pode ocorrer quando usado o córpus balanceado artificialmente por replicação.

### 5.2.4. Seleção de atributos

Na tarefa de classificação, foram utilizados 14 atributos, desde superficiais (como sobreposição de palavras) até atributos que necessitem de algum conhecimento linguístico (como contagem de classes de palavras). A fim de verificar se a tarefa poderia ser executada com menos atributos, mantendo mesmo desempenho ou superior, técnicas de seleção de atributos foram utilizadas.

Foi utilizada uma técnica que avalia cada atributo em relação ao ganho de informação relacionado à classe. Esses atributos são então ordenados de acordo com o valor obtido e o seguinte ranking é fornecido: 3, 6, 1, 7, 2, 12, 11, 10, 14, 4, 9, 8, 5 e 13. Apenas essa técnica de seleção de atributos foi empregada por ser a mais utilizada.

Foi feito um teste com o classificador multiclasse para as seis relações de conteúdo dos últimos experimentos (a partir da página 68) e verificou-se que, retirando os atributos 5 e 13, o valor aumentou em 0.0701%. Assim, esses dois atributos (5: Posição de S2 no texto; e 13: Diferença no número de verbos de atribuição entre S1 e S2) foram desconsiderados na criação dos classificadores. Se mais atributos (de trás para frente do ranking) forem desconsiderados, o desempenho do classificador começa a diminuir.

### 5.2.5. Os primeiros experimentos

Inicialmente, foram explorados alguns cenários na criação dos classificadores. Dois classificadores multiclasse foram criados, um considerando todas as relações CST (apresentando os problemas de desbalanceamento e sobreposição de classes) e outro com apenas algumas relações de "conteúdo", as mais frequentes. Dada a sobreposição de classes, característica da CST, criou-se um classificador multirótulo (que gerou mais desbalanceamento nas classes). Motivado pela hierarquia das relações, dois classificadores hierárquicos foram desenvolvidos, segundo duas abordagens diferentes, a saber: *top-down* e *big-bang*. Por fim, explorou-se a criação de classificadores binários, para as relações mais frequentes do córpus.

Visto que foram utilizadas três técnicas de aprendizado supervisionado (NaiveBayes, SVM e J48), utilizou-se o recurso Experimenter da ferramenta WEKA, e foi aplicado o teste de significância T pareado (com valor de confiança igual a 95%) para indicação da melhor técnica em cada experimento. Assim, as tabelas com os resultados para cada experimento correspondem à dos classificadores que obtiveram os melhores valores. Quando não houve diferença estatisticamente significativa, foi dada prioridade ao classificador simbólico (*J48*), por se tratar de uma técnica simbólica e permitir que o modelo gerado seja mais facilmente interpretado. Nas tabelas 12, 14, 16, 18 e 20 são apresentados os testes estatísticos para cada experimento. A ocorrência de (e) após o valor indica *empata*, (p) indica *perde* e (g) indica ganha, com relação à técnica NaiveBayes.

Vale salientar que, como será observado nos resultados no decorrer desta seção, a técnica J48 obtém resultados melhores que os da técnica SVM na maioria dos testes de significância e, em muitos casos em que o SVM obtém resultado acima do J48, essa diferença não é significativa. Portanto, para os experimentos, todos os classificadores foram criados utilizando o algoritmo J48.

### 5.2.5.1. Classificadores Multiclasse

Dois classificadores multiclasse foram construídos: o primeiro considera todas as relações CST e, portanto, sofre mais com o desbalanceamento e sobreposição das classes. No segundo, tratam-se esses dois problemas desconsiderando-se algumas relações, principalmente as relações de "forma".

Os resultados descritos a seguir mostram que o desbalanceamento e a sobreposição das classes diminuem o desempenho dos classificadores. As relações com poucos exemplos têm desempenho nulo. Esse experimento foi realizado inicialmente apenas para averiguar o que se esperava: baixos resultados para as relações menos frequentes. Como se analisará, tentativas de balanceamento não são viáveis neste caso, pois, por exemplo, a relação *Translation* apresenta frequência menor que 1% (0,43%) da relação mais frequente (*Overlap*).

#### *5.2.5.1.1. Todas as relações CST*

Como algumas relações não apresentam frequência suficiente para um bom aprendizado, os resultados para essas relações não são satisfatórios. Os resultados (Tabela 11) mostram isso pelos baixos valores para as relações *Modality*, *Indirect-Speech*, *Summary* e *Translation*. Já as relações com maior frequência tiveram medida-F maior que 0.3, como *Subsumption*, *Elaboration* e *Overlap*. Essas relações de "conteúdo", além da alta frequência no córpus, são passíveis de identificação pelos atributos utilizados, que refletem principalmente o conteúdo das sentenças.

Pelo teste de significância da Tabela 12, vemos que a técnica J48 empata com a NaiveBayes e se comporta melhor que a SVM. Assim, o classificador foi criado utilizando J48.

**Tabela 11 – Resultados do classificador multiclasse para todas relações**

| Relação | Precisão | Cobertura | Medida-F |
|---|---|---|---|
| *Subsumption* | 0.38 | 0.44 | 0.40 |
| *Elaboration* | 0.38 | 0.39 | 0.38 |
| *Attribution* | 0.15 | 0.10 | 0.12 |
| *Overlap* | 0.41 | 0.45 | 0.43 |
| *Follow-up* | 0.29 | 0.30 | 0.30 |
| *Historical-background* | 0.17 | 0.13 | 0.15 |
| *Contradiction* | 0.10 | 0.04 | 0.06 |
| *Identity* | 0.89 | 0.95 | 0.92 |
| *Equivalence* | 0.19 | 0.13 | 0.15 |
| *Modality* | 0 | 0 | 0 |
| *Indirect Speech* | 0 | 0 | 0 |
| *Summary* | 0 | 0 | 0 |
| *Translation* | 0 | 0 | 0 |
| **Media** | **0.23** | **0.22** | **0.22** |

**Tabela 12 - Resultado do teste de significância estatística**

| NaiveBayes | SVM | J48 |
|---|---|---|
| 0.3858 | 0.3402 (p) | 0.3700 (e) |

### *5.2.5.1.2. Relações de "conteúdo"*

Como exemplo do desbalanceamento, a relação *Citation* não tem ocorrências no córpus, pois um jornal não costuma citar a outro. Dessa forma, optou-se pela criação de um classificador que considere apenas as relações mais frequentes. Relações de "conteúdo" são estabelecidas entre sentenças que apresentam alguma sobreposição de informações entre si. Veja a Figura 10, em que é apresentada a tipologia das relações.

A Tabela 13 exibe os resultados do classificador (utilizando o algoritmo J48, com taxa de acerto de 0.402) para as relações de "conteúdo", excetuando-se as relações *Summary* e *Contradiction*. As relações *Overlap* e *Subsumption* tiveram medidas-F maiores que 0.48. Já a relação *Historical-background* teve o valor de 0.23, por ser pouco frequente e, dadas suas características, não ser satisfatoriamente identificada com os atributos utilizados.

A fim de diminuir a confusão na identificação das relações, escolheu-se um grupo menor de classes para compor o aprendizado. Esse conjunto é composto das seguintes

relações de "conteúdo": *Identity*, *Equivalence*, *Overlap*, *Subsumption*, *Historical-background*, *Follow-up* e *Elaboration*. Essa escolha foi motivada tanto pela frequência dessas relações quanto pela sua principal característica: sobreposição de informações. Essa sobreposição de informações é manifesta principalmente na superfície textual das sentenças por meio de palavras semelhantes. Os atributos escolhidos refletem bem essas características na superfície textual. Os resultados (Tabela 13) constituem o estado da arte da identificação das relações CST.

Para esse cenário, novamente a técnica J48 empata com o NaiveBayes e, agora, também com a SVM (Tabela 14). Foi utilizado o algoritmo J48 para a criação do classificador multiclasse.

**Tabela 13 - Resultados do classificador multiclasse para relações de "conteúdo"**

| Relação | Precisão | Cobertura | Medida-F |
|---|---|---|---|
| *Subsumption* | 0.485 | 0.560 | 0.520 |
| *Elaboration* | 0.386 | 0.382 | 0.384 |
| *Overlap* | 0.486 | 0.503 | 0.494 |
| *Follow-up* | 0.317 | 0.287 | 0.301 |
| *Historical back* | 0.243 | 0.221 | 0.231 |
| *Identity* | 0.941 | 0.941 | 0.941 |
| *Equivalence* | 0.207 | 0.154 | 0.176 |
| **Media** | **0.438** | **0.435** | **0.435** |

**Tabela 14 - Resultado do teste de significância estatística**

| NaiveBayes | SVM | J48 |
|---|---|---|
| 0.4161 | 0.3904 (e) | 0.3970 (e) |

Os atributos escolhidos para esses experimentos refletem principalmente a sobreposição do conteúdo das sentenças. As outras relações, que não são de "conteúdo", dependem de outros conhecimentos, como a estrutura sintática e informação semântica das sentenças. Inclusive, relações como *Follow-up* e *Historical-background*, embora sejam relações de "conteúdo", também se beneficiariam desses conhecimentos mais linguísticos na

identificação da evolução temporal de um evento e da elaboração histórica de um elemento.

### 5.2.5.2. Classificadores Multirótulo

Tendo em vista a sobreposição das classes, explorou-se uma metodologia de identificação multirótulo. Essa metodologia consiste em identificar os pares de sentenças que têm mais de uma relação CST (no máximo 2, como pode ser visto pelos rótulos gerados na Tabela 15) e gerar um novo rótulo contendo o nome das duas relações. No entanto, essa metodologia leva a um maior desbalanceamento das classes. A Figura 16 mostra a quantidade de cada classe (rótulos) obtida no processamento do córpus para obter os multirótulos.

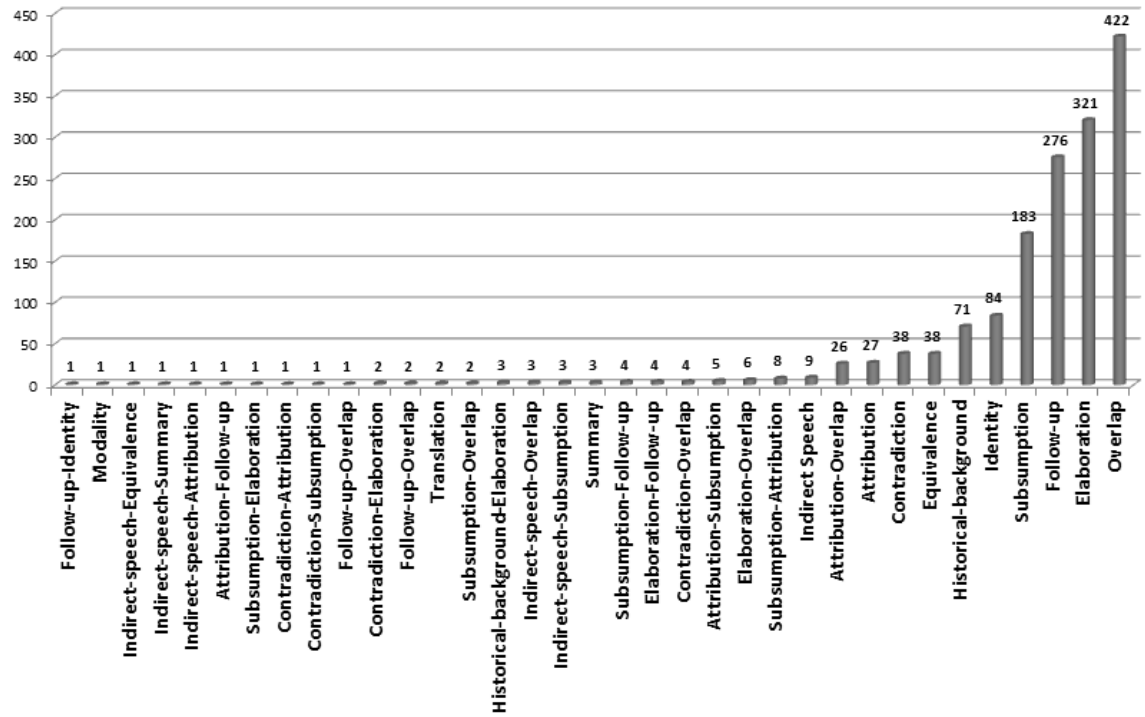


**Figura 16 - Frequência das classes no classificador multirótulo**

Os resultados da classificação multirótulo estão na Tabela 15. Vê-se que as novas classes criadas, contendo mais de uma relação CST, ficaram com valores nulos, exceto a classe "*attribution-overlap*" e "*subsumption-follow-up*", que tiveram uma frequência suficiente

para algum aprendizado. Assim, a abordagem multirótulo, da forma como foi realizada, é inviável dadas as características do único córpus disponível para o aprendizado.

