Anotação automática de papéis semânticos de
textos jornalísticos e de opinião sobre árvores
sintáticas não revisadas

Nathan Siegle Hartmann

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Nathan Siegle Hartmann

**Anotação automática de papéis semânticos de textos
jornalísticos e de opinião sobre árvores sintáticas não
revisadas**

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*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Sandra Maria Aluísio

**USP – São Carlos
Agosto de 2015**

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

H333a Hartmann, Nathan Siegle
Anotação automática de papéis semânticos de textos
jornalísticos e de opinião sobre árvores sintáticas
não revisadas / Nathan Siegle Hartmann; orientadora
Sandra Maria Aluísio. -- São Carlos, 2015.
128 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, 2015.

1. Anotação de Papéis Semânticos. 2. Processamento
de Línguas Naturais. 3. Português do Brasil. I.
Aluísio, Sandra Maria, orient. II. Título.

Nathan Siegle Hartmann

**Automatic semantic role labeling on non-revised syntactic
trees of journalistic and opinion texts**

Master dissertation submitted to the Instituto de
Ciências Matemáticas e de Computação - ICMC-
USP, in partial fulfillment of the requirements for the
degree of the Master Program in Computer Science
and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and
Computational Mathematics

Advisor: Profa. Dra. Sandra Maria Aluísio

**USP – São Carlos
August 2015**

Agradecimentos

À minha família, pelo seu amor, sua confiança e seu apoio em minhas decisões. Sei que sempre posso contar com vocês, independentemente da distância que nos separa.

À minha orientadora, Profa. Sandra Maria Aluísio, pela confiança depositada em mim, pela oportunidade dada de pesquisar em uma área muito interessante do Processamento de Língua Natural (PLN), pelas críticas e sugestões sobre meu trabalho e, principalmente, pela dedicação nas incontáveis horas de escrita de artigos e da monografia e nas correções em geral.

À Magali Sanches Duran, na qualidade de coorientadora informal, que participou ativamente em meu trabalho de mestrado com críticas, sugestões e, acima de tudo, discussões que me ajudaram amadurecer não apenas no meu trabalho de mestrado mas como pesquisador.

Aos integrantes do NILC, pelas conversas, reuniões, festas, *happy hours* e, em geral, todos os momentos de convivência. Obrigado por terem compartilhado comigo a sua experiência e motivação de pesquisar em uma área desafiadora como é o PLN.

À SAMSUNG Eletrônica da Amazônia LTDA por fomentar o projeto de pesquisa no qual meu mestrado está inserido e pelo apoio financeiro prestado ao meu mestrado. Também à FAFQ - Fundação de Apoio à Física e à Química - pelo repasse da verba das bolsas e por estar sempre à disposição para atender as necessidades do projeto.

“Minha energia é o desafio,
minha motivação é o impossível,
e é por isso que eu preciso ser, à
força e a esmo, inabalável.”

Augusto Branco

Resumo

Contexto: A Anotação de Papéis Semânticos (APS) é uma tarefa da área de Processamento de Línguas Naturais (PLN) que permite detectar os eventos descritos nas sentenças e os participantes destes eventos (Palmer et al., 2010). A APS responde perguntas como “Quem?”, “Quando?”, “Onde?”, “O quê?”, e “Por quê?”, dentre outras e, sendo assim, é importante para várias aplicações de PLN. Para anotar automaticamente um texto com papéis semânticos, a maioria dos sistemas atuais emprega técnicas de Aprendizagem de Máquina (AM). Porém, alguns papéis semânticos são previsíveis e, portanto, não necessitam ser tratados via AM. Além disso, a grande maioria das pesquisas desenvolvidas em APS tem dado foco ao inglês, considerando as particularidades gramaticais e semânticas dessa língua, o que impede que essas ferramentas e resultados sejam diretamente transportados para outras línguas. **Revisão da Literatura:** Para o português do Brasil, há três trabalhos finalizados recentemente que lidam com textos jornalísticos, porém com performance inferior ao estado da arte para o inglês. O primeiro (Alva-Manchego, 2013) obteve 79,6 de F1 na APS sobre o corpus PropBank.Br; o segundo (Fonseca, 2013), sem fazer uso de um *treebank* para treinamento, obteve 68,0 de F1 sobre o corpus PropBank.Br; o terceiro (Sequeira et al., 2012) realizou anotação apenas dos papéis Arg0 (sujeito prototípico) e Arg1 (paciente prototípico) no corpus CETEM-Público, com performance de 31,3 pontos de F1 para o primeiro papel e de 19,0 de F1 para o segundo. **Objetivos:** O objetivo desse trabalho de mestrado é avançar o estado da arte na APS do português brasileiro no gênero jornalístico, avaliando o desempenho

de um sistema de APS treinado com árvores sintáticas geradas por um *parser* automático (Bick, 2000), sem revisão humana, usando uma amostragem do *corpus* PLN-Br. Como objetivo adicional, foi avaliada a robustez da tarefa de APS frente a gêneros diferentes, testando o sistema de APS, treinado no gênero jornalístico, em uma amostra de revisões de produtos da web. Esse gênero não foi explorado até então na área de APS e poucas de suas características foram formalizadas. **Resultados:** Foi compilado o primeiro *corpus* de opiniões sobre produtos da web, o *corpus* Buscapé (Hartmann et al., 2014). A diferença de performance entre um sistema treinado sobre árvores revisadas e outro sobre árvores não revisadas – ambos no gênero jornalístico – foi de 10,48 pontos de F1. A troca de gênero entre as fases de treinamento e teste, em APS, é possível, com perda de performance de 3,78 pontos de F1 (*corpus* PLN-Br e Buscapé, respectivamente). Foi desenvolvido um sistema de inserção de sujeitos não expressos no texto, com precisão de 87,8% no *corpus* PLN-Br e de 94,5% no *corpus* Buscapé. Foi desenvolvido um sistema, baseado em regras, para anotar verbos auxiliares com papéis semânticos modificadores, com confiança de 96,76% no *corpus* PLN-Br. **Conclusões:** Foi mostrado que o sistema de Alva-Manchego (2013), baseado em árvores sintáticas, desempenha melhor APS do que o sistema de Fonseca (2013), independente de árvores sintáticas. Foi mostrado que sistemas de APS treinados sobre árvores sintáticas não revisadas desempenham melhor APS sobre árvores não revisadas do que um sistema treinado sobre dados *gold-standard*. Mostramos que a explicitação de sujeitos não expressos nos textos do Buscapé, um *corpus* do gênero de opinião de produtos na web, melhora a performance da sua APS. Também mostramos que é possível anotar verbos auxiliares com papéis semânticos modificadores, utilizando um sistema baseado em regras, com alta confiança. Por fim, mostramos que o uso do sentido do verbo, como *feature* de AM, para APS, não melhora a performance dos sistemas treinados sobre o PLN-Br e o Buscapé, por serem *corpus* pequenos.

Palavras-chave. Anotação de Papéis Semânticos, Processamento de Línguas Naturais, Português do Brasil.

Abstract

Background: Semantic Role Labeling (SRL) is a Natural Language Processing (NLP) task that enables the detection of events described in sentences and the participants of these events (Palmer et al., 2010). SRL answers questions such as “Who?”, “When?”, “Where?”, “What?” and “Why?” (and others), that are important for several NLP applications. In order to automatically annotate a text with semantic roles, most current systems use Machine Learning (ML) techniques. However, some semantic roles are predictable, and therefore, do not need to be classified through ML. In spite of SRL being well advanced in English, there are grammatical and semantic particularities that prevents full reuse of tools and results in other languages. **Related work:** For Brazilian Portuguese, there are three studies recently concluded that performs SRL in journalistic texts. The first one (Alva-Manchego, 2013) obtained 79.6 of F1 on the SRL of the PropBank.Br corpus; the second one (Fonseca, 2013), without using a treebank for training, obtained 68.0 of F1 for the same corpus; and the third one (Sequeira et al., 2012) annotated only the Arg0 (prototypical agent) and Arg1 (prototypical patient) roles on the CETEMPúblico corpus, with a performance of 31.3 of F1 for the first semantic role and 19.0 for the second one. None of them, however, reached the state of the art of the English language. **Purpose:** The goal of this master’s dissertation was to advance the state of the art of SRL in Brazilian Portuguese. The training corpus used is from the journalistic genre, as previous works, but the SRL annotation is performed on non-revised syntactic trees, i.e., generated by an automatic parser (Bick, 2000) without human re-

vision, using a sampling of the corpus PLN-Br. To evaluate the resulting SRL classifier in another text genre, a sample of product reviews from web was used. Until now, product reviews was a genre not explored in SRL research, and few of its characteristics are formalized. **Results:** The first corpus of web product reviews, the Buscapé corpus ([Hartmann et al., 2014](#)), was compiled. It is shown that the difference in the performance of a system trained on revised syntactic trees and another trained on non-revised trees – both from the journalistic genre – was of 10.48 of F1. The change of genres during the training and testing steps in SRL is possible, with a performance loss of 3.78 of F1 (corpus PLN-Br and Buscapé, respectively). A system to insert unexpressed subjects reached 87.8% of precision on the PLN-Br corpus and a 94.5% of precision on the Buscapé corpus. A rule-based system was developed to annotated auxiliary verbs with semantic roles of modifiers (ArgMs), achieving 96.76% confidence on the PLN-Br corpus. **Conclusions:** First we have shown that [Alva-Manchego \(2013\)](#) SRL system, that is based on syntactic trees, performs better annotation than [Fonseca \(2013\)](#)'s system, that is non-dependent on syntactic trees. Second the SRL system trained on non-revised syntactic trees performs better over non-revised trees than a system trained on gold-standard data. Third, the explicitation of unexpressed subjects on the Buscapé texts improves their SRL performance. Additionally, we show it is possible to annotate auxiliary verbs with semantic roles of modifiers, using a rule-based system. Last, we have shown that the use of the verb sense as a feature of ML, for SRL, does not improve the performance of the systems trained over PLN-Br and Buscapé corpus, since they are small.

Keywords. Semantic Role Labeling, Natural Language Processing, Brazilian Portuguese.

Sumário

Lista de Figuras	v
Lista de Tabelas	vii
Lista de Abreviações	xi
1 Introdução	1
1.1 Contextualização e Motivação	1
1.2 Objetivos e Hipóteses	7
1.3 Organização do Trabalho	9
2 Fundamentação Teórica	11
2.1 Papéis Semânticos	12
2.1.1 PropBank	13
2.1.2 PropBank.Br	15
2.2 Anotação de Papéis Semânticos	17
2.2.1 Anotação Manual de Papéis Semânticos	17
2.2.2 Anotação Automática de Papéis Semânticos	19
2.3 As Principais <i>Features</i> Utilizadas em APS Automática via AM	21
2.4 Considerações Finais	23
3 Trabalhos Relacionados	25

3.1	Sistemas Supervisionados	26
3.1.1	Trabalhos Pioneiros em APS	26
	Trabalho de Gildea and Jurafsky (2002)	26
	Trabalho de Palmer et al. (2005)	27
3.1.2	Trabalhos para o Português	28
	Trabalho de Alva-Manchego (2013)	28
	Trabalho de Fonseca (2013)	30
	Trabalho de Sequeira et al. (2012)	31
3.1.3	Estado da Arte em APS	33
3.1.4	Novidades na área: APS Intergêneros	34
3.2	Sistemas Não-Supervisionados	37
3.2.1	Abordagem Bilíngue: Inglês-Alemão (Titov and Klementiev, 2012b)	37
3.3	Sistemas Semissupervisionado	39
3.3.1	Trabalho de Alva-Manchego (2013)	39
3.3.2	Trabalho de (Titov and Klementiev, 2012c)	40
3.3.3	Trabalho de (Kozhevnikov and Titov, 2013)	42
3.4	Considerações Finais	43
4	Projeto Brazilis	47
4.1	Recursos	48
4.1.1	Córpus e Processos de Anotação Manual	49
	PropBank.Br (versão 1.1)	49
	Seleção do PLN-Br	49
	Anotação Manual de Papéis Semânticos sobre a Seleção do PLN-Br	53
	Seleção do Buscapé	57
	Anotação Manual de Papéis Semânticos sobre a Seleção do Buscapé	62
4.1.2	Regras para Identificação de Verbos Auxiliares	63
4.2	Uma Arquitetura Híbrida para APS Automática	67
4.2.1	Módulo de Inserção de Sujeitos Não Expressos	69
4.2.2	Anotação de Verbos Auxiliares com Papéis Semânticos Modifica- dores	74

4.2.3	Módulo de APS baseado em AM	75
4.3	Considerações Finais	76
5	Avaliações: Resultados e Discussão	79
5.1	Avaliação do Módulo de Inserção de Sujeitos Não Expressos	81
5.2	Avaliação do Sistema de APS em Verbos Auxiliares	84
5.3	Experimentos com o Uso do Sentido do Verbo como <i>Feature</i> de AM . . .	89
5.4	Experimentos de APS no Gênero Jornalístico: PropBank.Br e PLN-Br . .	91
5.5	Experimentos de APS no Gênero de Opiniões sobre Produtos da Web: Buscapé	98
5.6	Considerações Finais	105
6	Conclusões	107
6.1	Contribuições	109
6.2	Limitações	111
6.3	Trabalhos Futuros	112
	Referências Bibliográficas	112
A	Quantidade de Instâncias Seleccionadas por Verbo do Corpus Buscapé	125
B	Tabela com as Etiquetas do Guia de Anotação do PropBank.Br	127