**Tabela 15 - Resultados do classificador multirótulo**

| Relação | Precisão | Cobertura | Medida-f |
|---|---|---|---|
| *Subsumption* | 0.380 | 0.420 | 0.400 |
| *Elaboration* | 0.330 | 0.360 | 0.340 |
| *Historical-background-Elaboration* | 0 | 0 | 0 |
| *Attribution-Overlap* | 0.030 | 0.040 | 0.040 |
| *Subsumption-Follow-up* | 0.400 | 0.50 | 0.440 |
| *Overlap* | 0.410 | 0.440 | 0.420 |
| *Historical-background* | 0.300 | 0.310 | 0.300 |
| *Follow-up* | 0.300 | 0.300 | 0.300 |
| *Contradiction* | 0.190 | 0.110 | 0.140 |
| *Subsumption-Attribution* | 0 | 0 | 0 |
| *Identity* | 0.900 | 0.950 | 0.920 |
| *Equivalence* | 0.230 | 0.160 | 0.190 |
| *Follow-up-Identity* | 0 | 0 | 0 |
| *Elaboration-Overlap* | 0 | 0 | 0 |
| *Contradiction-Elaboration* | 0 | 0 | 0 |
| *Attribution* | 0 | 0 | 0 |
| *Modality* | 0 | 0 | 0 |
| *Indirect-speech-Overlap* | 0 | 0 | 0 |
| *Indirect-speech-Equivalence* | 0 | 0 | 0 |
| *Indirect Speech* | 0 | 0 | 0 |
| *Indirect-speech-Summary* | 0 | 0 | 0 |
| *Indirect-speech-Subsumption* | 0 | 0 | 0 |
| *Indirect-speech-Attribution* | 0 | 0 | 0 |
| *Elaboration-Follow-up* | 0 | 0 | 0 |
| *Attribution-Subsumption* | 0 | 0 | 0 |
| *Follow-up-Overlap* | 0 | 0 | 0 |
| *Translation* | 0 | 0 | 0 |
| *Summary* | 0 | 0 | 0 |
| *Attribution-Follow-up* | 0 | 0 | 0 |
| *Subsumption-Elaboration* | 0 | 0 | 0 |
| *Contradiction-Attribution* | 0 | 0 | 0 |
| *Contradiction-Subsumption* | 0 | 0 | 0 |
| *Contradiction-Overlap* | 0 | 0 | 0 |
| *Follow-up-Overlap* | 0 | 0 | 0 |
| *Subsumption-Overlap* | 0 | 0 | 0 |
| **Media** | **0.100** | **0.100** | **0.100** |

Neste cenário, a técnica J48 ganha tanto do NaiveBayes quanto do SVM (Tabela 16) e é utilizada na criação do classificador multirótulo.

**Tabela 16 - Resultado do teste de significância estatística**

| NaiveBayes | SVM | J48 |
|---|---|---|
| 0.3274 | 0.3495 (e) | 0.3765 (g) |

### 5.2.5.3. Classificadores Binários

Para algumas relações, foram criados classificadores binários. Esses classificadores têm de escolher entre apenas duas possíveis classes. Como foram consideradas, neste cenário, oito relações CST, foram criados oito classificadores. Cada um desses classificadores verifica a existência ou não de cada relação entre os pares de sentenças em análise. Assim, cada par de sentenças é verificado por cada classificador, ou seja, verifica-se a existência de cada relação independentemente. Após a aplicação desses classificadores, deve-se optar por apenas uma das relações identificadas.

As relações consideradas nesse experimento são: *Attribution*, *Contradiction*, *Elaboration*, *Equivalence*, *Follow-up*, *Historical-background*, *Overlap* e *Subsumption*. Cada classificador escolherá entre uma das relações citadas e uma classe chamada "Outra". Para gerar os dados de treinamento e teste para, por exemplo, o classificador binário que identifica a relação *Elaboration*, todas as outras relações foram consideradas como sendo a relação "Outra". Dessa forma, cada instância ou era da classe *Elaboration* ou da classe "Outra", permitindo a geração de classificadores binários, por conter apenas duas classes.

Como são utilizados oito classificadores binários e mais de um pode escolher uma relação CST (que não seja a relação "Outra"), o critério de escolha entre as relações identificadas, no caso de mais de uma, é a confiabilidade da escolha, informada pelos classificadores binários. Assim, por exemplo, se os classificadores das relações *Subsumption* e *Overlap* identificam essas relações, com respectivas confiabilidades 0.8 e 0.9, e os outros classificadores identificam a relação "Outra", opta-se pela relação *Overlap*, por ter maior confiabilidade em sua identificação.

Pode ocorrer que nenhum dos classificadores binários identifique alguma relação (diferente de "Outra") para dado par de sentenças em análise. Nesse caso, um classificador multiclasse, para as mesmas relações consideradas nesse experimento, é utilizado para

identificar alguma relação para o par de sentenças. Por exemplo, se todos os oito classificadores binários optarem pela relação "Outra" (ou seja, nenhuma relação CST foi identificada), um classificador multiclasse definirá uma relação CST para o par de sentenças.

Existem outras formas de adaptar um problema multiclasse em um problema binário. Essas outras formas variam em como combinar os rótulos, sempre de dois em dois. Por exemplo, ao invés de combinar um com todos os outros rótulos agrupados, pode-se combinar os rótulos do problema de dois em dois.

Os resultados de cada classificador binário (utilizando o algoritmo J48) estão na Tabela 17. As medidas-F médias ficaram acima de 65%, indicando que a abordagem é promissora, na identificação das relações mais frequentes do córpus. Obviamente, ao criar classificadores binários, a classe "Outra" (que contém todas as outras relações do córpus, exceto a que se objetiva identificar no classificador binário) torna-se majoritária (isto é, com maior quantidade de exemplos) e obtém resultados maiores, mas, mesmo assim, (exceto para a relação *Contradiction*) os valores de medida-F para as relações minoritárias de cada classificador ficaram acima de 48%. Por exemplo, o classificador da relação *Attribution* identifica a relação com uma precisão de 60%. Embora a relação "Outra" seja identificada com mais de 90% de precisão, a identificação da relação *Attribution* é maior que no classificador multiclasse para todas as relações CST.

Na Tabela 18, vemos que as técnicas J48 e SVM obtiveram resultados bem próximos para a maioria das relações. Como o algoritmo J48 obteve melhores resultados para os outros cenários e nesse esteve bem próximo ao SVM, optou-se por criar os classificadores binários com J48.

Quando utilizados pelo *parser*, cada classificador assimilará um rótulo a cada par de sentenças. Na escolha da classe, será necessário apenas escolher entre os rótulos que não sejam o rótulo "Outra". Como critério de seleção, será utilizado um valor de confiabilidade dado pelo classificador na escolha da classe, optando-se pelo maior.

**Tabela 17 - Classificadores Binários**

| **Attribution** | | | | **Follow-up** | | | |
|---|---|---|---|---|---|---|---|
| Precisão | Cobertura | Medida-F | Relação | Precisão | Cobertura | Medida-F | Relação |
| 0,950 | 0,976 | 0,963 | Outra | 0,814 | 0,887 | 0,848 | Outra |
| 0,600 | 0,413 | 0,489 | *Attribution* | 0,668 | 0,529 | 0,590 | *Follow-up* |
| **0,775** | **0,695** | **0,726** | **Media** | **0,741** | **0,708** | **0,719** | **Media** |
| **Contradiction** | | | | **Historical-background** | | | |
| Precisão | Cobertura | Medida-F | Relação | Precisão | Cobertura | Medida-F | Relação |
| 0.957 | 0.994 | 0.976 | Outra | 0.953 | 0.968 | 0.960 | Outra |
| 0.700 | 0.228 | 0.344 | *Contradiction* | 0.608 | 0.513 | 0.556 | *Historical-background* |
| **0.829** | **0.611** | **0.660** | **Media** | **0.781** | **0.741** | **0.758** | **Media** |
| **Elaboration** | | | | **Overlap** | | | |
| Precisão | Cobertura | Medida-F | Relação | Precisão | Cobertura | Medida-F | Relação |
| 0.810 | 0.844 | 0.827 | Outra | 0.759 | 0.763 | 0.761 | Outra |
| 0.677 | 0.622 | 0.648 | *Elaboration* | 0.697 | 0.693 | 0.695 | *Overlap* |
| **0.744** | **0.733** | **0.738** | **Media** | **0.728** | **0.728** | **0.728** | **Media** |
| **Equivalence** | | | | **Subsumption** | | | |
| Precisão | Cobertura | Medida-F | Relação | Precisão | Cobertura | Medida-F | Relação |
| 0.972 | 0.986 | 0.979 | Outra | 0.749 | 0.698 | 0.723 | *Subsumption* |
| 0.593 | 0.410 | 0.485 | *Equivalence* | 0.915 | 0.933 | 0.924 | Outra |
| **0.783** | **0.698** | **0.732** | **Media** | **0.832** | **0.816** | **0.824** | **Media** |

**Tabela 18 - Resultado do teste de significância estatística**

| **Classificador** | **NaiveBayes** | **SVM** | **J48** |
|---|---|---|---|
| *Attribution* | 0.9416 | 0.9458 (e) | 0.9580 (g) |
| *Contradiction* | 0.9066 | 0.9721 (g) | 0.9721 (g) |
| *Elaboration* | 0.5270 | 0.8004 (g) | 0.7637 (g) |
| *Equivalence* | 0.9210 | 0.9752 (g) | 0.9745 (g) |
| *Follow-up* | 0.4713 | 0.8371 (g) | 0.7813 (g) |
| *Historical-background* | 0.7228 | 0.9563 (g) | 0.9548 (g) |
| *Overlap* | 0.5364 | 0.7168 (g) | 0.6879 (g) |
| *Subsumption* | 0.8574 | 0.8752 (g) | 0.8700 (e) |

### 5.2.5.4. Classificadores Hierárquicos para todas as relações

Foram exploradas duas abordagens na criação de classificadores hierárquicos: *top-down* e *big-bang*. Em uma abordagem *top-down*, por exemplo, no primeiro nível, para definir entre "conteúdo" e "forma", um classificador é utilizado. Para definir entre "redundância",

"complemento" ou "contradição", outro classificador é utilizado, e assim para cada ramificação na tipologia. Os resultados de cada classificador são apresentados na Tabela 3. A Figura 17 mostra a localização de cada classificador na hierarquia das relações.

Todas as medidas-F médias tiveram valores acima de 0.45. Isso mostra a potencialidade dessa abordagem na identificação das relações. No entanto, algumas relações ainda tiveram resultados muito baixos, como *Modality*, *Translation* e *Summary*, devido à baixa frequência dessas relações no córpus. Considerando o classificador A: ele decide se o par de sentenças contém uma relação de "conteúdo" (com mais de 0.93 de precisão) ou "forma" (com 0.64 de precisão). Mesmo com o desbalanceamento das relações CST, quando elas são agrupadas, segundo a tipologia apresentada, os resultados ficam melhores.

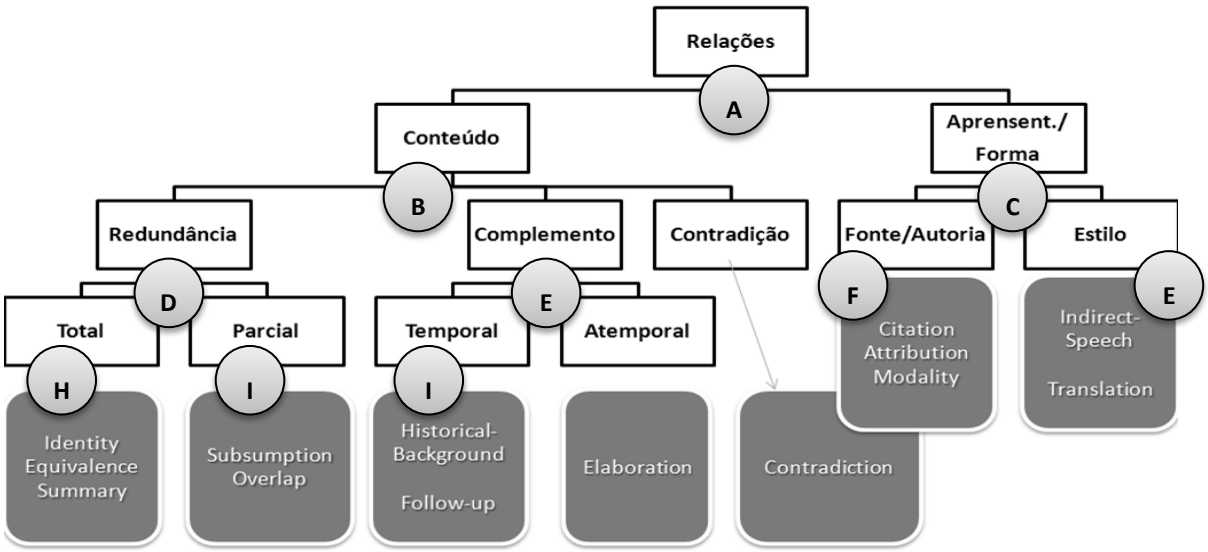


**Figura 17 - Localização dos classificadores na hierarquia**

Pela Tabela 20, vê-se que a técnica J48 obteve melhore resultados que o SVM na maioria dos casos. Portanto, os classificadores foram criados utilizando J48.

**Tabela 19 - Classificadores Hierárquicos**

| **Classificador A (Conteúdo x Forma)** | | | | **Classificador F (Attribution x Modality)** | | | |
|---|---|---|---|---|---|---|---|
| Precisão | Cobertura | Medida-F | Relação | Precisão | Cobertura | Medida-F | Relação |
| 0.933 | 0.974 | 0.953 | Conteúdo | 0.986 | 1 | 0.993 | *Attribution* |
| 0.640 | 0.394 | 0.488 | Forma | 0 | 0 | 0 | *Modality* |
| **0.7865** | **0.684** | **0.7205** | **Media** | **0.493** | **0.500** | **0.496** | **Media** |
| **Classificador B (Redundância x Compl. Contrad.)** | | | | **Classificador G (Indirect Speech x Translation)** | | | |
| Precisão | Cobertura | Medida-F | Relação | Precisão | Cobertura | Medida-F | Relação |
| 0.696 | 0.701 | 0.699 | Redundância | 0.895 | 0.944 | 0.919 | *Indirect Speech* |
| 0.661 | 0.679 | 0.670 | Complemento | 0 | 0 | 0 | *Translation* |
| 0.045 | 0.022 | 0.029 | *Contradiction* | **0.447** | **0.472** | **0.459** | **Media** |
| **0.470** | **0.467** | **0.466** | **Media** | **Classificador H (Identity x Equivalence X Summary)** | | | |
| **Classificador C (Fonte x Estilo)** | | | | Precisão | Cobertura | Medida-F | Relação |
| Precisão | Cobertura | Medida-F | Relação | 0.882 | 0.965 | 0.921 | *Identity* |
| 0.853 | 0.829 | 0.841 | Fonte | 0.800 | 0.718 | 0.757 | *Equivalence* |
| 0.400 | 0.444 | 0.421 | Estilo | 0 | 0 | 0 | *Summary* |
| **0.626** | **0.636** | **0.631** | **Media** | **0.561** | **0.561** | **0.559** | **Media** |
| **Classificador D (Parcial x Total)** | | | | **Classificador I (Subsumption x Overlap)** | | | |
| Precisão | Cobertura | Medida-F | Relação | Precisão | Cobertura | Medida-F | Relação |
| 0.953 | 0.985 | 0.969 | Parcial | 0.609 | 0.541 | 0.573 | *Subsumption* |
| 0.905 | 0.742 | 0.815 | Total | 0.806 | 0.846 | 0.825 | *Overlap* |
| **0.929** | **0.864** | **0.892** | **Media** | **0.707** | **0.693** | **0.699** | **Media** |
| **Classificador E (Atemporal x Temporal)** | | | | **Classificador J (Follow-up x Historical-background)** | | | |
| Precisão | Cobertura | Medida-F | Relação | Precisão | Cobertura | Medida-F | Relação |
| 0.851 | 0.918 | 0.884 | Atemporal | 0.871 | 0.922 | 0.896 | *Follow-up* |
| 0.440 | 0.286 | 0.346 | Temporal | 0.617 | 0.481 | 0.540 | *Historical-background* |
| **0.645** | **0.602** | **0.615** | **Media** | **0.744** | **0.7015** | **0.718** | **Media** |

**Tabela 20 - Resultado do teste de significância estatística**

| **Classificador** | **NaiveBayes** | **SVM** | **J48** |
|---|---|---|---|
| A | 0.9251 | 0.9299 (e) | 0.9440 (g) |
| B | 0.6291 | 0.6047 (e) | 0.6657 (g) |
| C | 0.7738 | 0.7858 (e) | 0.7835 (g) |
| D | 0.9494 | 0.8770 (p) | 0.9502 (e) |
| E | 0.8231 | 0.8276 (e) | 0.8119 (e) |
| F | 0.9857 | 0.9857 (e) | 0.9857 (e) |
| G | 0.9000 | 0.9000 (e) | 0.8450 (e) |
| H | 0.8916 | 0.6643 (p) | 0.8544 (e) |
| I | 0.7665 | 0.7199 (p) | 0.7616 (e) |
| J | 0.8046 | 0.8114 (e) | 0.8284 (e) |

Numa classificação hierárquica, utilizando a abordagem *top-down*, os classificadores serão utilizados sequencialmente de acordo com a escolha dos classificadores mais altos na hierarquia. Assim, o primeiro classificador (classificador A) escolherá entre "conteúdo" e "forma". Essa escolha definirá os próximos classificadores a serem utilizados. Se o rótulo escolhido for "forma", o classificador C será utilizado, escolhendo entre "fonte" e "estilo". Caso "estilo" seja escolhido, o classificador E será utilizado, escolhendo entre *Indirect Speech* e *Translation*. Como se chegou ao nível mais inferior da hierarquia, termina-se o processo, com a escolha de uma relação CST. Utilizando uma avaliação cruzada de 10 pastas, os classificadores foram combinados, como acima descrito, obtendo-se 0.351 de acurácia.

O classificador segundo a abordagem *big-bang* obteve uma porcentagem de acerto de 0.587. Esse classificador utiliza, quando mais vantajoso, rótulos internos da hierarquia, como classe escolhida. Assim, como se deseja atingir os rótulos mais inferiores da tipologia das relações, tentou-se aumentar a especificação do modelo aprendido alterando-se o valor de uma constante utilizada pelo algoritmo no cálculo da entropia (Clare, 2003) na geração do modelo. Mesmo aumentando o valor dessa constante, a especificação total não foi alcançada e o modelo não chega a todas as folhas da tipologia e o desempenho obtido, portanto, não corresponde a chegar a todas as folhas da hierarquia, podendo o algoritmo parar em algum nó interno da estrutura.

### 5.2.6. Resolvendo desbalanceamento e sobreposição de classes

Dado o conhecimento obtido nos experimentos anteriores, a fim de resolver o problema do desbalanceamento e sobreposição das classes, foram consideradas, nos experimentos seguintes, seis relações, a saber: *Equivalence*, *Subsumption*, *Overlap*, *Historical-background*, *Follow-up* e *Elaboration*. As demais relações foram desconsideradas, dentre essas, as relações *Summary*, *Modality* e um tipo de *Contradiction*. As relações *Identity*, um tipo de *Contradiction*, *Citation*, *Attribution*, *Indirect Speech* e *Translation* serão identificadas via regras, reportadas mais à frente.

Para cada um dos três experimentos apresentados a seguir, os classificadores foram criados a partir dos dados sem balanceamento (agora considerando apenas seis classes) e

dos dados balanceados pela replicação dos exemplos das relações com menos frequência até atingir numero igual ou próximo ao da relação de maior frequência, nesse caso, a relação *Overlap*, que contém 467 exemplos no córpus. A Figura 18 mostra a frequência de cada relação escolhida para compor os classificadores apresentados.

A Figura 19 apresenta as frequências após o balanceamento das relações. Como relações com poucos exemplos não foram consideradas, o balanceamento por replicação de exemplos se mostrou viável. Embora a relação *Equivalence* tenha poucos exemplos, comparando-se com as outras cinco relações escolhidas, acredita-se que os atributos escolhidos reconheçam bem essa relação e, por esse motivo, sua inclusão nesses experimentos foi considerada.

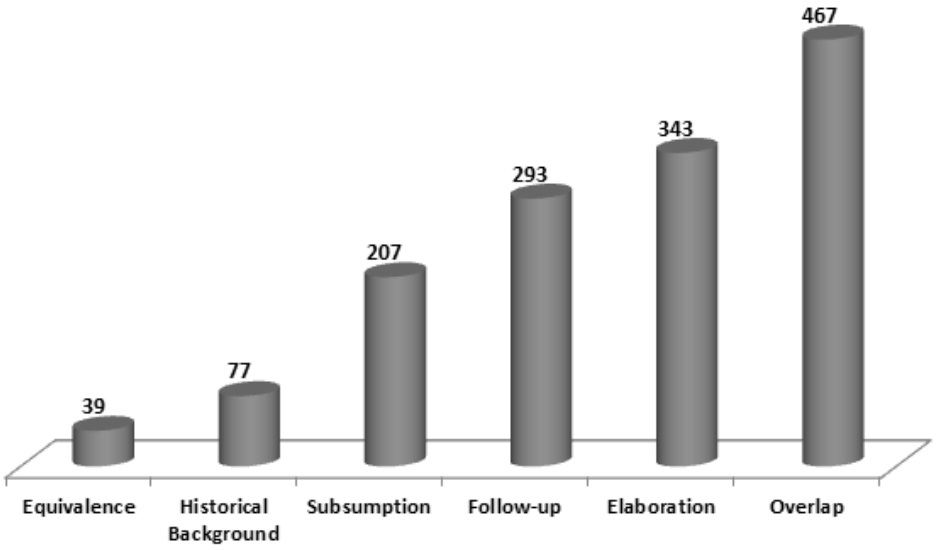


**Figura 18 - Frequência das relações de "conteúdo" escolhidas**

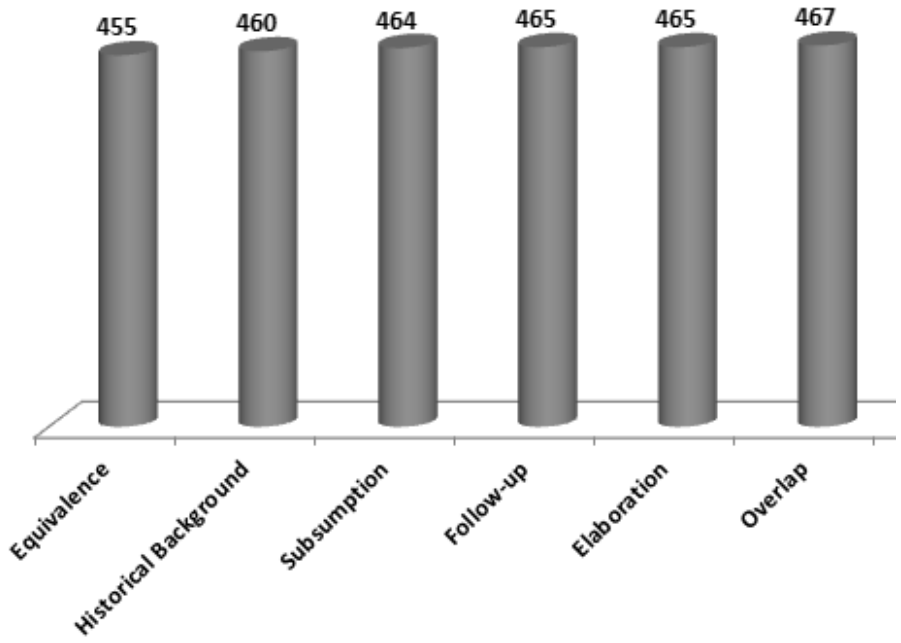


**Figura 19 - Frequência das relações escolhidas após balanceamento dos dados**

### 5.2.6.1. Classificador multiclasse para algumas relações de "conteúdo"

A exemplo do classificador multiclasse citado anteriormente, que considera apenas as relações de "conteúdo", o classificador aqui descrito considera apenas algumas das relações de "conteúdo". As relações *Identity*, *Summary* e *Contradiction* não foram consideradas: a primeira devido a uma regra manual que identifica a relação perfeitamente; a segunda devido a sua baixa frequência no córpus; e a terceira, além de ter baixa frequência, será tratada parcialmente por regras. A retirada dessas três relações nesse cenário diminui a confusão na classificação e aumenta o desempenho do classificador.

**Tabela 21 - Resultados do classificador multiclasse para algumas relações de conteúdo com dados não balanceados**

| Relação | Precisão | Cobertura | Medida-F |
|---|---|---|---|
| *Elaboration* | 0.405 | 0.385 | 0.395 |
| *Overlap* | 0.441 | 0.478 | 0.458 |
| *Follow-up* | 0.282 | 0.273 | 0.277 |
| *Historical-background* | 0.299 | 0.260 | 0.278 |
| *Subsumption* | 0.449 | 0.447 | 0.448 |
| *Equivalence* | 0.378 | 0.359 | 0.368 |
| **Média** | **0.376** | **0.367** | **0.371** |

**Tabela 22 - Matriz de confusão para o classificador multiclasse com dados não balanceados**

| Subsumption | Elaboration | Overlap | Follow-up | Hist.-back. | Equivalence | |
|---|---|---|---|---|---|---|
| 108 | 25 | 45 | 16 | 5 | 8 | Subsumption |
| 24 | 136 | 96 | 70 | 16 | 1 | Elaboration |
| 50 | 86 | 242 | 67 | 17 | 5 | Overlap |
| 28 | 80 | 93 | 74 | 15 | 3 | Follow-up |
| 8 | 22 | 16 | 16 | 15 | 0 | Historical-backgroud |
| 10 | 2 | 13 | 3 | 0 | 11 | Equivalence |

Pela matriz de confusão para esse classificador (Tabela 22), as relações *Follow-up* e *Equivalence* são classificadas erroneamente na maioria dos casos como relação *Overlap*. A relação *Historical-background* é confundida, na maioria das vezes com a relação *Elaboration*. Um dos motivos para isso são os atributos utilizados, que podem não distinguir bem entre as relações citadas.