Lista de Figuras

1	Exemplo de um <i>frameset</i> do PropBank para o verbo <i>to kick</i>	15
2	Modelo de APS manual definida por Palmer et al. (2005).	18
3	Árvore sintática para a sentença “Ele receberá o valor à vista após 30 dias.”.	20
4	Valores de F1 para a classificação de papéis semânticos nos sistemas Supervisionado, Não-Supervisionado, Semissupervisionado e SyntF (Titov and Klementiev, 2012c).	41
5	Sentença paralela em inglês e tcheco, anotada com papéis semânticos e com ligações entre palavras sinônimas entre as línguas (Kozhevnikov and Titov, 2013).	42
6	Exemplo de diferença entre os corpus PropBank.Br versão 1 (à esquerda) e PropBank.Br versão 1.1 (à direita). Circulada em vermelho, a anotação que demarca o sentido do verbo.	50
7	Exemplo de interface para anotação na ferramenta SALTO.	54
8	Dados para interpretação da estatística Kappa segundo Landis and Koch (1977).	56
9	Exemplos de textos do corpus Buscapé com seus problemas característicos para um <i>parser</i> dedicado à analisar textos bem escritos.	59
10	<i>Pipeline</i> de pré-processamento do corpus Buscapé.	61
11	Exemplo de anotação de verbos auxiliares, conforme o PropBank inglês.	66

12	Exemplo de anotação de verbos auxiliares, conforme proposto por Duran and Aluísio (2010).	67
13	Diagrama do sistema híbrido proposto para anotação de papéis semânticos.	68
14	Sentença “Agora, os soldados e a polícia estão trabalhando juntos para prender os traficantes.” no formato plano de colunas da CoNLL.	80
15	Valores para a métrica Suporte das regras utilizadas na identificação de verbos auxiliares no corpus PLN-Br.	86

Lista de Tabelas

1	Etiquetas ArgM definidas por Palmer et al. (2005).	14
2	Resultados do sistema supervisionado de Gildea and Jurafsky (2002) para testes com <i>parser</i> sintático automático.	27
3	Resultados do sistema de Palmer et al. (2005) para teste com dados passados por <i>parser</i> sintático automático, dados <i>gold-standard</i> e <i>gold-standard</i> com traços.	28
4	Resultados do sistema supervisionado de Alva-Mancheiro (2013) para teste com dados <i>gold-standard</i>	30
5	Resultados do sistema supervisionado de Fonseca (2013) para teste com dados <i>gold-standard</i>	31
6	Frequência de ocorrência de verbos, sujeitos e objetos nos corpúscos analisados por Sequeira et al. (2012).	32
7	Resultados da anotação de papéis semânticos de Sequeira et al. (2012) para o corpúscos BosqueUE.	33
8	Resultados da anotação de papéis semânticos de Sequeira et al. (2012) para o conjunto de teste do CoNLL-2004.	33
9	Resultados do sistema supervisionado de Toutanova et al. (2008) para teste com dados <i>gold-standard</i> e com <i>parser</i> sintático automático.	34

10	Resultados de Clercq et al. (2012) para a classificação de argumentos, com utilização de 45 mil palavras por gênero.	37
11	Resultados de Clercq et al. (2012) para a classificação de argumentos, utilizando 50 mil palavras por gênero para o primeiro caso e 50 mil palavras somadas às 45 mil do gênero em teste para o segundo caso. . . .	37
12	Resultados do sistema de Titov and Klementiev (2012b).	39
13	Resultados do sistema semissupervisionado de Alva-Manchego (2013) para teste com dados <i>gold-standard</i>	40
14	Valores de <i>Purity</i> , <i>Collocation</i> e F1 para a utilização de 300 sentenças como conjunto de treinamento (Titov and Klementiev, 2012c).	41
15	Resultados dos modelos <i>self-training</i> e <i>joint-probability</i> de Kozhevnikov and Titov (2013), em termos de F1.	43
16	Resumo das características dos trabalhos relacionados apresentados no Capítulo 3.	44
17	Estatísticas sobre os 34.110.689 <i>tokens</i> do corpus PLN-Br.	50
18	Distribuição das instâncias coletadas do PLN-Br.	53
19	Valores de Kappa para a Identificação de Argumentos, Classificação de Argumentos e escolha de sentido da anotação manual da seleção do PLN-Br.	55
20	Proporção da anotação do PLN-Br em que houve plena concordância entre os anotadores.	58
21	Estatísticas do corpus Buscapé, em comparação com o Corpo São Carlos.	59
22	Valores de Kappa para a Identificação de Argumentos, Classificação de Argumentos e Escolha de Sentido da anotação manual da seleção do Buscapé.	63
23	Proporção da anotação do Buscapé em que houve plena concordância entre os anotadores.	63
24	Regras para identificação de verbos auxiliares e etiqueta atribuída à função exercida pelo verbo auxiliar.	66
25	Resultados de uma análise preliminar sobre a inserção de sujeitos não expressos em uma seleção de 200 sentenças do corpus PropBank.Br. . . .	73

26	Fonte de erros da análise preliminar sobre a inserção de sujeitos não expressos em uma seleção de 200 sentenças do <i>córpus PropBank.Br</i>	73
27	Resultados de um experimento <i>10-fold cross-validation</i> com a seleção do <i>córpus Buscapé</i> , na análise do uso de sujeitos ocultos.	83
28	Resultados da inserção automática de papéis semânticos para verbos auxiliares.	85
29	Resultado da anotação de verbos auxiliares no <i>córpus PLN-Br</i> , via AM. .	89
30	Resultados de <i>10-fold cross-validation</i> para os conjuntos <i>Concordância</i> e <i>Adjudicados</i> do <i>PLN-Br</i> , ao utilizar o sistema de Alva-Manchego (2013). .	93
31	Resultados de <i>10-fold cross-validation</i> para o conjunto <i>Full</i> da seleção do <i>PLN-Br</i> , composto pelos conjuntos <i>Concordância</i> e <i>Adjudicados</i> , ao utilizar o sistema de Alva-Manchego (2013).	94
32	Resultados dos experimentos de APS verificando a troca de contexto entre as fases de treinamento e teste para árvores sintáticas revisadas e não revisadas, ao utilizar o sistema de Alva-Manchego (2013).	95
33	Resultados da anotação dos conjuntos <i>Concordância</i> e <i>Adjudicados</i> do <i>PLN-Br</i> utilizando o sistema de APS treinado sobre o <i>PropBank.Br</i> , ao utilizar o sistema de Alva-Manchego (2013)	96
34	Resultados de <i>10-fold cross-validation</i> para o conjunto <i>Full</i> da seleção do <i>PLN-Br</i> , composto pelos conjuntos <i>Concordância</i> e <i>Adjudicados</i> , utilizando o sistema de Fonseca (2013).	97
35	Resultados de <i>10-fold cross-validation</i> para os conjuntos <i>Concordância</i> e <i>Adjudicados</i> da seleção do <i>Buscapé</i> , ao utilizar o sistema de Alva-Manchego (2013).	99
36	Resultados de <i>10-fold cross-validation</i> para o conjunto <i>Full</i> da seleção do <i>Buscapé</i> , composto pelos conjuntos <i>Concordância</i> e <i>Adjudicados</i> , ao utilizar o sistema de Alva-Manchego (2013).	100
37	Resultados da anotação do conjunto <i>Concordância</i> do <i>Buscapé</i> utilizando os sistema de APS treinados sobre o <i>PropBank.Br</i> e sobre o <i>PLN-Br</i> , ao utilizar o sistema de Alva-Manchego (2013).	102

38	Resultados da anotação do conjunto Adjudicados do Buscapé utilizando os sistema de APS treinados sobre o PropBank.Br e sobre o PLN-Br, ao utilizar o sistema de Alva-Manchego (2013).	103
39	Resultados da anotação do conjunto <i>Full</i> do Buscapé utilizando os sistema de APS treinados sobre o o PropBank.Br. e sobre o PLN-Br, ao utilizar o sistema de Alva-Manchego (2013).	104
40	Resultados da anotação do conjunto <i>Full</i> do Buscapé utilizando os sistema de APS treinados sobre o o PropBank.Br. e sobre o PLN-Br, ao utilizar o sistema de Fonseca (2013).	105
41	Distribuição das instâncias selecionadas por verbo do corpus Buscapé. .	126
42	Conjunto de papéis semânticos utilizado nesse trabalho de mestrado. .	128

Lista de Abreviaturas

ADVP	<i>Adverbial Phrase</i>
AM	Aprendizagem de Máquina
APS	Anotação de Papéis Semânticos
CO	<i>Collocation</i>
F1	<i>F-measure</i> simples
PB	Português Brasileiro
PLN	Processamento de Línguas Naturais
PoS	<i>Part-of-speech</i>
PU	<i>Purity</i>
NP	<i>Noun Phrase</i>
PP	<i>Prepositional Phrase</i>
VP	<i>Verb Phrase</i>
WSJ	<i>Wall Street Journal</i>

Introdução

1.1 Contextualização e Motivação

A Anotação de Papéis Semânticos (APS), do inglês *Semantic Role Labeling*, é uma tarefa da área de Processamento de Línguas Naturais (PLN) responsável por detectar os eventos descritos nas sentenças e os participantes destes eventos (Palmer et al., 2010). Os eventos são realizados por predicadores como verbos, pronomes ou adjetivos, e os participantes são chamados de argumentos. No caso de predicadores verbais, podemos considerar um verbo simples como “arremessar” ou verbos complexos como “dar conta”. Por exemplo, para a sentença “João arremessou a pedra.”, temos o verbo “arremessar” e seus argumentos, “João” e “pedra”; para a sentença “João deu conta do trabalho.”, temos como argumentos “João” e “do trabalho”.

A tarefa de APS, mais especificamente, consiste em identificar/delimitar o predador e seus argumentos e atribuir a cada argumento um rótulo de papel semântico, de um conjunto pré-determinado de etiquetas. Para a sentença “João quebrou o vaso com o martelo.”, podemos atribuir aos argumentos do verbo *quebrar* o papel de *agente* a “João”, de *paciente* a “vaso” e de *instrumento* a “martelo”. Mesmo utilizando um *parser* para, primeiramente, identificar em uma sentença os argumentos de interesse da

anotação, ficamos na dependência do formalismo gramatical em que o *parser* se baseia para realizar a anotação desejada. Nem todos os *parsers* trabalham com a anotação de predicados complexos, por exemplo.

Segundo [Kipper \(2005\)](#), [Cançado \(2005\)](#) e [Scarton \(2013\)](#), não existe um consenso sobre o conjunto de papéis necessários para caracterizar os argumentos de um predicado, assim como não existe um critério para determinar quais papéis devem ser aplicados a um argumento em particular. [Kipper \(2005\)](#) define a seguinte lista de papéis semânticos como sendo os mais comuns: Agente, Paciente, Tema, Experienciador, Estímulo, Instrumento, Posição, Fonte, Meta, Receptor e Beneficiário. Porém, como não há consenso sobre o conjunto universal de papéis semânticos, outras teorias são estudadas e utilizadas para anotação de papéis semânticos. O projeto PropBank ([Palmer et al., 2005](#)) define um conjunto de 6 papéis semânticos numerados e livres de teorias semânticas, (Arg0–Arg5), todos variáveis, exceto o Arg0, que sempre é o agente, e Arg1, que ora é agente, ora paciente. Além disso, é definido um conjunto de 11 papéis semânticos para adjuntos, chamados de papéis modificadores. A VerbNet ([Kipper, 2005](#)) especifica um conjunto de 23 papéis semânticos. Por fim, a Framenet ([Baker et al., 1998](#)) define papéis específicos para cada verbo, não sendo comumente utilizada para tarefas de Aprendizagem de Máquina (AM) devido à sua alta granularidade, embora tenha sido utilizada em conjunto com a VerbNet e PropBank para a construção de um *parser* semântico robusto, via interconecção dos três recursos ([Giuglea and Moschitti, 2006](#)).

Para anotar automaticamente um texto com papéis semânticos, a maioria dos sistemas de APS atuais empregam técnicas probabilísticas de AM, uma vez que a tarefa pode ser considerada como um problema de classificação (abordagem supervisionada): para dado predicador (verbo ou predicado complexo) e um constituinte sintático, a tarefa é selecionar, de um conjunto pré-definido de papéis semânticos, o que melhor se enquadra ao constituinte ([Palmer et al., 2010](#)). A tarefa também pode ser considerada um problema de agrupamento (abordagem não-supervisionada) em que, devido ao usuário não possuir dados anotados, a técnica de AM separa os dados de acordo com alguma medida de similaridade e cabe a um especialista de domínio avaliar o resultado do agrupamento. A tarefa ainda pode ser modelada como um problema misto (abordagem semissupervisionada), que é utilizada em cenários em que há uma quantidade insufici-

ente de dados anotados, porém muitos dados não anotados. Nesse caso, são utilizadas técnicas tanto de classificação quanto de agrupamento para, por meio de um conjunto inicial de dados rotulados, agrupar dados não rotulados.