Para os outros classificadores não são apresentadas as matrizes de confusão, pois tanto os classificadores binários quanto os hierárquicos não apresentam, cada classificador, mais do que duas relações.

Verifica-se que os resultados para os dados balanceados ficaram bem melhores que para os dados não balanceados. Essa diferença nos resultados entre a Tabela 21 e a Tabela 23 deve-se ao *overfitting* sobre os dados replicados durante o desbalanceamento. Por exemplo, a medida-F para a relação *Historical-background* aumentou de 0.278 para 0.932.

**Tabela 23 - Resultados do classificador multiclasse para algumas relações de conteúdo com dados balanceados**

| Relação | Precisão | Cobertura | Medida-F |
|---|---|---|---|
| *Overlap* | 0.496 | 0.429 | 0.460 |
| *Elaboration* | 0.625 | 0.523 | 0.569 |
| *Follow-up* | 0.610 | 0.615 | 0.612 |
| *Subsumption* | 0.750 | 0.815 | 0.781 |
| *Historical-background* | 0.873 | 1.000 | 0.932 |
| *Equivalence* | 0.942 | 1.000 | 0.970 |
| **Média** | **0.716** | **0.730** | **0.721** |

O classificador foi gerado pelo algoritmo J48, por ter se apresentado melhor na maioria dos testes realizados. O uso do classificador treinado nos dados sem balanceamento será utilizado no *parser* final, pois reflete melhor o cenário real da tarefa.

### 5.2.6.2. Classificadores binários

Para as mesmas seis relações dos classificadores anteriores, seis classificadores binários foram criados, segundo a mesma abordagem já explicada, considerando para cada classificador uma relação CST, rotulando as demais como "Outra".

**Tabela 24 - Resultados dos classificadores binários para algumas relações de conteúdo com dados não balanceados**

| **Elaboration** | | | | **Historical-background** | | | |
|---|---|---|---|---|---|---|---|
| **Precisão** | **Cobertura** | **Medida-F** | **Relação** | **Precisão** | **Cobertura** | **Medida-F** | **Relação** |
| 0.380 | 0.347 | 0.363 | *Elaboration* | 0.953 | 0.991 | 0.972 | *Outra* |
| 0.799 | 0.821 | 0.809 | *Outra* | 0.478 | 0.143 | 0.220 | *Hist. Background* |
| **0.5895** | **0.584** | **0.586** | **Média** | **0.7155** | **0.567** | **0.596** | **Média** |
| **Equivalence** | | | | **Overlap** | | | |
| **Precisão** | **Cobertura** | **Medida-F** | **Relação** | **Precisão** | **Cobertura** | **Medida-F** | **Relação** |
| 0.977 | 0.991 | 0.984 | *Outra* | 0.743 | 0.773 | 0.758 | *Outra* |
| 0.368 | 0.179 | 0.241 | *Equivalence* | 0.493 | 0.452 | 0.472 | *Overlap* |
| **0.6725** | **0.585** | **0.6125** | **Média** | **0.618** | **0.6125** | **0.615** | **Média** |
| **Follow-up** | | | | **Subsumption** | | | |
| **Precisão** | **Cobertura** | **Medida-F** | **Relação** | **Precisão** | **Cobertura** | **Medida-F** | **Relação** |
| 0.818 | 0.880 | 0.848 | *Outra* | 0.900 | 0.930 | 0.915 | *Outra* |
| 0.346 | 0.246 | 0.287 | *Follow-up* | 0.485 | 0.388 | 0.431 | *Subsumption* |
| **0.582** | **0.563** | **0.5675** | **Média** | **0.6925** | **0.659** | **0.673** | **Média** |

Verificou-se que os resultados (utilizando o algoritmo J48) também aumentaram para os dados balanceados, possivelmente devido ao *overfitting* aos dados. Uma boa evidência é que relações com menos frequência nos dados não balanceados obtiveram melhores resultados nos dados balanceados que as relações que eram mais frequentes nos dados não balanceados e que sofreram, portanto, menos replicação de exemplos. Como exemplo, a relação *Elaboration*, uma das mais frequentes, teve sua medida-F aumentada de 0.363 (Tabela 24, para dados não balanceados) para 0.869 (Tabela 25, para dados balanceados por replicação), enquanto a relação *Equivalence* aumentou de 0.241 (Tabela 24) para 0.985 (Tabela 25), possivelmente causado pelo *overfitting*.

**Tabela 25 - Resultados dos classificadores binários para algumas relações de conteúdo com dados balanceados**

| **Elaboration** | | | | **Historical-background** | | | |
|---|---|---|---|---|---|---|---|
| **Precisão** | **Cobertura** | **Medida-F** | **Relação** | **Precisão** | **Cobertura** | **Medida-F** | **Relação** |
| 0.937 | 0.766 | 0.843 | *Outra* | 1 | 0.951 | 0.975 | *Outra* |
| 0.802 | 0.948 | 0.869 | *Elaboration* | 0.953 | 1 | 0.976 | *Hist. Background* |
| **0.8695** | **0.857** | **0.856** | **Média** | **0.9765** | **0.9755** | **0.9755** | **Média** |
| **Equivalence** | | | | **Overlap** | | | |
| **Precisão** | **Cobertura** | **Medida-F** | **Relação** | **Precisão** | **Cobertura** | **Medida-F** | **Relação** |
| 1 | 0.970 | 0.985 | *Outra* | 0.804 | 0.680 | 0.737 | *Outra* |
| 0.971 | 1 | 0.985 | *Equivalence* | 0.722 | 0.834 | 0.774 | *Overlap* |
| **0.9855** | **0.985** | **0.985** | **Média** | **0.763** | **0.757** | **0.7555** | **Média** |
| **Follow-up** | | | | **Subsumption** | | | |
| **Precisão** | **Cobertura** | **Medida-F** | **Relação** | **Precisão** | **Cobertura** | **Medida-F** | **Relação** |
| 0.983 | 0.748 | 0.849 | *Outra* | 0.997 | 0.884 | 0.937 | *Outra* |
| 0.796 | 0.987 | 0.881 | *Follow-p* | 0.896 | 0.998 | 0.944 | *Subsumption* |
| **0.8895** | **0.8675** | **0.865** | **Média** | **0.9465** | **0.941** | **0.9405** | **Média** |

Nesse cenário, são escolhidos os classificadores binários treinados com os dados não balanceados para compor o *parser* multidocumento.

### 5.2.6.3. Classificadores hierárquicos

Para gerar os classificadores hierárquicos considerando apenas relações de "conteúdo", a hierarquia de relações foi alterada, segundo a Figura 20, a fim de identificar apenas as seis relações citadas. Também foram explorados os dados sem e com balanceamento.

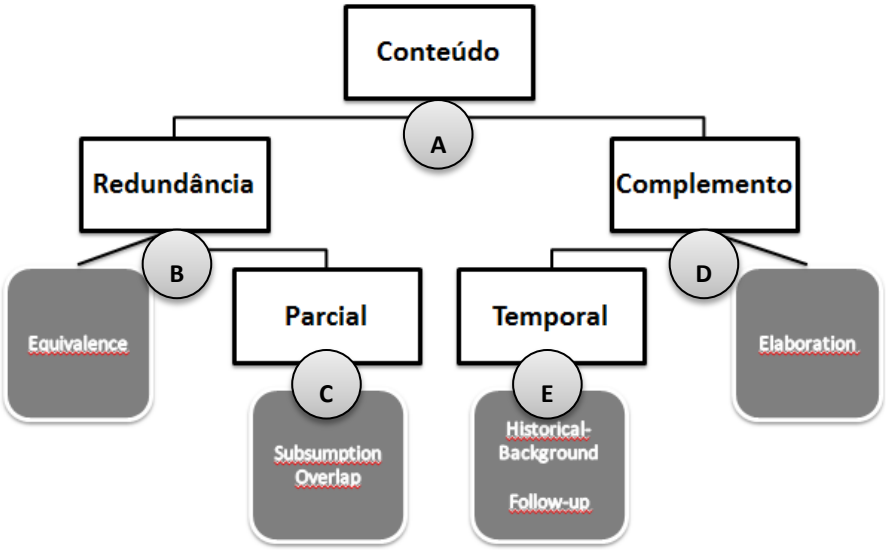


**Figura 20 - Tipologia das relações para algumas relações de "conteúdo"**

**Tabela 26 - Resultados dos classificadores hierárquicos para algumas relações de conteúdo com dados não balanceados**

| **Classificador A (Complemento x Redundância)** | | | | **Classificador D (Elaboration x Temporal)** | | | |
|---|---|---|---|---|---|---|---|
| **Precisão** | **Cobertura** | **Medida-F** | **Relação** | **Precisão** | **Cobertura** | **Medida-F** | **Relação** |
| 0.640 | 0.675 | 0.657 | Complemento | 0.548 | 0.601 | 0.573 | *Elaboration* |
| 0.656 | 0.621 | 0.638 | Redundância | 0.593 | 0.541 | 0.566 | Temporal |
| **0.648** | **0.648** | **0.6475** | **Média** | **0.5705** | **0.571** | **0.5695** | **Média** |
| **Classificador B (Parcial x Equivalence)** | | | | **Classificador E (Follow-up x Historical-background)** | | | |
| **Precisão** | **Cobertura** | **Medida-F** | **Relação** | **Precisão** | **Cobertura** | **Medida-F** | **Relação** |
| 0.958 | 0.988 | 0.973 | Parcial | 0.884 | 0.935 | 0.909 | *Follow-up* |
| 0.556 | 0.256 | 0.351 | *Equivalence* | 0.683 | 0.532 | 0.599 | *Hist. Background* |
| **0.757** | **0.622** | **0.662** | **Média** | **0.7835** | **0.7335** | **0.754** | **Média** |
| **Classificador C (Overlap x Subsumption)** | | | | | | | |
| **Precisão** | **Cobertura** | **Medida-F** | **Relação** | | | | |
| 0.824 | 0.852 | 0.838 | *Overlap* | | | | |
| 0.637 | 0.587 | 0.611 | *Subsumption* | | | | |
| **0.7305** | **0.7195** | **0.7245** | **Média** | | | | |

Aqui, pela Tabela 26 e Tabela 27, chega-se à mesma conclusão sobre os classificadores anteriores, ou seja, os classificadores criados sobre os dados balanceados podem sofrer de *overfitting* e, portanto, os classificadores criados sobre os dados não balanceados serão utilizados. O algoritmo J48 foi utilizado na criação dos classificadores.

**Tabela 27 - Resultados dos classificadores hierárquicos para algumas relações de conteúdo com dados balanceados**

| **Classificador A (Complemento x Redundância)** | | | | **Classificador D (Temporal x Elaboration)** | | | |
|---|---|---|---|---|---|---|---|
| **Precisão** | **Cobertura** | **Medida-F** | **Relação** | **Precisão** | **Cobertura** | **Medida-F** | **Relação** |
| 0.640 | 0.675 | 0.657 | Complemento | 0.607 | 0.599 | 0.603 | Temporal |
| 0.656 | 0.621 | 0.638 | Redundância | 0.603 | 0.611 | 0.607 | *Elaboration* |
| **0.648** | **0.648** | **0.6475** | **Média** | **0.605** | **0.605** | **0.605** | **Média** |
| **Classificador B (Parcial x Equivalence)** | | | | **Classificador E (Follow-up x Historical-background)** | | | |
| **Precisão** | **Cobertura** | **Medida-F** | **Relação** | **Precisão** | **Cobertura** | **Medida-F** | **Relação** |
| 1 | 0.952 | 0.976 | Parcial | 0.996 | 0.822 | 0.901 | *Follow-up* |
| 0.953 | 1 | 0.976 | *Equivalence* | 0.847 | 0.997 | 0.916 | *Hist. Background* |
| **0.9765** | **0.976** | **0.976** | **Média** | **0.9215** | **0.9095** | **0.9085** | **Média** |
| **Classificador C (Overlap x Subsumption)** | | | | | | | |
| **Precisão** | **Cobertura** | **Medida-F** | **Relação** | | | | |
| 0.886 | 0.77 | 0.824 | *Overlap* | | | | |
| 0.796 | 0.901 | 0.845 | *Subsumption* | | | | |
| **0.841** | **0.8355** | **0.8345** | **Média** | | | | |

## 5.2.7. Conclusões sobre os experimentos

Levando em consideração os resultados da Tabela 21 à Tabela 27, o desbalanceamento das relações CST é algo natural e o balanceamento dos dados por replicação (visto que o balanceamento por deleção também não se mostra viável, dada a baixa frequência da relação *Equivalence*, por exemplo) causou, possivelmente, o *overfitting* dos classificadores aos dados de treinamento.

A Tabela 28 apresenta as acurácias gerais dos classificadores nos primeiros experimentos, em que se exploraram todas as relações CST em alguns casos e, em outros, apenas as relações de "conteúdo". Já a Tabela 29 apresenta as acurácias gerais dos

classificadores para apenas o rol de seis relações (*Elaboration*, *Overlap*, *Subsumption*, *Equivalence, Historical-background* e *Follow-up*), sendo esses dados balanceados e não balanceados.

Considerou-se como método baseline a escolha da classe majoritária do córpus para todo par de sentenças. Um classificador que sempre indicasse a relação *Overlap* (classe majoritária) obteria os valores de 28%, 7%, 17% e 33% de acurácia (tabelas 27 e 28), respectivos aos dados não balanceados e balanceados, considerando todas as relações CST, e dados não balanceados e balanceados, considerando apenas as seis relações escolhidas para os segundos experimentos.

**Tabela 28 - Comparação entre as abordagens e técnicas dos experimentos realizados**

| Estratégia | Técnica | | |
|---|---|---|---|
| | **NB** | **SVM** | **J48** |
| **Sem balanceamento** | | | |
| Multiclasse todas relações | 0.3858 | 0.3402 | 0.3700 |
| Multiclasse relações “conteúdo” | 0.4161 | 0.3904 | 0.3970 |
| Multirótulo | 0.3274 | 0.3495 | 0.3765 |
| Binário | - | - | 0.2750 |
| Hierárquico top-down | - | - | 0.3530 |
| Hierárquico big-bang | - | - | 0.5870 |
| Baseline | 0.2829 | | |
| **Com balanceamento** | | | |
| Multiclasse todas relações | 0.5157 | 0.5109 | 0.8373 |
| Multiclasse relações “conteúdo” | 0.4506 | 0.4789 | 0.7252 |
| Multrotulo | - | - | - |
| Binário | - | - | 0.7243 |
| Hierárquico top-down | - | - | 0.7560 |
| Hierárquico big-bang | - | - | 0.7600 |
| Baseline | 0.0769 | | |

**Tabela 29 - Comparação entre as abordagens e técnicas dos experimentos realizados apenas para algumas relações de "conteúdo"**

| Estratégia | Técnica | | |
|---|---|---|---|
| | NB | SVM | J48 |
| **Sem balanceamento** | | | |
| Multiclasse relações "conteúdo" | 0.3906 | 0.4158 | 0.4109 |
| Binário | - | - | 0.7051 |
| Hierárquico top-down | - | - | 0.4270 |
| Hierárquico big-bang | - | - | 0.6150 |
| Baseline | 0.1667 | | |
| **Com balanceamento** | | | |
| Multiclasse relações "conteúdo" | 0.4525 | 0.4804 | 0.7287 |
| Binário | - | - | 0.6028 |
| Hierárquico top-down | - | - | 0.7416 |
| Hierárquico big-bang | - | - | 0.7070 |
| Baseline | 0.3274 | | |

Comparando-se com os valores médios das medidas-F obtidos por Zhang et al. (2003) e Zhang e Radev (2004), respectivamente 43%, 23% e 30% para precisão, cobertura e medida-F, os classificadores explorados neste documento tiveram valores acima desses últimos, mostrando a potencialidade do *parser* implementado. Vale salientar que são resultados obtidos para línguas e córpus de treinamento diferentes, mas essa comparação pode indicar possíveis caminhos para se melhorar os resultados na área.

Os resultados mostram também que classificação hierárquica das relações multidocumento é promissora. Além disso, fica evidenciado que a própria hierarquia proposta neste trabalho é válida, pois tem um impacto na qualidade dos classificadores existentes até então.

Para resolver o problema multirótulo, em que um mesmo par de sentenças terá uma relação de "conteúdo" e uma de "forma", a seguinte abordagem será utilizada: primeiramente as relações de "conteúdo" serão identificadas por algum dos classificadores bem sucedidos apresentados acima; posteriormente, as relações de "forma" serão identificadas por meio das regras.

O córpus CSTNews é o principal limitante na abordagem apresentada nessa seção, dado o desbalanceamento das classes que o compõem. Vale salientar, que, embora não desejável para a tarefa, o desbalanceamento, assim como a sobreposição das classes é algo natural nesse tipo de análise multidocumento. Portanto, o tratamento dessas questões foram consideradas.

## 5.3 Regras

As relações CST que são abordadas nesta subseção não apresentam frequência suficiente (no córpus utilizado) para a criação de classificadores, mas são passíveis de serem identificadas por regras definidas manualmente. As regras são aplicadas para todo par de sentenças que são analisados via classificadores.

A relação *Identity* indica uma igualdade total entre duas sentenças. As relações *Attribution* e *Citation* são semelhantes em sua definição: tratam da autoria de informação redundante entre sentenças, sendo que a segunda dá autoria a outro documento e a primeira a qualquer outra fonte que não seja um documento do conjunto de documentos analisados. A identificação de discurso direto e indireto, objetivo da relação *Indirect Speech*, é obtido a partir da identificação de padrões que indiquem a forma de discurso nas duas sentenças em análise. A relação *Contradiction* explícita consiste em verificar divergências numéricas entre sentenças que apresentam quase a mesma informação (diferente por informações numéricas). Por fim, a relação *Translation* é identificada através da verificação de mesmas informações compartilhadas entre sentenças, mas em línguas diferentes.

As regras foram obtidas a partir da análise manual dos exemplos extraídos do córpus. A análise foi feita para cada relação em particular. A relação *Identity*, no entanto, não necessitou de análise, visto a facilidade de tratamento da mesma: um casamento completo das sentenças. Esse processo, conhecido como engenharia de conhecimento, foi avaliado medindo-se a cobertura sobre os exemplos do córpus.

A relação *Translation*, por exemplo, tem apenas dois exemplos no córpus, mas pela definição da relação sua regra pode ser desenvolvida. Para a relação *Indirect-speech* a quantidade de exemplos permitiu uma análise mais instanciada e gerou diversas sub-regras, refletindo a diversidade dos exemplos da relação.

A seguir, as regras desenvolvidas são descritas e exemplificadas utilizando exemplos do córpus, por fim os resultados obtidos são apresentados e comentados.

### 5.3.1. *Identity*

A relação *Identity* é a mais simples de identificar, por se tratar de uma simples verificação de igualdade entre duas sentenças. Assim, no exemplo abaixo basta fazer a verificação S1 = S2. Caso afirmativo, a relação *Identity* é estabelecida entre as sentenças.

(S1) *As vítimas do acidente foram 14 passageiros e três membros da tripulação.*

(S2) *As vítimas do acidente foram 14 passageiros e três membros da tripulação.*

Essa regra é a primeira a ser aplicada no *parser* multidocumento, poupando esforços na verificação de outras relações entre o par de sentenças que são idênticas.

**Esquema da regra *Identity*:**

**Se** S1 igual a S2

**Regra** = *Identity*

*S1 e S2 representam as sentenças S1 e S2, respectivamente*

### 5.3.2. *Indirect Speech*, *Attribution* e *Citation*

Essas três relações são identificadas de forma bem similar, pequenas diferenças definem a relação a ser escolhida. Na relação *Attribution*, uma sentença atribui o conteúdo compartilhado a uma fonte qualquer, que não está contida na segunda sentença. Já a relação *Citation*, atribui um conteúdo a outro documento, que é o documento da sentença à

qual a atual sentença se relaciona. *Indirect Speech* é definida quando as sentenças apresentam sobreposição de informação e esta é apresentada em discurso indireto em uma das sentenças e em discurso direto na outra.

A similaridade entre as três relações consiste em que uma das sentenças atribui uma fala a um autor/fonte. Desta forma, uma mesma regra faz a diferenciação entre as relações.

Primeiro, verifica-se em S1 a presença de alguns padrões: presença de verbo de atribuição[10] seguido da palavra "que"; conjunção conformativa ("conforme", "para", "segundo", "de acordo") seguido de nome próprio ou pronome ou artigo definido; pontuação seguido de verbo de atribuição. Em seguida, verifica-se na sentença S2, se há verbo na primeira pessoa, tanto no singular quanto no plural.

Se houve sucesso na busca em S1 e em S2, tem-se a relação *Indirect Speech* de S1 para S2. Caso haja sucesso apenas em S1 e não em S2, tem-se a relação *Attribution*. Visto que essas relações têm direcionalidade, o mesmo processo é executado invertendo-se S1 e S2, a fim de identificar as mesmas relações com direcionalidade inversa.

Abaixo um exemplo da relação *Attribution* entre as sentenças S1 e S2, em que S1 atribui a informação compartilhada (*acidente aéreo*) pelas duas sentenças a uma fonte/autoria (*porta-voz das Nações Unidas*)

(S1) *Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo, matou 17 pessoas na quinta-feira à tarde, informou hoje um porta-voz das Nações Unidas.*

(S2) *Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.*

No exemplo abaixo, S1 apresenta uma fala do presidente em discurso indireto e S2, a mesma fala em discurso direto.

[10] Verbos que atribuem uma fala a um autor, por exemplo, os verbos *dizer*, *anunciar* e *retificar*.

(S1) *O presidente Luiz Inácio Lula da Silva afirmou nesta segunda-feira, durante o programa de rádio "Café com o Presidente", que vai anunciar obras de infra-estrutura que transformarão o Brasil em um "verdadeiro canteiro de obras".*

(S2) - *O dado concreto é que nós vamos fazer deste País um verdadeiro canteiro de obra em se tratando de infra-estrutura - disse o presidente.*

**Esquema da regra *Indirect Speech, Attribution*:**

**Em sentença S1**

Se contiver [verbo de atribuição] seguido de [“que”]

Indicador_S1 = OK

Se contiver [“conforme”, “para”, “segundo”, “de acordo”] seguido de [nome próprio, pronome, artigo definido]

Indicador_S1 = OK

Se contiver [pontuação] seguido de [verbo de atribuição]

Indicador_S1 = OK

**Em sentença S2**

Se contiver [verbo na primeira pessoa do singular ou plural]

Indicador_S2 = OK

Se Indicador_S1 = OK e Indicador_S2 = OK

**Regra** = *Indirect Speech de S1 para S2*

Se Indicador_S1 = OK e Indicador _S2 != OK

**Regra** = *Attribution de S1 para S2*

***Invertendo a ordem das sentenças***

**Em sentença S2**

Se contiver [verbo de atribuição] seguido de [“que”]

Indicador _S2 = OK

Se contiver [“conforme”, “para”, “segundo”, “de acordo”] seguido de [nome próprio, pronome, artigo definido]

Indicador _S2 = OK

Se contiver [pontuação] seguido de [verbo de atribuição]

Indicador _S2 = OK

**Em sentença S1**

Se contiver [verbo na primeira pessoa do singular ou plural]

Indicador _S1 = OK

Se Indicador _S1 = OK e Indicador _S2 = OK

**Regra** = *Indirect Speech de S2 para S1*

Se Indicador _S1 = OK e Indicador _S2 != OK

**Regra** = *Attribution de S2 para S1*

### 5.3.4. *Translation*

A regra básica para identificação dessa relação é verificar se existe algum termo ou expressão que não esteja em Português em uma das sentenças. Caso haja algum termo em outra língua, um tradutor automático é utilizado para verificar a possível tradução do termo ou expressão. Obtida a tradução, verifica-se na outra sentença do par em análise se esta contém a tradução. Se sim, é, então, atribuída a relação *Translation*.

Nessa regra, após obter a tradução de um termo em outra língua, a base de sinônimos TeP 2.0 é utilizada (Maziero et al., 2008) para obter todos os conjuntos de sinônimos (*synsets*) de cada termo traduzido. Assim, ao buscar a tradução em outra sentença, tem-se a relação *Translation* entre as sentenças na ocorrência de qualquer palavra que pertença a um dos *synsets* encontrados.

Abaixo há um exemplo da relação, em que "*Action Contre la Faim"* de S2, é traduzido em S1, como "Ação Contra a Fome", em S1.

(S1) *Quinze voluntários da ONG francesa Ação Contra a Fome (ACF) foram assassinados no nordeste do Sri Lanka, informou hoje um porta-voz da organização.*

(S2) *Segundo um representante do grupo Action Contre la Faim, os corpos foram encontrados no escritório da organização.*

**Esquema da regra *Translation***

Seja TRAD um vetor de palavras vazio

**Para cada palavra p1 de S1**

Se p1 não for nome próprio

Se não pertence ao Português

Identificar a língua de p1 e obter sua tradução

Adicionar a tradução a TRAD

**Para cada palavra p2 em S2**

Para cada palavra t1 em TRAD

Obter os *synsets* de p2 e t1

Se houver equivalência entre algum dos *synsets* de p2 e t1

**Regra** = *Translation*

***Invertendo a ordem das sentenças***

**Para cada palavra p2 de S2**

Se p2 não for nome próprio

Se não pertence ao Português

Identificar a língua de p2 e obter sua tradução

Adicionar a tradução a TRAD

**Para cada palavra p1 em S1**

Para cada palavra t1 em TRAD

Obter os *synsets* de p1 e t1

Se houver equivalência entre algum dos *synsets* de p1 e t1

**Regra** = *translation*

Foi utilizada um API do Google Translate[11], que, dada uma palavra, retorna a tradução da mesma. Ainda que uma mesma palavra possa ter diversas traduções, esse serviço retorna apenas uma tradução para cada palavra. Não é conhecido o processo de desambiguação lexical aplicada. Essa API também identifica a língua de uma dada palavra ou expressão.