Ao utilizar AM, a tarefa de APS é geralmente realizada sobre árvores sintáticas devido ao extenso conjunto de *features* já levantadas (cita-se a listagem apresentada no trabalho de [Palmer et al. \(2010\)](#)). No entanto, as árvores sintáticas de uma sentença são geradas por um *parser* sintático automático e, como essas ferramentas são sujeitas a erros, árvores sintáticas problemáticas são muitas vezes processadas na tarefa de APS. Portanto, os pesquisadores da área de APS muitas vezes corrigem manualmente as árvores sintáticas para lidar com menos dados ruidosos, mas que não representam um cenário real. Essas árvores sintáticas revisadas são chamadas de árvores *gold-standard*¹. Um exemplo da diferença de performance na APS sobre árvores sintáticas pode ser visto no trabalho de [Toutanova et al. \(2008\)](#), que obteve 91,2 de F1² ao anotar papéis semânticos em árvores corrigidas e 80,0 de F1 ao anotar árvores não corrigidas. Essa disparidade é um desafio para a tarefa.

Como os métodos de AM são sujeitos a erros (como é o caso do processamento sobre árvores sintáticas), nem sempre é interessante utilizar AM na anotação automática de papéis semânticos. Métodos de AM são utilizados para tratar problemas em que existe certa variabilidade, erros ou exceções nos dados. A classe dos verbos auxiliares, caracterizada pelos verbos que adicionam informação gramatical ao verbo em sequência (verbo alvo), no entanto, são altamente previsíveis, ou seja, podem ser identificados via utilização de regras. É o caso da sentença “João tentou ajudar seu irmão.”, em que o verbo “tentar” assume papel de auxiliar do verbo “ajudar”, via regra “*tentar* seguido de verbo no infinitivo”.

Segundo [Feng et al. \(2012\)](#), a APS captura o significado essencial das sentenças, o que a torna uma tarefa essencial para que os computadores possam extrair e utilizar efetivamente as informações expressas na forma textual. É o caso das aplicações que devem responder às perguntas “Quem?”, “Quando?”, “Onde?”, “O quê?”, e “Por quê?”

¹Um conjunto de dados *gold-standard*, geralmente revisado por especialistas, tendem a não conter erros para uma dada tarefa, por exemplo, *parsing* sintático.

²A medida F1 ($F\text{-measure}_1$) é a média ponderada simples entre as medidas de precisão e cobertura.

que, segundo a literatura, seriam os sistemas de Extração de Informação ([Surdeanu et al., 2003](#); [Chambers and Jurafsky, 2011](#)), Perguntas e Respostas ([Shen and Lapata, 2007](#); [Kaisser and Webber, 2007](#)), Sumarização Automática ([Knight and Marcu, 2002](#)), Tradução Automática ([Wu and Fung, 2009](#); [Liu and Gildea, 2010](#); [Lo and Wu, 2011](#); [Gao and Vogel, 2011](#); [Baker et al., 2012](#)), Sistemas de Diálogo ([Basili et al., 2009](#); [Van der Plas et al., 2011](#)) entre outros ([Palmer et al., 2010](#)).

Atualmente, várias aplicações de PLN utilizam APS dentro de seu procedimento. [Salza \(2013\)](#) propõe um sistema para anotação semântica de expressões em italiano que veiculam ideia de movimento. Após a anotação de papéis semânticos, a sentença é anotada com informações referentes ao verbo (que indica uma ação de movimento).

A sumarização automática de textos possibilita a redução da redundância de sentenças de dado texto e a geração de sumários que representam o mesmo conteúdo de uma forma mais coesa ([Knight and Marcu, 2002](#)). O trabalho de [Yoshikawa et al. \(2012\)](#) utiliza APS para cobrir fraquezas da árvore de dependências sintáticas na atividade de sumarização automática de textos. Os autores argumentam que a identificação de dependências sintáticas não é suficiente para sumarizar algumas sentenças complexas que contenham voz passiva, verbo auxiliar ou coordenação.

A resolução de anáforas trata da ligação de referências a um mesmo elemento em um texto e é foco de estudo da área de PLN há décadas. [Kong and Zhou \(2012\)](#) propuseram, via APS, um método para resolução de eventos pronominais, o tipo mais difícil de evento anafórico a ser resolvido para o inglês, devido à informação pouco discriminativa contida nos pronomes.

Os trabalhos mais recentes em Tradução Automática têm utilizado métodos probabilísticos, com bons resultados. [Jones et al. \(2012\)](#) propuseram um tradutor automático estatístico-semântico que trabalha em duas etapas: na primeira é analisado o significado da sentença da língua origem e então é feita uma decodificação do significado para a língua destino. Na segunda etapa, a sentença é anotada semanticamente sobre uma estrutura sintática, utilizando o modelo PropBank de APS ([Palmer et al., 2005](#)). Já em outro trabalho, [Feng et al. \(2012\)](#) propuseram dois modelos para tradução automática de textos, utilizando APS para reordenar os argumentos do verbo da língua de origem para a destino.

Xu et al. (2012) propuseram um *framework* para identificação de argumentos de discurso por meio de um anotador de papéis semânticos intersentencial e intrasentencial. O primeiro é baseado em regras heurísticas. O segundo utiliza APS para identificar os argumentos do predador.

O processamento de histórias infantis, como contos de fadas, tem chamado a atenção da comunidade de PLN, pois pode revelar novos problemas (questões de pesquisa) para aplicações da área. Maarouf and Villaneau (2012) desenvolveram um corpus anotado nos níveis sintático e semântico (com papéis semânticos) para representar o gênero de histórias infantis.

Com as exemplificações de trabalhos recentes que utilizam APS, percebemos que a tarefa de APS é necessária a muitas aplicações, o que estimula estudos para o desenvolvimento de sistemas de APS para diversas línguas, como o trabalho de Hart et al. (2012) em que propuseram um anotador de papéis semânticos para a língua árabe. Essa língua possui características muito específicas, por exemplo, morfologia rica, em que substantivos e adjetivos codificam informações sobre número, gênero, caso e flexão, e os verbos (ainda mais ricos) codificam tempo, voz, modo, gênero, número e pessoa. Os autores utilizaram *features* (atributos) clássicas da literatura³ e acrescentaram outras específicas da língua árabe, melhorando o desempenho do classificador.

Johansson et al. (2012) desenvolveram o primeiro anotador de papéis semânticos para o sueco a utilizar a *Swedish FrameNet*⁴ como recurso. O conjunto de *features* utilizado estão presentes no trabalho de Johansson (2008). Devido ao corpus utilizado ser considerado pequeno para realizar um treinamento supervisionado, foi proposto um modelo de relaxamento da FrameNet, em que rótulos similares, porém em *frames* diferentes, são considerados idênticos (ao contrário do modelo original). Testes mostraram que, com a utilização de dados *gold-standard*, a fase de identificação de argumentos obteve precisão de 70,6% e cobertura de 64,8% e a fase de classificação dos papéis semânticos para os argumentos obteve acurácia de 64,3%.

Para a língua portuguesa do Brasil, Fonseca (2013) obteve 68,0 de F1 para a tarefa conjunta de identificação de argumentos e classificação de papéis semânticos em seu

³Exemplos de *features* clássicas para APS podem ser vistos em Palmer et al. (2010).

⁴Disponível em <http://spraakbanken.gu.se/eng/swefn>.

sistema baseado em redes neurais. Esse trabalho não utilizou recursos de PLN, por exemplo, *parser* sintático. [Alva-Manchego \(2013\)](#) obteve 79,6 de F1 em seu sistema supervisionado utilizando máxima entropia (também conhecido como regressão logística) e 80,5 de F1 em seu sistema semissupervisionado, utilizando *self-training* e máxima entropia. Os resultados de [Alva-Manchego \(2013\)](#) são para dados *gold-standard*.

As pesquisas sobre a APS para o inglês, ao contrário das demais línguas, encontra-se mais consolidada na tarefa de anotação de papéis semânticos. O trabalho de [Toutanova et al. \(2008\)](#) obteve 91,2 de F1 para a tarefa conjunta de identificação e classificação de papéis semânticos em seu sistema supervisionado, que considerou a interdependência entre os argumentos do verbo, ao avaliá-lo sobre dados *gold-standard*, e 80,0 de F1 ao avaliá-lo sobre árvores sintáticas não revisadas.

A APS para línguas diferentes do inglês possui um desempenho muito inferior aos sistemas do estado da arte do inglês, devido à falta de dados anotados para essas línguas ([Fürstenau and Lapata, 2009](#)). Além desse problema, [Palmer et al. \(2010\)](#) listam limitações dos sistemas atuais, como a baixa capacidade de generalização e a baixa portabilidade dos sistemas desenvolvidos. Um exemplo da falta de generalização está no trabalho de [Pradhan et al. \(2008\)](#), que demonstra a perda de desempenho na troca do gênero das sentenças dos *corpus* das fases de treinamento e teste, chegando a 10% na medida F1. [Toutanova et al. \(2008\)](#) comentam, citando a tese de doutorado de [Roland \(2002\)](#), que se o sentido do verbo em uma sentença for desambiguado, a anotação dos papéis semânticos torna-se mais fácil, melhorando a qualidade da anotação. Como o baixo desempenho da APS compromete os resultados de aplicações que fazem uso de conhecimento semântico, há a necessidade contínua de estudos em busca de melhoria da qualidade da anotação de papéis semânticos, possibilitando uma melhor qualidade dos sistemas que utilizam esse conhecimento.

Este projeto de mestrado, chamado “*Brazilis*”, está inserido em um projeto maior denominado *Processamento Semântico de Textos em Português Brasileiro*, no qual o ICMC/USP e a SAMSUNG Eletrônica da Amazônia LTDA estão desenvolvendo recursos e ferramentas de base para a criação de aplicações envolvendo o uso de PLN. O objetivo geral desse projeto maior é avançar o estado da arte em processamento semântico de textos/documentos escritos em português brasileiro, mais especificamente, permitir a ano-

tação de papéis semânticos e a desambiguação lexical de sentido de verbos e, com base nesses recursos e ferramentas, construir aplicações de mineração e sumarização de textos, com especial enfoque em opiniões sobre produtos encontrados na web.

1.2 Objetivos e Hipóteses

A literatura tem deixado várias pistas sobre como melhorar o desempenho de um sistema de APS:

- O corpus Penn Treebank (Marcus et al., 1993) utiliza *null-elements* para tratar elementos suprimidos. Palmer et al. (2005) mostra que o uso desses *null elements* aumenta o desempenho do sistema de APS.
- Toutanova et al. (2008) demonstra que há **interdependência** entre a anotação dos papéis semânticos. Lidando com esse fenômeno, conseguiu o estado da arte para a tarefa.
- Toutanova et al. (2008) afirma que a **desambuiguização do sentido do verbo** alvo da anotação facilita a tarefa de APS.
- Pradhan et al. (2008) mostra que **na troca de gênero** entre o treinamento e teste de um sistema de APS que utiliza AM, há uma perda de desempenho de até **10% na medida F1**.

Neste cenário, os objetivos deste trabalho de mestrado são:

- Avançar o estado da arte na APS de textos jornalísticos em português brasileiro;
- Avaliar a tarefa de APS automática para o gênero de textos de opinião sobre produtos na Web.

Textos de opiniões sobre produtos na web podem ser obtidos de websites como o Buscapé⁵, o Mercado Livre⁶ e o Reclame Aqui⁷. Esse gênero não foi explorado até então na

⁵Acessível em <http://buscape.com.br>.

⁶Acessível em <http://mercadolivre.com.br>

⁷Acessível em <http://reclameaqui.com.br>

área de APS, e poucas de suas características foram formalizadas.

As hipóteses a serem verificadas neste trabalho de mestrado são:

1. A explicitação de sujeitos ocultos, potenciais candidatos à papéis semânticos que demarcam o sujeito do verbo, melhora a medida F1 de um sistema de APS.

Essa hipótese é motivada pelo uso de *null elements* no trabalho de [Palmer et al. \(2005\)](#), que melhorou a performance do sistema de APS. Também nos baseamos no trabalho de [Toutanova et al. \(2008\)](#) que mostra a interdependência entre os argumentos de APS, ou seja, que a ausência do sujeito impacta a anotação dos demais argumentos ao utilizar a arquitetura de interdependência de [Toutanova et al. \(2008\)](#).

2. A anotação automática de verbos auxiliares com papéis semânticos modificadores pode ser realizada via regras com precisão similar à obtida com um sistema de AM.

Ao anotar verbos auxiliares como modificadores de verbos plenos, eles deixam de ser tratados como verbos alvo de APS e passam a ser argumentos. Assim, não inflamos a nota de desempenho do sistema de APS anotando um conjunto altamente previsível.