### 5.3.5. *Contradiction* explícita

Na relação *Contradiction* não há pistas tão explicitas quanto nas relações anteriormente mencionadas. No exemplo abaixo, verifica-se a contradição entre "não confirmar as mortes" e "confirmou a morte..." das sentenças S1 e S2, respectivamente.

[11] http://translate.google.com

(S1) *Até o momento, as autoridades do Sri Lanka não confirmaram as mortes ou esclareceram o que acontece na cidade de Muttur.*

(S2) *O diretor da ACF no Sri Lanka, Benoit Miribel, confirmou a morte de seus funcionários e afirmou, comovido, que a ONG "não sofreu uma perda similar em seus mais de 25 anos de existência".*

Na relação em questão, no entanto, verifica-se apenas a contradição existente entre os números das sentenças em análise, como no exemplo apresentado no primeiro exemplo da Introdução desta dissertação (página 1).

**Esquema da regra *Contradiction***

Converter todos os números por extenso das sentenças em dígitos

Seja ER uma lista de padrões quantitativos (por exemplo: #hs, em que # indica 1 ou mais dígitos seguidos das letras hs, o que indica um horário)

**Para cada expressão *e* em ER**

Se i ocorrer na sentença 1, armazenar o texto encontrado em P1

Se i ocorrer na sentença 2, armazenar o texto encontrado em P2

**Para cada valor *p* em P1**

**Para cada valor *q* em P2**

Se *p* e *q* forem do mesmo tipo de padrão

Se *p* diferente de *q*

**Regra** = *Contradiction*

### 5.3.6. Resultados

A regra *Translation* teve cobertura de 0.5, pois no córpus há apenas dois exemplos dessa relação e, para um deles a relação ocorreu pela existência da tradução de "Hezbolá" para "Hezbollah" e o que não foi traduzido pelo tradutor utilizado. Essa relação teve resultado nulo no uso de todos os tipos de classificadores, devido à quantidade de exemplos no córpus.

A regra *Indirect Speech / Attribution / Citation* obteve um bom desempenho: 0.58, desempenho bem acima do obtido utilizando classificadores. Por exemplo, a relação *Indirect Speech* obteve resultado nulo na identificação automática pelo uso de classificadores multiclasse para todas as relações CST e multirótulo.

**Tabela 30 - Resultados das regras**

| Regra | Cobertura | Precisão | Medida-F |
|---|---|---|---|
| *Indirect Speech* / *Attribution* / *Citation* | 0.6322 | 0.5288 | 0.5759 |
| *Translation* | 0.5000 | 0.5000 | 0.5000 |
| *Contradiction* Explícita | 0.1765 | 0.2728 | 0.2143 |
| **Media** | **0.4363** | **0.4339** | **0.4301** |

A regra *Contradiction* explícita teve um desempenho baixo, tendo em vista a dificuldade da tarefa. As regras sofrem com erros do etiquetador morfossintático, que etiquetou algumas palavras erroneamente. Por exemplo, alguns números escritos por extenso foram etiquetados como verbo. Isso impossibilitou a identificação da contradição numérica entre duas sentenças que continham esses numerais.

A Tabela 31 apresenta o suporte para as regras criadas, isto é, o número de exemplos utilizados na elaboração das regras.

Esses resultados foram obtidos nos mesmos pares de sentenças que originaram as regras. Isso foi feito devido à baixa quantidade de exemplos disponíveis, não permitindo separar exemplos distintos para geração e teste das regras.

**Tabela 31 - Suporte para as regras**

| Regra | Suporte |
|---|---|
| *Indirect Speech* / *Attribution* / *Citation* | 78 |
| *Translation* | 2 |
| *Contradiction* Explícita | 17 |
| **Media** | **32.33** |

O *parser* apresentou acurácia geral de 68,13% utilizando os classificadores binários para as relações *Overlap*, *Subsumption*, *Elaboration*, *Equivalence*, *Historical-background* e *Follow-up* e as regras para as relações *Contradiction* explícita, *Attribution*, *Indirect Speech*, *Translation*. O cálculo da acurácia geral foi realizado utilizando a média ponderada da acurácia geral dos classificadores binários (70,51% - Tabela 29) e da medida-F média das regras (43,01% - Tabela 30). A ponderação foi baseada no número de instâncias classificadas pelos classificadores binários (1426 instâncias) e as utilizadas pelas regras (135 instâncias). Esse resultado é considerado bom devido à subjetividade inerente à tarefa de identificar as relações multidocumento.

### 5.3.7. Apresentação da análise

A fim de tornar o *parser* mais acessível, uma interface *web*[12] está disponível. Como passo inicial em sua utilização, o usuário pode escolher por procurar automaticamente os textos na *web* ou enviar os textos que tenha agrupado. A Figura 21 apresenta esse passo inicial.

[12] http://www.nilc.icmc.usp.br/~erick/CSTParser

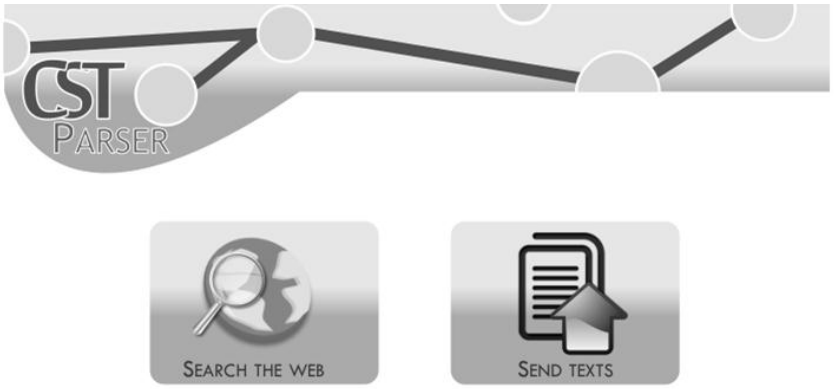


**Figura 21 - Tela inicial do *parser* online**

Caso o usuário opte por uma busca automática dos textos (Figura 22), a interface solicita uma ou mais palavras-chave que serão utilizadas na busca. Neste trabalho, são utilizadas APIs (*Application Programming Interface*) de acesso aos resultados de sistemas de busca como o Bing[13] para obter textos que tratem sobre a consulta do usuário. O serviço do Bing foi escolhido por ser gratuito (serviços como o YahooNews[14] são pagos) e pela API do GoogleNews[15] estar obsoleta (*deprecated*).

[13] http://www.bing.com/toolbox/bingdeveloper/

[14] http://news.yahoo.com

[15] http://code.google.com/intl/pt-BR/apis/newssearch/

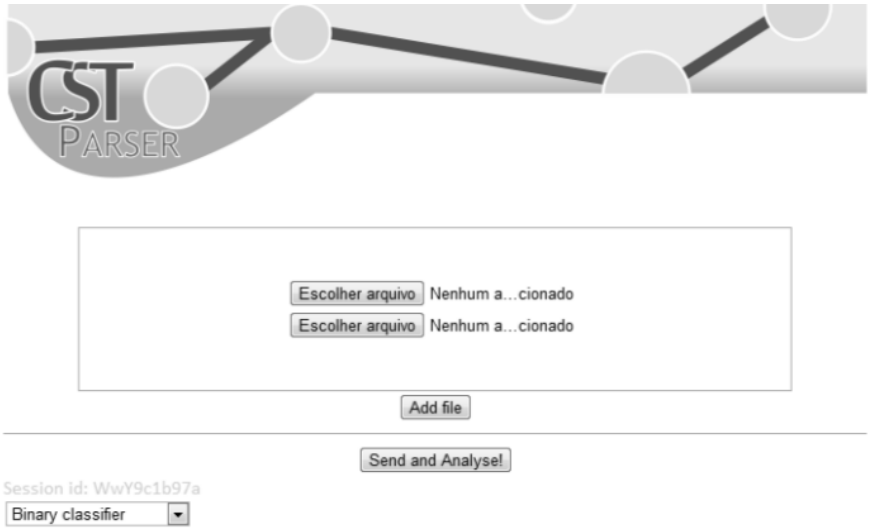


**Figura 22 - Busca e agrupamento automáticos dos textos**

Caso o usuário tenha o grupo de textos e queira submetê-lo à analise, deve carregar cada um dos textos, como mostrado na Figura 23.

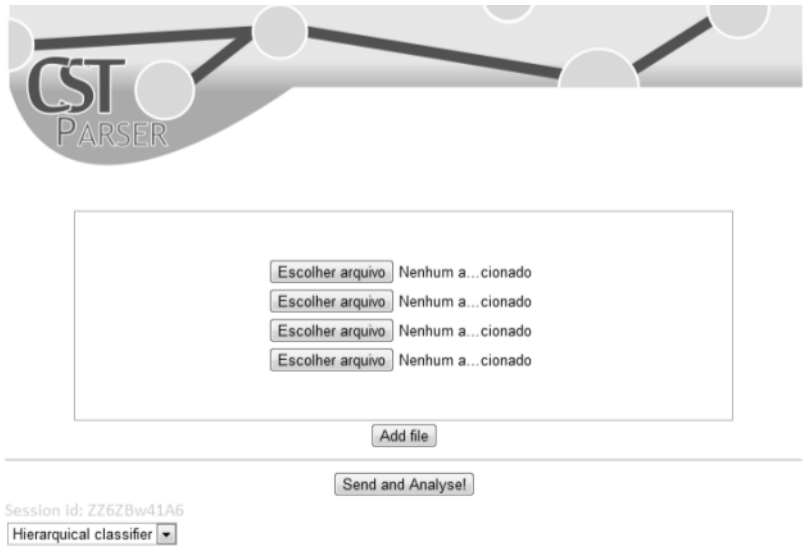


**Figura 23 - Envio manual do grupo de textos**

Passada essa etapa de obtenção do grupo de documentos, seja por busca na *web*, seja por envio manual, procede-se à análise CST automática. Como pode ser visto nas duas figuras

anteriores, o usuário pode escolher o tipo de classificador a ser utilizado: binário, hierárquico ou multiclasse.


**Figura 24 - Tela inicial da apresentação da análise**

Após a aplicação dos classificadores e das regras, o resultado da análise é apresentado ao usuário da interface como na Figura 24. Nessa interface, são apresentados os textos envolvidos na análise e um deles deve ser escolhido para exibição de suas sentenças. A Figura 25 mostra as sentenças do primeiro texto da tela anterior. O usuário pode escolher uma das sentenças para ver as sentenças com as quais há alguma relação CST. No exemplo, a primeira sentença é escolhida e duas sentenças relacionadas com as relações *Overlap* e *Attribution* são exibidas (Figura 26).