3. O uso do sentido do verbo alvo de APS, como *feature* de AM, melhora a medida F1 de um sistema de APS.

Nos baseamos no trabalho de [Toutanova et al. \(2008\)](#), que mostrou que a desambiguação do sentido do verbo alvo de anotação facilita a tarefa de APS.

4. Em um ambiente em que o sistema de APS deve atuar sobre árvores sintáticas não revisadas por humanos, ou seja, árvores que podem conter erros, é necessário treinar o sistema de APS com árvores não revisadas.

Nunca um sistema de APS em PB foi avaliado sobre árvores sintáticas não revisadas. Esse trabalho de mestrado é o primeiro a fazer tal avaliação.

5. A troca de gênero entre as fases de treinamento e teste de um sistema de APS, treinado com um corpus jornalístico e testado com um corpus do gênero de opinião sobre produtos na web, degrada a medida F1 em 10% na APS realizada por esse sistema.

Nos baseamos no trabalho de [Pradhan et al. \(2008\)](#), que mostrou que há uma perda de 10% ao trocar o gênero dos corpúsculos de treinamento e teste na tarefa de APS.

As seguintes questões de pesquisa são levantadas e motivam a investigação das hipóteses apresentadas.

- Caso o uso do sentido do verbo alvo como *feature* não agregue performance ao sistema de APS, qual é a razão?
- Os ruídos contidos em árvores sintáticas não revisadas por humanos são suficientemente significativos para inviabilizar o uso de um sistema treinado com as árvores sintáticas desses dados? Será que um sistema de APS que não faz uso de árvores sintáticas obtém melhor performance com dados ruidosos do que um sistema que faz uso de árvores sintáticas?
- Quais os motivos que degradam a performance do sistema de APS ao treinar em um gênero e testar em outro?
- O desempenho de um sistema treinado no gênero jornalístico e testado no gênero de opiniões sobre produtos web é superior ao de um sistema treinado e testado sobre o segundo gênero?

Este trabalho de mestrado se beneficia do projeto maior (em parceria entre universidade e instituição privada) ao utilizar o corpúsculo Buscapé, composto por textos do gênero de opinião sobre produtos em português brasileiro, e também o conhecimento e discussões de uma grande equipe de professores e bolsistas.

1.3 Organização do Trabalho

As seções que seguem são divididas da seguinte maneira: no Capítulo 2 são apresentados conceitos sobre Léxicos Computacionais e Papéis Semânticos, assim como exemplos importantes da literatura que utilizam esses conceitos e também o modelo de anotação manual de papéis semânticos utilizado no PropBank inglês, contrastando-o com o utilizado na anotação da versão brasileira, o PropBank.Br; mostramos o modelo de

anotação automática de papéis semânticos utilizado por [Palmer et al. \(2005\)](#); definimos as três vertentes da área de Aprendizagem de Máquina e são listadas as principais *features* utilizadas para anotação de papéis semânticos, segundo a abordagem de AM. No Capítulo [3](#) são apresentados os trabalhos mais importantes da área de APS e os trabalhos desenvolvidos ao longo dos últimos quatro anos. A organização dos trabalhos relacionados é feita pela abordagem de AM utilizada: sistemas supervisionados, não-supervisionados e semissupervisionados. No Capítulo [4](#) é apresentado o projeto *Brazilis*, os recursos utilizados e os métodos desenvolvidos. No Capítulo [5](#) é apresentado o planejamento de avaliações e os resultados obtidos. Finalmente, no Capítulo [6](#) são apresentadas as conclusões deste trabalho de mestrado.

Fundamentação Teórica

Um léxico de uma língua é um dos recursos linguísticos primários no que diz respeito à Engenharia de Linguagem ([Zavaglia, 2003, 2006](#)). Em PLN, o termo léxico se refere ao conjunto de dados que contém informações adequadas e codificadas a fim de que um programa computacional, ou um algoritmo, possa processá-los e utilizá-los. Segundo [Zavaglia \(2003, 2006\)](#), as informações contidas em um léxico “podem ser de vários níveis linguísticos (morfológico, morfossintático, sintático, semântico), sendo que para cada um deles existe um tipo de codificação de dados”. Palavras classificadas como verbos, por exemplo, possuem uma grande carga de informação sintática e semântica, afinal, a partir do conhecimento do comportamento de um verbo é possível saber como pode ser a construção da sentença ([Scarton, 2013](#)), identificando-se:

- A relação de verbos que admitem objetos diretos.
- A relação dos verbos que admitem objetos indiretos e quais as preposições admitidas pelos mesmos.
- A relação de verbos que admitem objetos diretos e indiretos simultaneamente.
- A relação de verbos que admitem objeto direto, mas podem ser usados intransitivamente (sem objeto).

- A relação de verbos que admitem objeto direto, mas podem ser usados intransitivamente (sem objeto) com transferência do objeto para posição do sujeito.
- A relação dos verbos que admitem uso causativo.

Logo, os verbos devem integrar os recursos, pois por meio deles é possível adquirir muito conhecimento sobre uma sentença. Papéis semânticos, por exemplo, podem ser centrados em um verbo da sentença, chamado de verbo alvo.

2.1 Papéis Semânticos

Um conjunto de informação muito importante para a área de PLN, composto pelas relações semânticas entre um predador e seus argumentos, recebe o nome de conjunto dos **papéis semânticos**. Um predador pode ser representado por nomes, adjetivos, advérbios e também verbos. Nesse trabalho lidamos apenas com papéis semânticos para verbos. A tarefa de identificar quais palavras atuam como argumentos da ação de um verbo é denominada anotação de papéis semânticos ([Shamsfard and Mousavi, 2007](#)).

Para a área de PLN, a anotação de papéis semânticos, utilizando corpus anotados para auxiliar a tarefa, foi idealizada primeiramente por [Gildea and Jurafsky \(2001, 2002\)](#), empregando a *Framenet* ([Baker et al., 1998](#)) como corpus de treinamento. A partir desse marco, vários projetos desenvolvendo APS foram realizados para diversas línguas, dos quais citam-se trabalhos desenvolvidos para o inglês: [Gildea and Palmer \(2002\)](#), [Gildea and Hockenmaier \(2003\)](#), [Surdeanu et al. \(2003\)](#), [Palmer et al. \(2005\)](#), [Yi et al. \(2007\)](#), [Toutanova et al. \(2008\)](#) e [Pradhan et al. \(2008\)](#).

Segundo [Palmer et al. \(2005, 2010\)](#), um dos trabalhos mais conceituados na área de APS utilizando corpus anotado, a anotação de papéis semânticos sobre sentenças pode favorecer aplicações como: extração de informação, tradução automática e sistemas de perguntas e respostas, o que motivou o desenvolvimento de um recurso lexical chamado *Propositional Bank*¹ ([Palmer et al., 2005](#)), adequado à abordagem de aprendizagem de máquina para a tarefa de criação de classificadores de alto desempenho.

¹O *Propositional Bank* é comumente conhecido como PropBank.

2.1.1 PropBank

O banco proposicional PropBank foi desenvolvido por [Palmer et al. \(2005\)](#) para a língua inglesa. O termo *proposition*, que dá nome ao recurso lexical, foi proposto por [Fillmore \(1967\)](#) e consiste de um verbo e um ou mais sintagmas nominais, associados ao verbo em uma particular relação de caso, por exemplo, agente, beneficiário, locativo. Como as listas de casos se diferem de um teórico para outro e mesmo em trabalhos de um mesmo teórico, o PropBank fez a proposta de utilizar um conjunto de papéis com propósito geral. Nessa linha, o foco do PropBank é a identificação dos argumentos dos verbos e seu resultado é a anotação de um cópús com uma classe de relações predicado-argumento: os papéis semânticos. O projeto PropBank adicionou uma camada de informação verbo-argumento às estruturas sintáticas do subcópus financeiro do Penn Treebank² ([Marcus et al., 1993](#)), totalizando 71.109 instâncias de anotação e 3.185 verbos alvos distintos. Além desse recurso, um léxico de verbos e seus sentidos foi construído para auxiliar a anotação manual de papéis semânticos e está disponível para consulta³.

Na tentativa de definir um conjunto neutro de papéis semânticos, o PropBank anota os argumentos semânticos de um verbo com etiquetas numeradas de zero a cinco (ArgN), não sendo obrigatória a ocorrência de todas as etiquetas em uma mesma sentença (o objeto do verbo pode não estar realizado na sentença, por exemplo). Sendo assim, a APS independente das teorias que definem um dado conjunto de papéis. Para exemplificar, geralmente o elemento que contém características de um Agente Prototípico recebe a etiqueta Arg0, enquanto um Paciente Prototípico (Tema) recebe a etiqueta Arg1 ([Palmer et al., 2010](#)). As demais etiquetas (Arg2, Arg3, Arg4 e Arg5) representam características específicas de um verbo na sentença, não possuindo um comportamento genérico, muito menos obrigatório⁴. Nesse caso, uma consulta ao léxico de verbos e sentidos (detalhado abaixo) permite identificar quais papéis o verbo aceita e em quais circunstâncias. Outro conjunto de etiquetas ainda é utilizado, composto por etiquetas

²Disponível em <http://www.cis.upenn.edu/~treebank/>.

³Disponível em <http://verbs.colorado.edu/verb-index/>.

⁴Tais argumentos podem ser os adjuntos de uma oração.

chamadas de modificadoras⁵ (ou ArgMs), apresentadas na Tabela 1.

Etiqueta	Descrição
ArgM-Loc	Local
ArgM-Ext	Extensão
Argm-Dis	Conectivos Discursivos
ArgM-Adv	Propósito Geral
ArgM-Neg	Marcador de Negação
ArgM-Mod	Verbo Modal
ArgM-Cau	Causa
ArgM-Tmp	Tempo
ArgM-Pnc	Propósito
ArgM-Mnr	Maneira
ArgM-Dir	Direção

Tabela 1: Etiquetas ArgM definidas por [Palmer et al. \(2005\)](#).

Fonte: [Alva-Manchego \(2013\)](#).

Ao associar um conjunto de papéis semânticos ao sentido⁶ de um verbo, [Palmer et al. \(2005\)](#) definiram uma formalização chamada *roleset*, que é o conjunto de papéis semânticos válidos para um sentido do verbo. Outro construtor do recurso é o *frameset*, que é composto por um identificador único, um sentido do verbo e os papéis semânticos aceitos para tal sentido (*roleset*), juntamente com exemplos. Diferentes significados de um verbo são representados em um *framefile* (isto é, conjunto de todos os *framesets* válidos para um determinado verbo) ([Palmer et al., 2005](#)). Para exemplificar, foi selecionado um *frameset* que representa a estrutura semântica de uma regra de utilização do verbo *to kick*, nomeado “kick.01”, apresentado na Figura 1.

⁵Etiquetas modificadoras possuem um papel mais específico, refinando os argumentos ArgN com novas características.

⁶Um verbo pode possuir mais de um sentido, consequentemente mais de uma estrutura sintática pode ser válida para esse verbo. O verbo “jogar”, por exemplo, pode ser sinônimo de “arremessar” ou então de “praticar”. Logo, o verbo “jogar” pode ter mais de um conjunto de papéis semânticos válidos.

Frameset kick.01 “drive or impel with the foot”
 Arg0: Kicker
 Arg1: Thing kicked
 Arg2: Instrument (defaults to foot)
 Ex1: [ArgM-DIS But] [Arg0 two big New York banks_i] seem [Arg0 *trace*_i] to have
kicked [Arg1 those chances] [ArgM-DIR away], [ArgM-TMP for the moment], [Arg2 with
 the embarrassing failure of Citicorp and Chase Manhattan Corp. to deliver
 \$7.2 billion in bank financing for a leveraged buy-out of United Airlines parent
 UAL Corp]. (wsj_1619)
 Ex2: [Arg0 John_i] tried [Arg0 *trace*_i] to *kick* [Arg1 the football], but Mary pulled it
 away at the last moment.

Figura 1: Exemplo de um *frameset* do PropBank para o verbo *to kick*.

Fonte: Palmer et al. (2005).

2.1.2 PropBank.Br

Uma versão do PropBank foi construída para o português brasileiro, intitulada PropBank.Br⁷ (Duran et al., 2010; Duran and Aluísio, 2011, 2012). O PropBank.Br possui anotações de papéis semânticos sobre um subconjunto do córpus Bosque da Floresta Sintá(c)tica⁸, isto é, um *treebank* composto por textos do córpus CETENFolha⁹ automaticamente anotados pelo *parser* sintático PALAVRAS (Bick, 2000) e manualmente corrigido por linguistas. O PropBank.Br não possui sentenças em que ocorrem verbos auxiliares, pois como a identificação desses verbos pode ser feita automaticamente com base em um léxico de auxiliares (Duran and Aluísio, 2010), não é necessário esforço humano na anotação dessa classe de verbos. Em sua primeira versão, o PropBank.Br continha 5.831 instâncias¹⁰ anotadas e 1.023 verbos alvos distintos. O PropBank.Br encontra-se em sua segunda versão (PropBank.Br versão 1.1), contendo 5.942 instâncias anotadas e 1.025 verbos alvos distintos. O córpus PropBank.Br teve 1.396 sentidos distintos anotados, o que dá uma média de 1,36 sentido por verbo. Foi identificado que 80% dos verbos presentes no córpus apresentam apenas um sentido; 12,8% apresentam dois sentidos; 3,4% apresentam três sentidos e 3,7% apresentam quatro ou mais sentidos. Dos 1.396 sentidos criados, apenas 109 não tiveram um *roleset* equivalente

⁷Disponível em <http://nilc.icmc.usp.br/portlex/index.php/pt/projetos/propbankbr>.

⁸Disponível em <http://www.linguatca.pt/Floresta/principal.html>.

⁹Disponível em http://www.linguatca.pt/cetenfolha/index_info.html.

¹⁰Uma instância de anotação consiste na APS para um determinado predicado. Assim, uma sentença pode conter diversas instâncias de anotação.

identificado no PropBank. É interessante observar que o percentual de verbos com mais de um sentido é de 20%, exatamente igual ao observado no PropBank. Seis novos campos de anotação para as instâncias do PropBank.Br versão 1.1 são mostrados abaixo.

- **PB-roleset**: sentido do verbo no repositório do PropBank inglês, equivalente ao sentido anotado em português, que permite herdar as definições dos papéis semânticos e seus mapeamentos para os papéis semânticos e classes da VerbNet.
- **Nota**: campo utilizado esporadicamente quando é necessário fazer alguma observação sobre o sentido do verbo pelo os anotadores.
- **Nota do exemplo**: esse campo é utilizado esporadicamente para ressaltar algum elemento do exemplo.
- **Predicate lemma**: campo obrigatório na primeira ocorrência de um sentido, em que se coloca o nome do predicator, incluindo predicados complexos (idiomáticos ou não), por exemplo "abrir mão".
- **Sentido**: Sentido do verbo em uma sentença, em relação aos *frames* do recurso léxico-semântico Verbo-Brasil ([Duran et al., 2013](#)).
- **t-glosa**: campo obrigatório na primeira ocorrência de um sentido, em que se coloca uma definição clara do sentido do verbo que permite ao anotador distinguir um sentido de outro.

Entre as anotações listadas acima, as autoras acreditam que a “Sentido” contribui para APS, pois reduzir o grau de incerteza que existe quanto ao sentido do verbo na anotação de papéis semânticos.

Vale ressaltar que o PropBank.Br não possui um equilíbrio no número de instâncias para cada verbo ou mesmo para seus sentidos. Segundo [Dligach and Palmer \(2009\)](#), modelos de língua podem ser utilizados na criação de um *ranking* das instâncias dos verbos de acordo com sua probabilidade de ocorrência no *corpus*. Como as instâncias de sentidos raros são menos frequentes do que as instâncias do sentido predominante, eles assumem que deve haver uma alta concentração de instâncias de sentidos raros entre as instâncias que têm baixa probabilidade. Logo deve haver uma melhor distribuição dos

verbos alvo de anotação em um *córpus*, de modo a ter um quantidade representativa dessa instâncias raras.

2.2 Anotação de Papéis Semânticos

Em uma abordagem que realiza a anotação de papéis semânticos sobre *treebanks*¹¹, seja a anotação realizada manual ou automaticamente, são analisados os relacionamentos dos constituintes de cada sentença de um texto, em particular, como o verbo determina o comportamento dos demais constituintes. Por exemplo, podemos perguntar aos verbos de uma sentença: “Quem?”, “O quê?”, “Para quem?”, “Quando?” e “Onde?” (entre outras perguntas possíveis), e as respostas serão dadas por constituintes sintáticos da sentença: sujeito, objeto direto, objeto indireto e adjuntos adverbiais.

Desde o trabalho de [Gildea and Jurafsky \(2002\)](#), a tarefa de anotação de papéis semânticos tem sido dividida em duas etapas: (1) a identificação de argumentos do predador; (2) a classificação desses argumentos, segundo o conjunto de papéis semânticos estabelecido pelo projeto de anotação. Para a primeira tarefa, é requerido um maior conhecimento sintático sobre o posicionamento dos constituintes em uma sentença e, para a segunda, um maior conhecimento no âmbito da semântica ([Pradhan et al., 2008](#)) para decidir o papel semântico de um constituinte.

2.2.1 Anotação Manual de Papéis Semânticos

O modelo de APS proposto pelo PropBank toma como base um grande repositório de verbos, separados por sentidos. Sabe-se que, quanto mais completo for o repositório lexical e quanto melhor for a distinção que ele faz dos sentidos do predador, mais simples a tarefa se tornará para os anotadores, aumentando a probabilidade de se atingir bons índices de concordância entre eles ([Duffield et al., 2007](#); [Palmer et al., 2007](#)). A Figura 2 apresenta um modelo para APS manual definido por [Palmer et al. \(2005\)](#). O modelo proposto consiste na repetição de 3 etapas:

1. Para uma sentença, são selecionados todos os verbos alvo de anotação e são geradas “instâncias de anotação”, em que cada instâncias é composta pela árvore

¹¹*Treebank* é um *córpus* com anotação sintática e/ou semântica.

Figura 2: Modelo de APS manual definida por [Palmer et al. \(2005\)](#).

sintática de sentença, que aponta o verbo alvo da anotação.

2. Para cada instância de anotação, o anotador deve buscar o sentido do verbo alvo no repositório lexical de suporte à tarefa, que indica os papéis semânticos esperados para o sentido identificado (*roleset*). O anotador ainda dispõe de um manual de anotação, que contém diretrizes de como proceder com a anotação, com o intuito de manter uma uniformidade na anotação.
3. Anotam-se os argumentos de cada instância, com base no *roleset* identificado, marcando os constituintes sintáticos da sentença com etiquetas ArgN e/ou ArgM.
4. Repete-se os passos de 1 à 3 até que todas as instâncias tenham sido anotadas.

[Duran et al. \(2013\)](#) apresentam a criação do Verbo-Brasil, um repositório lexical para dar suporte à APS do português brasileiro, a partir de um corpus já anotado com papéis semânticos (o PropBank.Br versão I), cuja anotação não se beneficiou da etapa 2 de construção de um corpus de papéis semânticos¹². Com um repositório construído, os responsáveis pelas próximas anotações podem consultar o arquivo predicator (chamado

¹²Dessa forma, o repositório Verbo-Brasil, na sua primeira versão, possui o mesmo número de verbos que o PropBank.Br versão I.

de *framefile*) para buscar qual o sentido que coincide com o sentido que está sendo anotado (*frameset*) e ver quais os papéis semânticos previstos para esse sentido.

Por exemplo, o verbo “jogar” possui dois *framesets*, apresentados a seguir. No caso em que um anotador teria que anotar uma instância contendo “jogar” como verbo alvo, caberia ao anotador identificar qual o sentido do verbo que melhor caracteriza a sentença de anotação, ou seja, qual o *frameset* correto. Para a sentença “João joga futebol com seus amigos”, é correto afirmar que o *frameset* “jogar.01” representa o sentido correto do verbo “jogar”.

Frameset **jogar.01** “participar de um jogo”

Arg0: Quem joga

Arg1: O que é jogado

Ex: [O menino_{Arg0}] jogou [futebol_{Arg1}] [ontem de tarde_{ArgM-Tmp}]

Frameset **jogar.02** “arremessar”

Arg0: Quem arremessa

Arg1: Objeto arremessado

Arg2: Local em que o objeto foi arremessado

Ex: [A pedra_{Arg1}] foi jogada [pelo menino_{Arg0}] [na janela do vizinho_{Arg2}]

Na criação do PropBank, a APS manual utilizou árvores sintáticas extraídas do Penn Treebank, ao invés de sentenças sem algum pré-processamento. A abstração da sentença na forma de uma árvore sintática aumenta a precisão do sistema de APS automático, devido às *features* que se pode extrair dessa árvore (As *features* comumente utilizadas são apresentadas na Seção 2.3). Um exemplo de representação de sentença no formato de árvore sintática pode ser visto na Figura 3.

2.2.2 Anotação Automática de Papéis Semânticos

Segundo Palmer et al. (2010), a tarefa de APS pode ser definida na seguinte frase:

“Considerando um predador e cada constituinte de uma árvore sintática, selecionam-se, de um conjunto pré-definido, as etiquetas semânticas para

Figura 3: Árvore sintática para a sentença “Ele receberá o valor à vista após 30 dias.”.

Fonte: Alva-Manchego (2013).

cada constituinte em relação ao verbo em foco, em uma dada sentença.” (p. 31).

Assim, tomando a especificação acima da tarefa de anotação de papéis semânticos, temos, para a sentença “João quebrou a janela com a pedra” (Alva-Manchego, 2013):

1. Identificação do predicado (nesse trabalho de mestrado, o verbo):
João [**quebrou_V**] a janela com a pedra.
2. Identificação de Argumentos: geralmente, extraem-se atributos da árvore sintática da sentença para identificar quais grupos de palavras não podem ser divididos, pois representam um único constituinte e, portanto, não podem ser candidatos individuais a argumentos (sintagmas podem ser divididos; palavras não):
[João_{ARG}] [**quebrou_V**] [**a janela_{ARG}**] [**com a pedra_{ARG}**].
3. Classificação de Argumentos: consiste em extrair atributos dos candidatos a argumentos (por exemplo, caminho na árvore sintática entre o argumento e o verbo), identificados no passo anterior, para determinar qual etiqueta de papel semântico correspondente será atribuída a cada um deles:
[João_{ARG0}] [**quebrou_V**] [**a janela_{ARG1}**] [**com a pedra_{ARG2}**].

Usualmente, os sistemas de anotação de papéis semânticos são treinados com conjuntos de dados que não possuem erros, denominados *gold-standard*. Comumente, a

entrada de um sistema de APS é a árvore sintática da sentença, logo um conjunto *gold-standard* possui árvores sintáticas corretas para um conjunto de sentenças. Uma alternativa ao uso do conjunto *gold-standard* é a utilização de *parsers* sintáticos sem pós-processamento das árvores, porém essas ferramentas estão sujeitos a erros que irão se propagar durante todo o processo de anotação de papéis semânticos. Apesar de ser comum utilizar conjuntos *gold-standard* para treinamento e teste e árvores sintáticas não revisadas somente para teste, estudos não foram realizados no PB para avaliar o impacto do treinamento de um sistema de APS sobre árvores sintáticas não revisadas, nem sobre a avaliação nesse conjunto de dados.

2.3 As Principais *Features* Utilizadas em APS Automática via AM

Para anotar automaticamente um texto com papéis semânticos, a maioria dos sistemas atuais emprega técnicas de Aprendizagem de Máquina (AM) na realização da tarefa, uma vez que esta pode ser considerada como um problema de classificação: para dado predador e um constituinte sintático, a tarefa é selecionar, de um conjunto pré-definido de papéis semânticos, o papel do constituinte que melhor emerge de sua relação com o verbo (Palmer et al., 2010).

Duffield et al. (2007) e Palmer et al. (2007) também reforçam que o bom desempenho de um sistema de APS automática é obtido com distinções adequadas de sentidos, isto é, para a anotação manual do corpus deve-se ter uma granularidade ideal. A divisão de sentidos (granularidade) não pode nem ser muito genérica, pois não faria distinção, nem muito específica, pois poderia ser difícil de distinguir, tanto por anotadores humanos quanto por classificadores automáticos. Por isso, a escolha de *features* adequadas, ou seja, aquelas que distinguem bem os argumentos e o sentido de um verbo, comumente acarreta em melhores resultados.

O sistema de Gildea and Jurafsky (2002) foi o primeiro em empregar técnicas estatísticas ao extrair informações de um grande corpus de sentenças anotadas e utilizá-las na tarefa de APS. Uma de suas contribuições mais importantes foi o conjunto de *features* por eles empregado, frequentemente utilizado como conjunto base por sistemas de APS desenvolvidos posteriormente. Um bom conjunto de *features* dá condição à classifica-

ção dos argumentos de um verbo, possibilitando uma anotação de papéis semânticos automática de boa qualidade. A seguir, apresentamos as *features* utilizadas por [Gildea and Jurafsky \(2002\)](#):

- **Tipo de Sintagma:** indica a categoria sintática do sintagma que expressa o papel semântico. Dada a árvore sintática, encontra-se o constituinte que abrange o mesmo conjunto de palavras que cada *frame element*¹³ anotado e a etiqueta não terminal do constituinte é tomada como o tipo de sintagma.
- **Categoria Principal:** indica se determinado sintagma nominal (SN) é sujeito ou objeto direto do verbo. De acordo com a anotação sintática do Penn Treebank, os nós do tipo SN que se encontram embaixo de nós S (raíz da árvore) são geralmente sujeitos gramaticais, e nós SN embaixo de nós Sintagma Verbal (SV) são geralmente objetos. Assim, sobe-se pela árvore sintática desde o constituinte que corresponde a um *frame element* até encontrar um nó S ou SV, o que determina o valor deste atributo.
- **Caminho na árvore sintática:** Representado por uma cadeia de nós não-terminais da árvore sintática, unidos por símbolos que indicam movimentos ascendentes ou descendentes na árvore. Essa *feature* contém o caminho do verbo até o constituinte a ser anotado.
- **Posição:** Indica se o constituinte a ser anotado está à direita ou à esquerda do verbo alvo de anotação. Sujeitos geralmente aparecerão antes do verbo e objetos depois.
- **Voz:** Indica se o verbo está na forma ativa ou passiva. Essa identificação é feita por meio da utilização de padrões definidos pelos autores.
- **Núcleo do Sintagma:** Indica o núcleo do sintagma do constituinte a ser anotado. No caso dos sintagmas nominais, fornece informação adicional que pode ser usada como restrição de seleção. Vale mencionar que, no caso dos sintagmas preposicionais, o núcleo é a preposição.