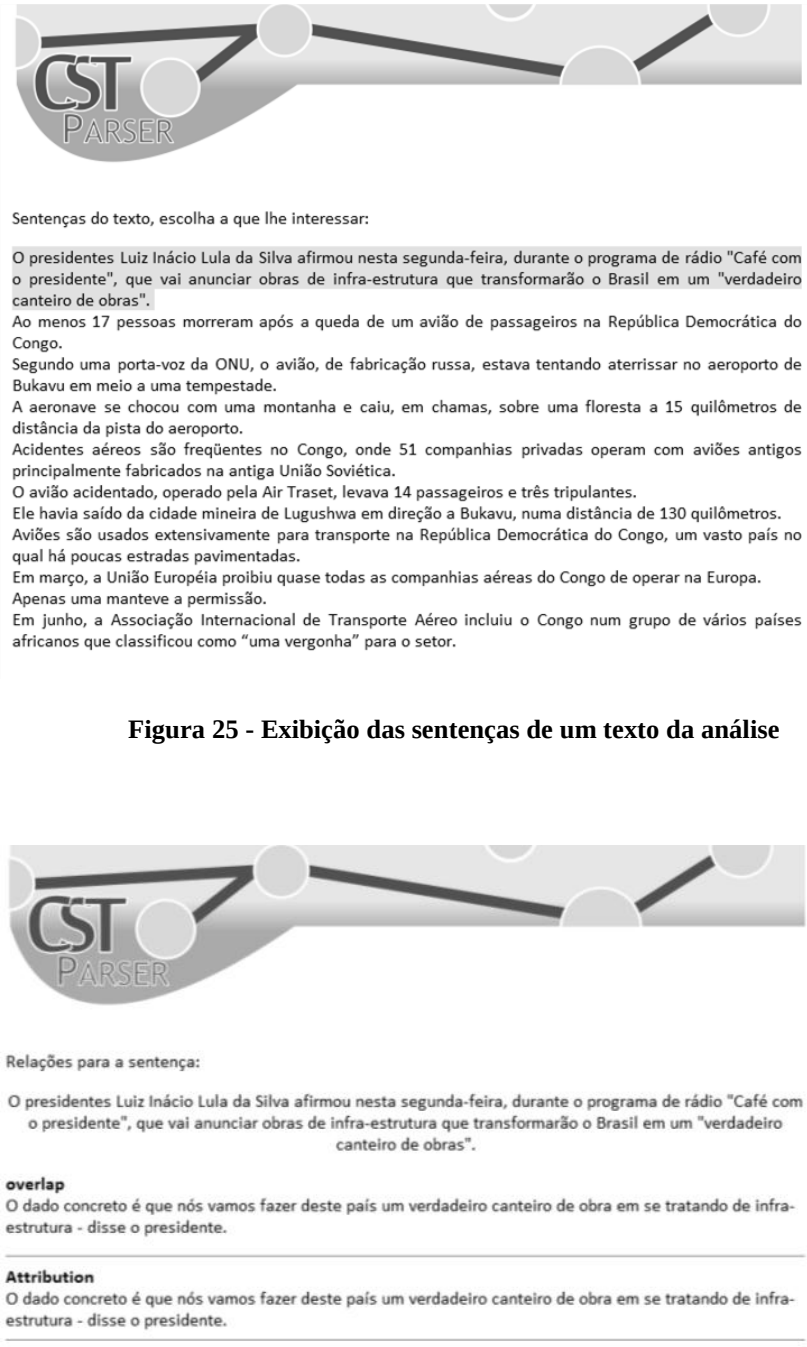


**Figura 25 - Exibição das sentenças de um texto da análise**




**Figura 26 - Relações identificadas para uma das sentenças**

## 5.4. Considerações finais

A metodologia descrita anteriormente automatiza a análise CST, fornecendo a usuários e aplicações uma estruturação semântico-discursiva de um grupo de textos fornecidos.

Vale salientar que a etapa que mais consume tempo é a de extração dos atributos para aplicação dos classificadores. Nessa etapa, um etiquetador e um *parser* morfossintático são utilizados. Essas são as ferramentas cuja execução mais demora em todo o processo de análise CST automática.

As regras mostraram boa cobertura (média de 0.44) nos exemplos do córpus e constituem uma ótima alternativa à falta de frequência de algumas relações. No entanto, o desenvolvimento dessas regras só foi possível para as relações que têm uma especificação bem formal e não necessitam de inferência textual, como a necessária à identificação da relação *Contradiction* por inferência.

Como verificado na literatura, nenhum outro trabalho realizou a identificação de relações CST por meio de regras diretamente sobre as sentenças em análise. Miyabe et al. (2008) utilizam uma abordagem com regras na identificação da relação *Transition* (equivalente à *Contradiction*) a partir da identificação prévia da relação *Equivalence* por classificadores.

# 6. Conclusões e trabalhos futuros

Neste documento, apresentou-se o desenvolvimento de uma metodologia de identificação automática de relações multidocumento aplicada ao Português, segundo a teoria CST. Uma série de modelos multidocumento foi estudada e a que se mostrou mais viável foi a CST, por ser computacionalmente tratável e amplamente utilizada em outros trabalhos que necessitam de tratamento multidocumento.

Como contribuição prática, este trabalho permitirá o tratamento automático de múltiplos documentos em Português, seja por usuários finais, seja por aplicações multidocumento. Vale ressaltar que, no cenário apresentado na introdução deste documento, esse tipo de tratamento (multidocumento) tem se tornado cada vez mais importante e indispensável no processamento automático da língua, principalmente considerando o volume sempre crescente das informações produzidas.

No decorrer deste trabalho, foi gerado um córpus anotado segundo a teoria em questão, o que permitiu abordar a tarefa com o uso de aprendizado automático. Além disso, as relações foram formalizadas e descritas (Anexo A) e foram organizadas em uma hierarquia, que se mostrou adequada, dados os resultados obtidos na identificação automática segundo essa hierarquia (classificadores hierárquicos). Isso confirma, inclusive, a hipótese de que é possível estabelecer uma tipologia genérica de relações multidocumento aplicável a qualquer texto.

Os resultados obtidos confirmaram a hipótese de que a teoria CST é aplicável à língua portuguesa e sua automatização se mostrou viável. Inclusive, os resultados ficaram acima do estado da arte para a língua inglesa. Embora a metodologia descrita nessa dissertação tenha sido aplicada apenas à língua portuguesa, a metodologia pode ser adequada a outras línguas. Para isso seria necessária a anotação de um córpus na língua desejada assim como as ferramentas e recursos da língua para extração dos atributos utilizados pelos classificadores. As regras teriam de ser adequadas com um estudo do córpus criado.

A hipótese de que estratégias híbridas seriam necessárias se confirmou: foram necessários o uso de classificadores e regras. Classificadores para as relações mais frequentes no córpus e regras para as relações menos frequentes.

A principal limitação encontrada foi o desbalanceamento (algo natural para a tarefa) encontrado no córpus utilizado para o aprendizado automático. Essa limitação foi superada para algumas relações por meio da criação de regras. Essas regras obtiveram bons resultados, permitindo que relações bem formalizadas pudessem ser identificadas automaticamente, embora com baixa frequência no córpus.

Inicialmente, realizou-se a investigação de técnicas de aprendizado automático, para geração de classificadores para as relações CST. Caso fossem investigadas, simultaneamente, a criação de classificadores e regras, a arquitetura do *parser* poderia ser diferente. Isso, pois algumas relações foram apenas tratadas por classificadores e, dado o bom desempenho dos mesmos, não foram exploradas regras na identificação dessas relações.

A investigação inicial de técnicas de aprendizado de máquina, também, levou à desconsideração de alguns atributos que poderiam ser bem úteis à classificação. Por exemplo, no estudo do córpus CSTNews sobre os exemplos da relação *Contradiction* para a geração das regras, a desigualdade dos valores dos numerais das sentenças e expressões de negação não foram consideradas como atributos e foram apenas identificados na confecção das regras.

As relações desconsideradas neste trabalho poderão ser abordadas futuramente, aumentando o número de exemplos desses fenômenos multidocumento identificados para a língua portuguesa. A relação *Summary*, por exemplo, pode ser identificada utilizando conhecimentos e medidas da área de sumarização automática, visto que nesta relação objetiva-se identificar se uma sentença é um sumário de outra sentença.

Um aprendizado semissupervisionado pode ser explorado a fim de aumentar a acurácia dos classificadores e, inclusive, tratar as relações pouco frequentes do córpus. Os classificadores atuais podem ser utilizados na rotulação de novos exemplos, que serão utilizados na criação de novos classificadores.

Dado que a abordagem *big-bang* obteve bons resultados na classificação hierárquica, mas lembrando que ele não chega até as folhas da tipologia de relações, a abordagem *top-down* pode ser utilizada para, a partir do ponto em que o classificador

segundo a abordagem *big-bang* pare, continuar até atingir uma das folhas da hierarquia, ou seja, uma relação CST.

Outras abordagens multirótulo podem ser exploradas (Mulan[16]), inclusive juntando a abordagem de aprendizado semissupervisionado, que criará mais exemplos das relações menos frequentes.

O aprimoramento na interface *web*, principalmente na apresentação da análise é algo que possibilitará uma melhor navegação pelas sentenças dos textos e suas relações. Uma possível melhoria está em apresentar ao usuário a sentença que contenha mais relações CST e permitir que, através dela se navegue por outras sentenças escolhendo alguma relação. Essa apresentação, ao invés de ser feita com apenas texto, pode ser implementada com elementos gráficos que tornem mais intuitiva a navegação no grafo gerado na análise CST.

[16] http://mlkd.csd.auth.gr/multilabel.html

# Referências bibliográficas


Afantenos, S.D.; Doura, I.; Kapellou, E.; Karkaletsis, V. (2004). Exploiting Cross-Document Relations for Multi-document Evolving Summarization. In the *Proceedings of SETN*. pp. 410-419.

Aires, R.V.X.; Aluísio, S.M.; Kuhn, D.C.S.; Andreeta, M.L.B.; Oliveira Jr., O.N. (2000). Combining Multiple Classifiers to Improve Part of Speech Tagging: A Case Study for Brazilian Portuguese. In the *Proceedings of the Brazilian AI Symposium*, pp. 20-22.

Aleixo, P. e Pardo, T.A.S. (2008a). *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory)*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 326. São Carlos – São Paulo, Maio. 12p.

Aleixo, P. e Pardo, T.A.S. (2008b). CSTTool: um *parser* multidocumento automático para o Português do Brasil. In the *Proceedings of the IV Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence – WTDIA*. Salvador, Bahia. Outubro, 30.

Aleixo, P. e Pardo, T.A.S. (2008c). Finding Related Sentences in Multiple Documents for Multidocument Discourse Parsing of Brazilian Portuguese Texts. In *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana – TIL*, Vila Velha, Espírito Santo. Outubro. pp. 26-28.

Allan J. (1996). Automatic Hypertext Link Typing. In *Proceedings of the Seventh ACM Conference on Hypertext*, New York-NY. pp. 42-52.

Batista, G. E. A. P. A.; Prati, R. C.; Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, United States of America. Junho. v. 6, n. 1, 20p.

Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis. Aarhus University. Denmark University Press. Outubro. p. 412.

Cardoso, P.C.F.; Maziero, E.G.; Jorge, M.L.C.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011a). CSTNews - A Discourse-Annotated

Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In the *Proceedings of the 3rd RST Brazilian Meeting*, Outubro, Cuiabá – Mato Grosso, pp. 88-105.

Cardoso, P.C.F.; Pardo, T.A.S.; Nunes, M.G.V. (2011b). Métodos para Sumarização Automática Multidocumento Usando Modelos Semântico-Discursivos. In the *Proceedings of the 3rd RST Brazilian Meeting*,. Outubro, Cuiabá - Mato Grosso. pp. 59-74.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: the Kappa Statistic. *Computational Linguistics*, v. 22, no. 2, pp. 249-254.

Chawla, N. V.; Bowyer , K. W.; Hall, L. O.; Kegelmeyer, W. P. (2002). Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. v.16, pp. 321-357.

Cherman, E. A.; M. C. Monard. (2009). Um Estudo sobre Métodos de Classificação Multirrótulo. *Anais do IV Congresso da Academia Trinacional de Ciências*. Foz do Iguaçú – Paraná. pp. 1-10.

Clare, A. (2003). *Machine learning and data mining for yeast functional genomics*. PhD Thesis. University of Wales Aberystwyth. p. 210.

Dagan, I.; Oren G.; Bernardo M. (2005). The pascal recognising textual entailment challenge. In the *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

Freitas, A.; Carvalho, A.C.P.F. (2007). A Tutorial on Hierarchical Classification with Applications in Bioinformatics. In *Research and Trends in Data Mining Technologies and Applications: Advances in Data Warehousing and Mining*. D. Taniar (Editor). v. 1, Idea Group Inc, Hershey, USA. ISBN-10: 159904272X, ISBN-13: 978-1599042725, pp. 176-209.

Freund, Y. e Schapire, R.E. (1997) A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*. v. 55. pp.119–139.

Gantz, J. e Reinsel, D. (2011). Extracting Value from Chãos. *International Data Corporation iView*.

Holte, R.C. (1993). *Very simple classification rules perform well on most commonly used datasets*. Machine Learning. v. 11, pp. 63-91.

John, G. H.; Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo. pp. 338-345.

Jorge, M.L.C.; Agostini, V.; Pardo, T.A.S. (2011). Multi-document Summarization Using Complex and Rich Features. In *Anais do VIII Encontro Nacional de Inteligência Artificial*, Natal, Rio Grande do Norte. Julho. pp. 1-12.

Jijkoun, V. e de Rijke, M. (2005). Recognizing textual entailment using lexical similarity. In the *Proceedings of the First PASCAL Challenges Workshop*.

Landis, J.R.; e Koch, G.G. (1977). "The measurement of observer agreement for categorical data". *Biometrics 33.* pp. 159–174.

Lin, D. (1998). An information-theoretic definition of similarity. In the *Proceedings of International Conference on Machine Learning*.

Mann, W.C. e Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS. pp. 87-190.

Maziero, E.G.; Pardo, T.A.S.; Di Fellipo, A.; Dias da Silva, B.C. (2008). A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In *Anais VI Workshop em Tecnologia da Informação e da Linguagem Humana,* Vila Velha. pp. 390-392.

MacCartney, B.; Trond G.; Marie-Catherine de M.; Daniel C.; Christopher D. M. (2006). Learning to recognize features of valid textual entailments. In the *Proceedings of HLT/NAACL*.

Marsi, E. e Krahmer, E. (2005). Classification of semantic relations by humans and machines. In the *Proceedings of ACL-05 - Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. Ann Arbor-Michigan. Junho. pp. 1-6.