¹³*Frame elements* são os tipos de entidades que podem participar no *frameset* e que podem ser considerados como papéis semânticos mais refinados ([Alva-Manchego, 2013](#)).

- **Subcategorização:** indica a regra da estrutura do sintagma que expande o nó pai do predicado evocador na árvore sintática, para diferenciar usos transitivos e intransitivos do verbo. Este atributo só é usado para predicados verbais.

Além das *features* propostas por [Gildea and Jurafsky \(2002\)](#), outras *features* são comumente utilizadas por sistemas mais recentes e são listadas por [Palmer et al. \(2010\)](#):

- **Ordem do argumento:** *feature* introduzida por [Fleischman et al. \(2003\)](#), é utilizada para casos em que não se utiliza a árvore sintática para a tarefa de APS. Indica a posição do constituinte em relação aos argumentos do verbo.
- **Regra anterior:** *feature* introduzida por [Fleischman et al. \(2003\)](#), indica o papel semântico atribuído ao argumento anterior ao em análise, com a finalidade de auxiliar a inferência do papel semântico ao argumento atual. Cria uma dependência entre a anotação dos argumentos, estratégia aprovada por [Toutanova et al. \(2008\)](#).
- **Part of speech do núcleo:** Devido ao fato das categorias *part of speech* (PoS) do Penn Treebank distinguirem singular de plural e nomes próprios de substantivos comuns, ter o conhecimento sobre o *part of speech* do núcleo de um SN pode refinar qual é o tipo do sintagma nominal. Essa *feature* foi proposta por [Surdeanu et al. \(2003\)](#).
- **Entidades nomeadas em constituintes:** Introduzida por [Pradhan et al. \(2005\)](#), essa *feature* identifica entidades como: Pessoa, Organização, Localidade, Porcentagem, Dinheiro, Tempo e Data. Trata a esparsidade dos dados devido à baixa ocorrência desse tipo de entidades no corpus.

2.4 Considerações Finais

Nesta seção foi apresentada a teoria de papéis semânticos de [Palmer et al. \(2005\)](#), que esse trabalho de mestrado optou por seguir. Essa teoria atualmente é a mais difundida devido ao seu conjunto de papéis semânticos ser o mais enxuto, o que propicia melhor performance para métodos de AM na APS automática. Também foi apresentado

o modelo de anotação manual de papéis semânticos utilizado no projeto PropBank, contrastando com o modelo utilizado no projeto PropBank.Br. Ainda foi apresentado o modelo de anotação automática de papéis semânticos comumente utilizado na literatura, proposto por [Palmer et al. \(2005\)](#). Por fim, foi apresentado um extenso conjunto de *features* que é comumente utilizado pela literatura na tarefa de anotação de papéis semânticos.

É interessante observar que a maioria dos sistemas atuais de APS utilizam métodos de AM e processam as árvores sintáticas das sentenças. No entanto, poucos trabalhos relatam as experiências com anotação sobre árvores sintáticas não revisadas, ou sejam, árvores que contêm erros. Acreditamos que esses erros impactam a performance da anotação humana, assim como observado pela literatura quanto ao impacto na anotação automática de papéis semânticos.

Trabalhos Relacionados

Para a área de AM, formalmente, 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 o seu desempenho em tarefas de T , medido por P , melhora com a experiência E (Mitchell, 1997). Assim, um sistema de aprendizado tem a função de analisar as informações de E e generalizá-las para a extração de novos conhecimentos (Monard and Baranauskas, 2003).

Um sistema baseado em AM pode seguir três paradigmas: supervisionado, semi-supervisionado ou não-supervisionado. Em um modelo de **aprendizagem supervisionada**, temos um conjunto de categorias etiquetadas conhecidas e um conjunto de dados de treinamento que serão etiquetados entre as categorias conhecidas, ou seja, há conhecimento prévio ao treinamento (Jannach et al., 2010). A esse tipo de problema se dá o nome de classificação.

No modelo de **aprendizagem não-supervisionada**¹, não se tem um agente externo para acompanhar o processo de aprendizado, ou seja, não se tem um conjunto prévio de categorias etiquetadas (Ricci et al., 2011). Nesse caso, os problemas abordados são os de agrupamento (*clustering*), no qual as instâncias são separadas de acordo com

¹Sistemas não-supervisionados também são conhecidos como auto-supervisionados (Ferneda, 2003).

alguma medida de similaridade. O significado do agrupamento de dados é abstrato para a máquina, necessitando da análise posterior de um especialista.

O modelo de **aprendizagem semissupervisionada** emprega estratégias que unificam o aprendizado supervisionado e o não-supervisionado, incluindo características de ambos os paradigmas. Esse tipo de aprendizado pode ter como objetivo treinar um classificador com parte dos dados anotados e, com a parte não anotada, utilizar aprendizado não-supervisionado. Assim, é possível obter um melhor classificador do que se fosse treinado apenas com poucos dados anotados. Nesse cenário, assume-se tipicamente que existem muito mais dados não anotados do que anotados.

A seguir, serão apresentados os trabalhos pesquisados relacionados a esse tema de mestrado, divididos nos três paradigmas de aprendizagem de máquina recém apresentadas.

3.1 Sistemas Supervisionados

3.1.1 Trabalhos Pioneiros em APS

Trabalho de [Gildea and Jurafsky \(2002\)](#)

[Gildea and Jurafsky \(2002\)](#) foram os primeiros a idealizar a tarefa de anotação automática de papéis semânticos. Segundo [Johansson \(2008\)](#), a forma como os autores lidaram com a tarefa de APS fez com que seu trabalho se tornasse uma referência para a área, justificando a quantidade de trabalhos subsequentes, como [Gildea and Palmer \(2002\)](#); [Gildea and Hockenmaier \(2003\)](#); [Surdeanu et al. \(2003\)](#); [Palmer et al. \(2005\)](#); [Yi et al. \(2007\)](#); [Toutanova et al. \(2008\)](#); [Pradhan et al. \(2008\)](#).

O corpus utilizado por [Gildea and Jurafsky \(2002\)](#) foi a FrameNet ([Baker et al., 1998](#)), que consiste de 50 mil sentenças, obtidas do *British National Corpus*², manualmente anotadas com papéis semânticos em estruturas intituladas *frame elements*³. O conjunto de *features* utilizadas já foi apresentado na Seção 2.3. Os autores utilizaram o

²Disponível em <http://www.natcorp.ox.ac.uk/>.

³*Frame element* é o papel semântico definido especificamente para um *frame*. Também chamado de “unidade básica” de um *frame*.

parser sintático de [Collins \(1999\)](#) para gerar árvores sintáticas para as sentenças da FrameNet e então realizar experimentos de classificação: em um primeiro cenário, tendo em mãos os nós sintáticos que são argumentos do predicado, os autores apenas classificarão os papéis semânticos das constituintes já conhecidas; em um segundo cenário, os autores identificaram as constituintes e classificaram os papéis semânticos. Os resultados são apresentados na Tabela 2.

Tarefa	F1	Acurácia
Classificação	-	82,0%
Identificação e Classificação	62,9	-

Tabela 2: Resultados do sistema supervisionado de [Gildea and Jurafsky \(2002\)](#) para testes com *parser* sintático automático.

Os resultados desse trabalho estão abaixo dos valores dos sistemas atuais para a língua inglesa, como o de [Toutanova et al. \(2008\)](#), mas são importantes graças aos esforços de seus idealizadores que iniciaram a área de pesquisa em anotação de papéis semânticos.

Trabalho de [Palmer et al. \(2005\)](#)

[Palmer et al. \(2005\)](#), além de desenvolverem o *corpus* PropBank (cujo modelo de anotação é comumente utilizado pela comunidade de PLN para a atividade de APS), treinaram um anotador de papéis semânticos utilizando esse *corpus* e verificaram a qualidade da identificação e classificação de argumentos. Os autores utilizaram o sistema supervisionado desenvolvido por [Gildea and Jurafsky \(2002\)](#) e o *parser* sintático automático de [Collins \(1999\)](#). Os autores realizaram testes com árvores sintáticas geradas por *parser* automático, com conjunto *gold-standard* (árvores sintáticas geradas pelo *parser*, porém com pós-processamento para correção de erros) e com a inserção de traços semânticos obtidos do Penn Treebank (por exemplo, explicitar um elemento e/ou reposicionar constituintes para a ordem canônica “sujeito-verbo-objeto”) ao conjunto *gold-standard*. Os resultados dos testes são apresentados na Tabela 3. Todos os valores são para a tarefa conjunto entre identificação das constituintes e classificação dos papéis

semânticos.

Cenário	Precisão	Cobertura	F1	Acurácia
<i>Parser automático</i>	68,6%	57,8%	62,7	79,9%
<i>Gold-standard</i>	74,3%	66,4%	70,1	82,0%
<i>Gold-standard_{traces}</i>	80,6%	71,6%	75,8	-

Tabela 3: Resultados do sistema de [Palmer et al. \(2005\)](#) para teste com dados passados por *parser* sintático automático, dados *gold-standard* e *gold-standard* com traços.

Os resultados divulgados não apresentam valores para a acurácia do sistema *gold-standard* com a utilização de traços semânticos. É interessante ressaltar que a adição de traços semânticos incrementou a precisão e cobertura da tarefa conjunta do sistema em 5%.

3.1.2 Trabalhos para o Português

Trabalho de [Alva-Manchego \(2013\)](#)

O trabalho de [Alva-Manchego \(2013\)](#), que usa o corpus PropBank.Br (em sua primeira versão) para treinamento, empregou uma abordagem supervisionada para anotação automática de papéis semânticos em sentenças escritas em português do Brasil. O conjunto de *features* utilizado pelo autor é listado a seguir:

- **Caminho:** Caminho na árvore sintática entre o verbo e o argumento.
- **Caminho Parcial:** Caminho entre o verbo e o menor antepassado em comum com o argumento.
- **Contexto do Predicado:** Uma palavra antes e uma depois do verbo com suas etiquetas PoS.
- **Distância em Constituintes na Árvore:** Número de constituintes encontrados no caminho sintático entre o verbo e o argumento.
- **NEG:** Indica se existe indicador de negação na sentença.

- **Núcleo, Lema do Núcleo, PoS do Núcleo** do constituinte a ser anotado.
- **Núcleo do Sintagma Preposicional:** Somente se o constituinte é um Sintagma Preposicional (SP).
- **Número de Orações:** Quantidade de orações no caminho sintático entre o verbo e o argumento.
- **Número de Sintagmas Verbais:** Quantidade de SV (Sintagma Verbal) no caminho sintático entre o verbo e o argumento.
- **Palavras do Constituintes:** A forma, o lema e a etiqueta PoS para as três primeiras palavras que formam o constituinte.
- **Parentes do Constituinte:** Atributos que indicam o tipo de sintagma, núcleo e PoS do núcleo do pai, irmãos esquerdo e direito do constituinte a ser anotado.
- **Posição:** Indica se o constituinte está antes ou depois do verbo.
- **Pontuação:** Sinal de pontuação a esquerda e a direita do constituinte.
- **Predicado:** Forma, lema e etiqueta PoS do verbo.
- **Primeira e Última Palavra/PoS no Constituinte:** Primeira e última palavra do constituinte juntamente com sua etiqueta PoS.
- **Bag of words** de substantivos, adjetivos e advérbios no constituinte.
- **SE na Oração do Verbo:** Indica a presença da partícula SE na oração.
- **Sequência PoS:** Cadeia formada pelas etiquetas PoS das palavras que formam o constituinte.
- **Sequência TOP:** Corresponde ao lado direito da regra que expande o nó do constituinte.
- **Subcategorização:** Regra da estrutura do sintagma que expande o nó pai do verbo.
- **Tipo de Sintagma:** Categoria sintática (SN, SP, SV, etc.) do constituinte.

- **Voz:** Indica se a oração está na voz ativa ou passiva.
- **Lema do Predicado + Caminho, Lema do Predicado + Núcleo, Lema do Predicado + Tipo de Sintagma, Voz + Posição.**

Os resultados obtidos pelo sistema supervisionado na tarefa conjunta de identificação e classificação de argumentos, utilizando o *PropBank.Br* como conjunto *gold-standard* para treinamento e também para teste, são apresentados na Tabela 4.

Tarefa	Precisão	Cobertura	F1	Acurácia
Identificação	94,9%	94,0%	94,5	-
Classificação	-	-	-	81,7%
Conjunta	80,0%	79,3%	79,6	-

Tabela 4: Resultados do sistema supervisionado de [Alva-Manchego \(2013\)](#) para teste com dados *gold-standard*.

Os autores não realizaram experimentos com a utilização de *parsers* sintáticos automáticos. O autor espera que, com a utilização de um *corpus* maior e com uma melhor distribuição de papéis semânticos, o sistema supervisionado possa obter melhores resultados futuros.

Trabalho de [Fonseca \(2013\)](#)

[Fonseca \(2013\)](#) desenvolveu em seu trabalho de mestrado um sistema para anotação de papéis semânticos para a língua portuguesa brasileira⁴, evitando a dependência de ferramentas externas de PLN, como a de um *parser* sintático. Nesse trabalho, foi seguida a arquitetura do sistema SENNA ([Collobert and Weston, 2008](#); [Collobert et al., 2011](#)), que obteve bons resultados em PLN para a língua inglesa. A arquitetura do SENNA modela dados em um espaço vetorial e utiliza redes neural, com uma camada de convolução, para treinamento supervisionado. Para tanto, um grande *corpus* de domínio geral foi compilado a partir da Wikipédia⁵ e do *corpus* PLN-BR ([Bruckschen](#)

⁴Disponível em <http://www.nilc.icmc.usp.br/nlpnet>.

⁵Disponível em <http://pt.wikipedia.org/>.

et al., 2008), sobre o qual foi treinado um modelo de língua em espaço vetorial (*word embeddings*) para representação de sentenças de entrada para o sistema anotador de papéis semânticos. O corpus utilizado para treinamento do anotador foi o PropBank.Br. O autor adotou uma rede neural com utilização de uma camada de convolução para a implementação do sistema de APS. Os resultados de seu trabalho podem ser vistos na Tabela 5.

Segundo o autor, o gargalo da qualidade do sistema encontra-se na fase de identificação de argumentos e que um aumento da quantidade de dados de treino melhoraria os resultados do anotador. Na medida em que o sistema aprende melhor a reconhecer os limites de constituintes sintáticos, é facilitada a anotação dos argumentos do verbo (Fonseca, 2013). Testes em que os argumentos ArgN e ArgM fossem anotados separadamente poderiam avaliar se o desempenho da anotação dessas classes é distinto.

Tarefa	Precisão	Cobertura	F1	Acurácia
Identificação	80,1%	79,9%	79,1	-
Classificação	-	-	-	88,1%
Conjunta	68,9%	67,0%	68,0	-

Tabela 5: Resultados do sistema supervisionado de Fonseca (2013) para teste com dados *gold-standard*.

Trabalho de Sequeira et al. (2012)

Sequeira et al. (2012) realizaram experimentos preliminares na classificação de papéis semânticos de sujeito (Arg0) e objeto (Arg1) para o português brasileiro, utilizando as *features*: *bag-of-words*, entidades nomeadas, etiquetas sintáticas e morfossintáticas. Os autores extraíram *features* sobre trigramas centrados no verbo alvo de anotação. O corpus utilizado pelos autores foi uma seleção do corpus CETEMPúblico⁶, denominado de BosqueUE, composta por 4.416 sentença.

Os autores realizaram experimentos com os métodos SVMCM (Support Vector Machine com Conditional Markov Models) e CRF (Conditional Random Fields). Ambos os

⁶Disponível em <http://www.linguatca.pt/cetempublico/>.

métodos estão disponíveis no pacote MinorThird (Cohen, 2004), utilizado pelos autores. O MinorThird é um *software open source* que desempenha tarefas como classificação de textos e extração de entidades nomeadas.

A avaliação dos métodos para o corpus do português foi feita utilizando a métrica *10-fold cross-validation*. Para fins de comparação com a anotação de papéis semânticos para o inglês, Sequeira et al. (2012) utilizaram os corpus de treinamento e teste do CoNLL-2004. A Tabela 6 apresenta a frequência de verbos e papéis semânticos avaliados nos corpus analisados. A Tabela 7 apresenta os resultados na classificação de papéis semânticos para o corpus BosqueUE e a Tabela 8 apresenta os resultados da classificação de papéis semânticos para o conjunto de teste do CoNLL-2004. Percebe-se que o método que utiliza *Support Vector Machine* é levemente superior ao que utiliza CRF para ambas as línguas. Percebe-se também que os autores não atingiram resultados similares aos do inglês, muito menos aos dos demais trabalhos desenvolvidos para o português, apresentados nesta seção. Sequeira et al. (2012) atribuem o desempenho inferior de seu sistema ao volume restrito de dados do BosqueUE, mas, ao comparar a quantidade de *features* utilizadas por Sequeira et al. (2012) e Alva-Manchego (2013), é possível supor que as *features* utilizadas por Sequeira et al. (2012) não são suficientes para a aprendizagem de máquina.

	BosqueUE	CoNLL-2004	
Papel Semântico		Treinamento	Teste
Verbo	7.268	19.098	3.627
Arg0	4.673	12.709	1.671
Arg1	3.802	18.046	3.429

Tabela 6: Frequência de ocorrência de verbos, sujeitos e objetos nos corpus analisados por Sequeira et al. (2012).

Categoria	SVMCMM			CRF		
	Precisão	Cobertura	F1	Precisão	Cobertura	F1
Verbo	58,8%	47,7%	52,7	64,8%	47,7%	54,9
Arg0	38,8%	25,9%	31,1	43,4%	23,7%	30,6
Arg1	26,9%	14,7%	19,0	31,7%	11,7%	17,1

Tabela 7: Resultados da anotação de papéis semânticos de [Sequeira et al. \(2012\)](#) para o corpus BosqueUE.

Categoria	SVMCMM			CRF		
	Precisão	Cobertura	F1	Precisão	Cobertura	F1
Verbo	85,0%	82,3%	83,6	84,2%	80,5%	82,3
Arg0	59,9%	46,4%	52,3	69,9%	46,3%	55,7
Arg1	37,2%	17,0%	23,4	41,4%	15,1%	22,1

Tabela 8: Resultados da anotação de papéis semânticos de [Sequeira et al. \(2012\)](#) para o conjunto de teste do CoNLL-2004.

3.1.3 Estado da Arte em APS

Trabalho de ([Toutanova et al., 2008](#))

[Toutanova et al. \(2008\)](#) comentam em seu trabalho sobre a pouca utilização de desambiguação de verbos nos sistemas de anotação de papéis semânticos. Os autores argumentam, baseando-se em [Roland \(2002\)](#), que a identificação do sentido do verbo em uma sentença elimina uma variável importante na anotação de papéis semânticos e que a não identificação do sentido do verbo faz com que os argumentos do verbo possuam maior interdependência, dificultando a tomada de decisão sobre a anotação de papéis semânticos.

[Toutanova et al. \(2008\)](#) apresentam um modelo de APS que modela o conjunto de argumentos semânticos de determinado verbo como uma estrutura conjunta, cujos constituintes possuem fortes dependências entre si. O algoritmo de atribuição conjunta

modela dependências entre as etiquetas dos constituintes e entre cada etiqueta e os atributos de entrada dos outros constituintes. Esse algoritmo emprega uma abordagem de *re-ranking* (Collins and Koo, 2000), que seleciona as n atribuições conjuntas⁷ mais prováveis de acordo com um modelo de anotação local⁸.

Para a realização do treinamento, os autores utilizaram o *cópus* PropBank como conjunto de dados *gold-standard* e, para avaliar o sistema sobre um ambiente real, os autores utilizaram o *parser* sintático automático de Charniak (2000) sobre o *cópus* disponibilizado no CoNLL-2005 *shared tasks*. Os resultados dos testes realizados são apresentados na Tabela 9. É importante destacar que, apesar da queda de desempenho na anotação de papéis semânticos com a utilização de *parser* automático, os resultados obtidos com dados *gold-standard* ainda estão no estado da arte para a tarefa de APS.

Tarefa	Gold-standard		Parser automático	
	F1	Acurácia	F1	Acurácia
Identificação	95,0	-	83,4	-
Classificação	-	91,4%	-	92,0%
Conjunta	91,2	-	80,0	-

Tabela 9: Resultados do sistema supervisionado de Toutanova et al. (2008) para teste com dados *gold-standard* e com *parser* sintático automático.

3.1.4 Novidades na área: APS Intergêneros

Trabalho de (Clercq et al., 2012)

Clercq et al. (2012) verificaram os benefícios que a utilização de um *cópus* multi-gênero provê para a APS. Os gêneros escolhidos foram: administrativo (ADM), *auto-*

⁷Considera interdependência entre as constituintes.

⁸Considera que as constituintes são independentes entre si.

*cue*⁹ (AUTO), comunicação externa¹⁰ (EXT), textos instrucionais¹¹ (INST), jornalístico (JOUR) e da *wikipedia* holandesa (WIKI). Para tanto, os autores utilizaram o *cópus* SoNaR, composto por 500.850 palavras em 29.418 sentenças levemente desbalanceado entre os gêneros e o sistema anotador de papéis semânticos de Stevens et al. (2007). O conjunto de *features* utilizado é listado a seguir:

- **Lema:** Lema do verbo.
- **PoS:** Etiqueta PoS do argumento.
- **Voz** da sentença (ativa ou passiva).
- **Categoria** do candidato a argumento (substantivo, pronome, etc.).
- **Dependência:** Relação de dependência entre o candidato a argumento e o verbo (sujeito, modificador, etc.).
- **Posição:** Indica se o candidato a argumento está posicionando antes ou depois do verbo.
- **Lema do pai** sintático do candidato a argumento juntamente com seu PoS.
- **Primeira e última palavra:** Se um argumento é composto por mais de uma palavra, utiliza-se a primeira e última palavra, juntamente com seus PoS.
- **Relação de dependência** dos elementos da cadeia que liga cada candidato a argumento com os elementos à esquerda e à direita.
- **Relação de constituinte** dos elementos da cadeia que liga cada candidato a argumento com os elementos à esquerda e à direita.
- **Rótulo de constituinte** juntamente com o de dependência do candidato a argumento.

⁹O gênero *autocue* pode ser exemplificado como os textos apresentados na televisão para que o apresentador, ou âncora de jornal, leia.

¹⁰O gênero *comunicação externa* é composto por materiais de *websites*, comunicados da imprensa e cartas.

¹¹O gênero instrutivo é composto por manuais, bulas de medicamentos e descrição de procedimentos.

Os testes realizados por Clercq et al. (2012) utilizam *10-fold cross-validation* apenas para a fase de classificação de argumentos. Os autores argumentam que, pelo fato de terem disponível um conjunto de dados *gold-standard* (córpus SoNaR), a identificação de argumentos se torna uma tarefa simples, podendo concentrar seus experimentos na tarefa de classificação de argumentos.

Na Tabela 10, verificamos que o resultado do anotador de papéis semânticos treinado em cinco gêneros conjuntos e testado em um sexto (*Out-of-domain* (45k)) teve resultados pouco inferiores quando comparado ao teste em um único gênero (*In-domain* (45k)). Já o experimento em conjunto com os seis gêneros, tanto para treinamento como para teste (*In & Out-of-domain* (45k)), teve resultados superiores aos dos dois primeiros testes, mostrando o potencial em se transferir conhecimento entre gêneros.

No primeiro experimento apresentado no conjunto de testes da Tabela 11, os autores utilizaram 250 mil palavras, 50 mil para cada gênero. Já no segundo experimento, o gênero em teste recebeu um reforço com a adição de 45 mil palavras. Verifica-se que os resultados do teste *Out-of-domain* (250k) foram superiores ao antecessor utilizando apenas 45 mil palavras (*Out-of-domain* (45k)) e pode-se até comparar ao *In & Out-of-domain* (45k), com performance levemente inferior. Os melhores resultados foram obtidos no teste *In & Out-of-domain* (295k).

Os experimentos mostraram que é necessária a utilização de um largo conjunto de dados para o treinamento de um sistema de APS, mas que a inclusão de material específico de um gênero é crucial para otimizar a anotação de papéis semânticos. Testes também revelaram a importância de incluir textos de mais de um domínio quando o objetivo é ter um APS robusto.

Gênero	<i>In-domain</i> (45k)			<i>Out-of-domain</i> (45k)			<i>In & Out-of-domain</i> (45k)		
	Precisão	Cobertura	F1	Precisão	Cobertura	F1	Precisão	Cobertura	F1
ADM	73,2%	72,6%	72,9	70,3%	70,0%	70,1	74,0%	73,9%	74,0
AUTO	72,6%	72,2%	72,4	70,2%	68,3%	69,2	73,1%	72,2%	72,6
EXT	72,1%	71,1%	71,6	73,3%	72,5%	72,9	73,3%	72,5%	72,9
INST	77,5%	77,1%	77,3	73,5%	73,1%	73,3	77,9%	77,6%	77,7
JOUR	72,1%	71,3%	71,7	71,5%	70,6%	71,1	73,0%	72,3%	72,6
WIKI	73,0%	71,8%	72,4	71,0%	69,6%	70,3	73,6%	72,3%	73,0

Tabela 10: Resultados de [Clercq et al. \(2012\)](#) para a classificação de argumentos, com utilização de 45 mil palavras por gênero.