Maziero, E.G.; Jorge, M.L.C.; Pardo, T.A.S. (2010). Identifying Multidocument Relations. In the *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science - NLPCS*, Junho 8-12, Funchal/Madeira, Portugal. pp. 60-69.

Mitchell, T. M. (1997) *Machine Learning*, McGraw-Hill.

Miyabe, Y.; Takamura, H.; Okumura, M. (2008). Identifying a Cross-Document Relation between Sentences. In the *Proceedings of the Third International Joint Conference on Natural Language Processing*, v. 1, pp. 141-148.

Murakami, K.; Nichols, E.; Mizuno, J.; Watanabe, Y.; Goto, H.; Ohki, M.;Matsuyoshi, S.; Inui, K.; Matsumoto, Y. (2010). Automatic Classification of Semantic Relations between Facts and Opinions. In the *Second International Workshop on NLP Challenges* In the *Information Explosion Era* - NLPIX 2010. Beijing, China.

Ohki, M.; Nichols, E.; Matsuyoshi, S.; Murakami, K.; Mizuno, J.; Masuda, S.; Inui, K.; Matsumoto, Y. (2011). Recognizing Confinement in Web Texts. In the *Proceedings of the Ninth International Conference on Computational Semantics*. Oxford – UK. pp. 215-224.

Otterbacher, J.C.; Radev, D.R.; Luo, A. (2002). Revisions that improve cohesion in multi-document summaries: a preliminary study. In the *Proceedings of the Workshop on Automatic Summarization*. Philadelphia. pp. 27-36.

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In the *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistic*s. Philadelphia. pp. 311-318

Pardo, T.A.S. (2006). *SENTER: Um Segmentador Sentencial Automático para o Português do Brasil*. Série de Relatórios do NILC. NILC-TR-06-01. São Carlos – São Paulo, Janeiro, 6p.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA.

Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross-document structure. In the *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*. Hong Kong. pp. 74-83.

Radev, D.R. e McKeown, K. (1998). Generating natural language summaries from multiple online sources. *Computational Linguistics*. v. 24, no. 3, pp. 469-500.

Radev, D.R.; Otterbacher, J.; Zhang, Z. (2004). CST Bank: A Corpus for the Study of Cross-document Structural Relationships. In the *Proceedings of Fourth International Conference on Language Resources and Evaluation*, Lisboa, Portugal.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In the *Proceedings of the First Empirical Methods in NLP Conference*. pp. 133-142.

Salton, G. e Lesk, M.E. (1968). *Computer evaluation of indexing and text processing. Journal of the ACM*. Janeiro. v. 15, no. 1, pp. 8-36.

Trigg, R. (1983). A Network-Based Approach to Text Handling for the Online Scientific Community. PhD. *Thesis. University of Maryland Technical Report, TR-1346*. College Park MD.

Trigg, R.; Weiser, M. (1986). TEXTNET: A Network-Based Approach to Text Handling. In *ACM Transactions on Office Information Systems*. v. 6.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.

Witten, I.H. e Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann. ISBN 978-0-12-374856-0. p. 629.

Zahri, N. e Fukumoto, F. (2011). Multi-document Summarization Using Link Analysis Based on Rhetorical Relations between Sentences. *Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science*. v. 6609, pp. 328-338.

Zhang, Z.; Goldenshon, S.B.; Radev, D.R. (2002). Towards CST-Enhanced Sumarization. In *Proceedings of the 18th National Conference on Artificial Intelligence* (AAAI-2002). Edmonton/Canadá.

Zhang, Z.; Otterbacher, J.; Radev, D.R. (2003). Learning Cross-document Structural Relationships using Boosting. In the *Proceedings of the Twelfth International Conference on Information and knowledge Management*. New York. pp. 124-130.

Zhang, Z. e Radev, D.R. (2005). Combining Labeled and Unlabeled Data for Learning Cross-Document Structural Relationships. In the *Proceedings of IJCNLP 2004*. pp. 32-41.

# ANEXO A – DEFINIÇÃO DAS RELAÇÕES CST

Abaixo são apresentadas as relações consideradas neste trabalho. Para cada relação coloca-se sua posição na tipologia definida, sua direcionalidade, as restrições de aplicação e alguns comentários. Após cada tabela, é mostrado um par de sentenças exemplo relacionadas pela relação apresentada. A relação *Citation* não ocorreu no córpus e, portanto, não dispõe de algum exemplo.

| |
|---|
| **Nome da relação:** *Identity* |
| **Tipo:** Conteúdo->Redundância->Total<br>**Direcionalidade:** Nenhuma<br>**Restrições:** As sentenças devem ser idênticas<br>**Comentários:** |
| S1: *As vítimas do acidente foram 14 passageiros e três membros da tripulação.*<br>S2: *As vítimas do acidente foram 14 passageiros e três membros da tripulação.* |

| |
|---|
| **Nome da relação:** *Equivalence* |
| **Tipo:** Conteúdo->Redundância->Total<br>**Direcionalidade:** Nenhuma<br>**Restrições:** As sentenças apresentam o mesmo conteúdo, mas expresso de forma diferente<br>**Comentários:** |
| S1: *"É um par de irmãos admirável, cada um com cerca de 1% da massa do Sol", disse Jayawardhana.*<br>S2: *"Este é um par de gêmeos verdadeiramente de destaque, já que cada um tem uma massa de apenas 1% de nosso Sol", declarou Jayawardhana.* |

| |
|---|
| **Nome da relação:** *Summary* |

**Tipo:** Conteúdo->Redundância->Total

**Direcionalidade:** S1<-S2

**Restrições:** S2 apresenta o mesmo conteúdo que S1, mas de forma mais compacta.

**Comentários:** *Summary* é um tipo de *equivalence*, mas *summary* deve haver diferença significativa de tamanho entre as sentenças.

---

S1: *Lula disse que o critério para o investimento nas cidades será técnico, não partidário.*

S2: *"O critério é eminentemente técnico, ou seja, eu não quero saber se o prefeito é do PFL, do PT, do PMDB, do PSDB, do PTB, do PR, do PC do B.*

---

**Nome da relação:** *Subsumption*

---

**Tipo:** Conteúdo->Redundância->Parcial

**Direcionalidade:** S1->S2

**Restrições:** S1 apresenta as informações contidas em S2 e informações adicionais.

**Comentários:** S1 contém X e Y, S2 contém X.

---

S1: *Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas.*

S2: *Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.*

---

**Nome da relação:** *Overlap*

---

**Tipo:** Conteúdo->Redundância->Parcial

**Direcionalidade:** Nenhuma

**Restrições:** S1 e S2 apresentam informações em comum e ambas apresentam informações adicionais distintas entre si.

**Comentários:** S1 contém X e Y, S2 contém X e Z.

---

S1: *Se a eleição fosse hoje, o presidente Luiz Inácio Lula da Silva, candidato à reeleição, teria 44% das intenções de voto, contra 25% do tucano Geraldo Alckmin, de acordo com a pesquisa CNI/Ibope divulgada nesta sexta-feira.*

S2: *De acordo com a pesquisa, Lula (PT) tem 44% das intenções de voto, contra 25% de Geraldo Alckmin (PSDB) e 11% de Heloísa Helena (PSOL).*

---

**Nome da relação:** *Historical-background*

---

**Tipo:** Conteúdo->Complemento->Temporal

**Direcionalidade:** S1<-S2

**Restrições:** S2 apresenta informações históricas sobre algum elemento presente em S1.

**Comentários:** O elemento explorado em S2 deve ser o foco de S2; se forem apresentadas informações repetidas, considere outra relação (por exemplo, *overlap*).

---

S1: *Em 1996, uma falha no reverso foi a causa do acidente com o Fokker-100 da TAM, ocorrido segundos depois da decolagem, também em Congonhas.*

S2: *O problema teria sido detectado pelo sistema eletrônico de checagem do próprio avião, e ainda assim a aeronave da TAM, um Airbus A320, continuou voando, com o reverso direito desligado.*

---

**Nome da relação:** *Follow-up*

---

**Tipo:** Conteúdo->Complemento->Temporal

**Direcionalidade:** S1<-S2

**Restrições:** S2 apresenta acontecimentos que acontecem após os acontecimentos em S1; os acontecimentos em S1 e em S2 devem ser relacionados e ter um espaço de tempo relativamente curto entre si.

**Comentários:**

---

S1: *Na hipótese de um segundo turno com a candidata Heloísa Helena, Lula também teve uma redução nas intenções de voto de 57% para 53%.*

S2: *Em junho, Lula tinha 57% e Heloísa 21%.*

---

**Nome da relação:** *Elaboration*

---

**Tipo:** Conteúdo->Complemento->Atemporal

**Direcionalidade:** S1<-S2

**Restrições:** S2 detalha/refina/elabora algum elemento presente em S1, sendo que S2 não deve repetir informações presentes em S1.

**Comentários:** O elemento elaborado em S2 deve ser o foco de S2; se forem apresentadas informações repetidas, considere outra relação (por exemplo, *overlap*); se forem apresentadas informações temporais, pondere sobre a relação *historical-background*.

---

S1: *O maior corpo celeste, com uma massa sete vezes maior do que a de Júpiter, foi detectado a cerca de 400 anos-luz de nosso sistema solar.*

S2: *A lista dos chamados exoplanetas, mundos localizados fora do sistema solar, têm um extraordinário novo membro.*

---

**Nome da relação:** *Contradiction*

---

**Tipo:** Conteúdo->Contradição

**Direcionalidade:** Nenhuma

**Restrições:** S1 e S2 divergem sobre algum elemento das sentenças.

**Comentários:**

---

S1: *O prédio da secretaria da Fazenda, no centro, foi atingido por três bombas caseiras.*

S2: *A Secretaria da Fazenda também foi atingida por uma bomba.*

---

**Nome da relação:** *Citation*

---

**Tipo:** Apresentação/Forma->Fonte/Autoria

**Direcionalidade:** S1<-S2

**Restrições:** S2 cita explicitamente informação proveniente de S1.

**Comentários:** Dada a natureza desta relação, ela não pode co-ocorrer com relações de redundância total.

---

**Nome da relação:** *Attribution*

---

**Tipo:** Apresentação/Forma->Fonte/Autoria

**Direcionalidade:** S1<-S2

**Restrições:** S1 e S2 apresentam informação em comum e S2 atribui essa informação a uma fonte/autoria.

**Comentários:** Dada a natureza desta relação, ela não pode co-ocorrer com relações de redundância total.

---

S1: *Fontes policiais e médicas informaram anteriormente que pelo menos 80 pessoas tinham morrido no acidente.*

S2: *Uma colisão entre dois trens de passageiros provocou a morte de pelo menos 80 pessoas e deixou 165 feridas.*

---

**Nome da relação:** *Modality*

---

**Tipo:** Apresentação/Forma->Fonte/Autoria

**Direcionalidade:** S1<-S2

**Restrições:** S1 e S2 apresentam informação em comum e em S2 a fonte/autoria da informação é indeterminada/relativizada/amenizada

**Comentários:** Dada a natureza desta relação, ela não pode co-ocorrer com relações de redundância total.

---

S1: *Até 9h30m foram registrados oito pontos de alagamento, dois deles intransitáveis - na Marginal Pinheiros, na altura da Ponte João Dias, e na Marginal Tietê, no acesso à Rodovia dos Bandeirantes.*

S2: *O CGE (Centro de Gerenciamento de Emergências) da Prefeitura de São Paulo registrava oito pontos de alagamento na cidade, às 9h30 desta segunda-feira.*

---

**Nome da relação:** *Indirect Speech*

---

**Tipo:** Apresentação/Forma->Estilo

**Direcionalidade:** S1<-S2

**Restrições:** S1 e S2 apresentam informação em comum; S1 apresenta essa informação em discurso direto e S2 em discurso indireto.

**Comentários:**

---

S1: *Lula disse que, na próxima semana, quando voltar da viagem que está fazendo pela América Central, irá começar a anunciar as obras de infra-estrutura em transporte como estradas, ferrovias, gasodutos a portos e aeroportos.*

S2: *"Quando eu voltar desta viagem pela América Central, nós vamos começar a anunciar as obras de infra-estrutura no que diz respeito a estradas, às ferrovias, a gasodutos, a tudo, a portos, a aeroportos, ou seja, tudo que tiver de infra-estrutura na área de transportes nós vamos anunciar também e começar a liberar o dinheiro para que as obras comecem a acontecer."*

---

**Nome da relação:** *Translation*

---

**Tipo:** Apresentação/Forma->Estilo

**Direcionalidade:** Nenhum

**Restrições:** S1 e S2 apresentam informação em comum em línguas diferentes.

**Comentários:**

---

S1: *Quinze voluntários da ONG francesa Ação Contra a Fome (ACF) foram assassinados no nordeste do Sri Lanka, informou hoje um porta-voz da organização.*

S2: *Segundo um representante do grupo Action Contre la Faim, os corpos foram encontrados no escritório da organização.*