Gênero	<i>Out-of-domain</i> (250k)			<i>In & Out-of-domain</i> (295k)		
	Precisão	Cobertura	F1	Precisão	Cobertura	F1
ADM	72,5%	72,7%	72,6	74,6%	74,7%	74,7
AUTO	73,1%	72,0%	72,5	74,6%	73,7%	74,2
EXT	73,5%	73,0%	73,3	74,5%	73,7%	74,1
INST	75,6%	76,0%	75,8	79,2%	79,2%	79,2
JOUR	73,4%	72,9%	73,2	74,4%	73,9%	74,2
WIKI	74,5%	73,7%	74,1	75,8%	75,0%	75,4

Tabela 11: Resultados de [Clercq et al. \(2012\)](#) para a classificação de argumentos, utilizando 50 mil palavras por gênero para o primeiro caso e 50 mil palavras somadas às 45 mil do gênero em teste para o segundo caso.

3.2 Sistemas Não-Supervisionados

3.2.1 Abordagem Bilíngue: Inglês-Alemão ([Titov and Klementiev, 2012b](#))

[Titov and Klementiev \(2012b\)](#) desenvolveram um trabalho com o objetivo de demonstrar que conjuntos paralelos de dados podem ser úteis em APS, via aprendizado não-supervisionado, utilizando como exemplo o par de línguas inglês-alemão. A principal motivação para o desenvolvimento desse trabalho foi a falta de um conjunto rele-

vante de dados anotados para a maioria das línguas, o que impossibilita o aprendizado supervisionado de qualidade.

Esse trabalho fez uso do sistema bayesino não supervisionado monolíngue, em estado da arte, de [Titov and Klementiev \(2012a\)](#), que possui 83,9 de F1 para a tarefa de agrupamento de argumentos¹², e o foco foi na melhoria da performance desse sistema utilizando características multilíngues. Esse trabalho é o primeiro a considerar características multilíngues na anotação de papéis semânticos, por meio da abordagem não-supervisionada de aprendizagem de máquina.

Devido ao fato de não haver dados anotados sintaticamente para muitas línguas, foi gerado automaticamente um grafo de dependências sintáticas para ambas as línguas alvo. A anotação de argumentos semânticos foi feita utilizando a estrutura por eles definida como *argument keys*. Para cada argumento de uma sentença, a estrutura *argument keys* é criada, composta por:

- **Voz** ativa ou passiva do verbo.
- **Posição** do argumento relativa ao predicado.
- **Relação sintática** do pai do argumento.
- **Preposição** utilizada para a realização do argumento.

A anotação de papéis semânticos no sistema não-supervisionado foi tratada por meio do agrupamento de *argument keys*, a fim de encontrar centros que representassem o comportamento dos rótulos dos argumentos de um verbo.

Os experimentos realizados nesse trabalho utilizaram a seção inglês-alemão do corpus multilíngue-paralelo Europarl v6 ([Koehn, 2005](#)) e os corpus WSJ para o inglês e SALSA ([Burchardt et al., 2006](#)) para o alemão. Para a fase de validação, foram utilizados os conjuntos de treinamento do CoNLL-2009 *shared tasks*.

Na Tabela 12 são apresentados os resultados obtidos por [Titov and Klementiev \(2012b\)](#) em seus testes. Os dados MonoBayes são os que foram anotados semanticamente, independente da outra língua, ou seja, sem influência multilíngue. Já os dados

¹²Utiliza-se o termo “agrupamento” ao invés de “classificação” devido ao sistema ser não-supervisionado.

Sistema	Inglês			Alemão		
	PU	CO	F1	PU	CO	F1
MonoBayes	87,5%	80,1%	83,6	86,8%	75,7%	80,9
MultiBayes	86,8%	80,7%	83,7	85,0%	80,6%	82,7
SyntF	81,5%	79,4%	80,4	83,1%	79,3%	81,2

Tabela 12: Resultados do sistema de [Titov and Klementiev \(2012b\)](#).

MultiBayes representam a anotação bidirecional das duas línguas. O SyntF (*syntactic function baseline*) é um experimento *baseline* para comparação que realiza o agrupamento dos argumentos do predicado apenas com as relações que assumem com seus pais sintáticos. Os autores acreditam que a baixa expressividade das estruturas das *argument keys* limitam a qualidade da identificação de papéis semânticos em um ambiente multilíngue. Uma alternativa seria melhorar a representação das *argument keys* com um modelo mais expressivo de *frames* sintáticos e da relação sintaxe-semântica existente entre os constituintes. Apesar dos autores considerarem fracos os resultados apresentados na Tabela 12, eles acreditam que se pode melhorar o sistema do estado da arte de [Titov and Klementiev \(2012a\)](#) com a utilização de relações multilíngues de papéis semânticos.

3.3 Sistemas Semissupervisionado

3.3.1 Trabalho de [Alva-Manchego \(2013\)](#)

O trabalho de [Alva-Manchego \(2013\)](#), que usa o corpus PropBank.Br versão I para treinamento, além de empregar uma abordagem supervisionada, utiliza a semissupervisionada para anotação automática de papéis semânticos em sentenças escritas do português do Brasil. No sistema semissupervisionado de [Alva-Manchego \(2013\)](#), empregou-se o algoritmo *self-training* com modelos de Regressão Logística (ou Máxima Entropia) e medidas de similaridade entre os constituintes das sentenças para realizar o aprendizado semissupervisionado. O autor não realizou experimentos com a utilização de

parsers sintáticos automáticos. Sua hipótese era a de que a abordagem semissupervisionada superaria o desempenho da supervisionada e, de fato, isso se concretizou, entretanto com uma margem muito pequena.

Os resultados atingidos pelo seu anotador semissupervisionado de papéis semânticos superaram o de seu anotador supervisionado em 0,9 pontos de F1 (verificar resultados apresentados na Tabela 4 e Tabela 13). Não foi verificado se há diferença estatística significativa entre os resultados dos sistemas supervisionado e semissupervisionado desenvolvidos.

Tarefa	Precisão	Cobertura	F1	Acurácia
Identificação	94,4%	94,0%	94,2	-
Classificação	-	-	-	83,2%
Conjunta	80,7%	80,4%	80,5	-

Tabela 13: Resultados do sistema semissupervisionado de [Alva-Manchego \(2013\)](#) para teste com dados *gold-standard*.

3.3.2 Trabalho de ([Titov and Klementiev, 2012c](#))

[Titov and Klementiev \(2012c\)](#) propuseram um modelo semissupervisionado com o intuito de mostrar que, até certa quantidade de instâncias de treinamento, a aprendizagem semissupervisionada possui melhores resultados do que a supervisionada. Algumas considerações são tomadas sobre o agrupamento, como a identificação de uma granularidade razoável para os dados anotados e a limitação da influência dos dados não anotados na granularidade do agrupamento.

Os autores avaliaram o sistema semissupervisionado proposto utilizando a distribuição CoNLL 2009 ([Hajič et al., 2009](#)) do cópús Penn Treebank. O cópús foi dividido em 20 mil sentenças para treinamento e 19.279 para teste. Os dados foram anotados sintaticamente com a utilização do sistema LTH ([Johansson and Nugues, 2008](#)). Para a anotação semântica, os autores selecionaram todos os verbos não-auxiliares como predicados, descartando os verbos auxiliares. A identificação dos argumentos dos predicados foi realizada com a utilização da heurística proposta por [Lang and Lapata \(2011\)](#). A

qualidade da identificação de argumentos com os textos do CoNLL-2009 *shared tasks*, segundo a métrica F1, foi de 82,7. O sistema proposto foi comparado com o modelo supervisionado, em estado da arte, de [Johansson and Nugues \(2008\)](#).

Figura 4: Valores de F1 para a classificação de papéis semânticos nos sistemas Supervisionado, Não-Supervisionado, Semissupervisionado e SyntF ([Titov and Klementiev, 2012c](#)).

Pode-se verificar na Figura 4 que o modelo semissupervisionado superou o modelo supervisionado quando o tamanho do corpus anotado utilizado foi inferior a 300 sentenças. Ao aumentar o tamanho do corpus de treinamento, o modelo semissupervisionado não atingiu resultados semelhantes, tendo uma performance inferior ao supervisionado. A Tabela 14 apresenta os valores de *Purity*, *Collocation* e F1 para o ponto de intersecção entre a performance F1 dos sistemas semissupervisionado e supervisionado, apresentado na Figura 4, que ocorre com a utilização de 300 sentenças para o conjunto de treinamento.

Sistema	PU	CO	F1
Supervisionado de Johansson and Nugues (2008)	88,0%	84,5%	86,2
Semissupervisionado de Titov and Klementiev (2012c)	89,6%	83,0%	86,2

Tabela 14: Valores de *Purity*, *Collocation* e F1 para a utilização de 300 sentenças como conjunto de treinamento ([Titov and Klementiev, 2012c](#)).

Os autores acreditam que a abordagem semissupervisionada de APS, com a explora-

ção de dados rotulados por métodos não supervisionados, é uma promissora direção de pesquisa.

3.3.3 Trabalho de (Kozhevnikov and Titov, 2013)

Kozhevnikov and Titov (2013) desenvolveram um trabalho seguindo a linha de Titov and Klementiev (2012b), tendo como diferencial um sistema semissupervisionado e por utilizar de similaridade de estruturas semânticas entre sentenças bilíngues a fim de gerar um par de modelos de anotação de papéis semânticos. O modelo proposto foi avaliado em quatro pares de línguas: inglês-alemão, inglês-espanhol, inglês-tcheco e inglês chinês. A Figura 5 apresenta um exemplo da relação de papéis semânticos entre uma mesma sentença em línguas distintas.

Figura 5: Sentença paralela em inglês e tcheco, anotada com papéis semânticos e com ligações entre palavras sinônimas entre as línguas (Kozhevnikov and Titov, 2013).

O sistema proposto por Kozhevnikov and Titov (2013) é composto por um conjunto de classificadores binários e multiclasse, treinados por meio da utilização do *Liblinear*, uma biblioteca para classificadores lineares (Fan et al., 2008). Os corpus multilíngues utilizados foram o Europarl v6, para os alinhamentos inglês-alemão, inglês-espanhol e inglês-tcheco; e o MultiUN, para o alinhamento inglês-chinês.

Pode-se verificar, na Tabela 15, os resultados obtidos em termos de F1 do teste dos algoritmos *self-training* e *joint-probability* (que considera a interdependência entre os argumentos para melhorar a qualidade do classificador (Roth and Yih, 2005)), sobre diversos pares de línguas. Houve uma melhora significativa entre o primeiro modelo semissupervisionado, utilizando *self-training*, e o modelo utilizando *joint-probability*. Os autores acreditam que com a utilização de um maior conhecimento linguístico para casar os papéis semânticos nas sentenças bilíngues, pode-se aumentar fortemente a performance do anotador.

Línguas	<i>Self-Training</i>	<i>Joint-Probability</i>
Inglês-Tcheco	62,1	70,2
Inglês-Alemão	68,3	79,2
Inglês-Espanhol	64,5	75,4
Inglês Chinês	73,5	76,7
Tcheco-Inglês	71,6	79,7
Alemão-Inglês	71,3	77,3
Espanhol-Inglês	71,4	79,5
Chinês-Inglês	72,2	76,5

Tabela 15: Resultados dos modelos *self-training* e *joint-probability* de [Kozhevnikov and Titov \(2013\)](#), em termos de F1.

3.4 Considerações Finais

Neste capítulo foram apresentados trabalhos de APS que utilizam aprendizagem de máquina, organizados segundo as abordagens supervisionada, não-supervisionada e semi-supervisionada. A Tabela 16 apresenta um resumo dos trabalhos destacados nesta seção, com a língua alvo do trabalho, o modelo de aprendizagem de máquina empregado e os corpúis utilizados.

Trabalhos importantes para a área como o de [Gildea and Jurafsky \(2002\)](#), [Palmer et al. \(2005\)](#) e [Toutanova et al. \(2008\)](#) serviram de base para os sistemas modernos, que se baseiam no aprendizado multilíngue, trazendo características de uma língua em busca de uma melhor qualidade na APS de uma segunda língua. Enquanto as pesquisas em APS para a língua inglesa estão avançadas, as pesquisas para algumas línguas como o árabe e o sueco recém desenvolveram seus primeiros sistemas ([Hart et al., 2012](#); [Johansson et al., 2012](#)). Pesquisas em APS para língua portuguesa brasileira desenvolveram três sistemas nos últimos anos ([Sequeira et al., 2012](#); [Alva-Manchego, 2013](#); [Fonseca, 2013](#)), porém ainda com desempenho inferiores aos da língua inglesa.

Percebe-se, com a exposição dos trabalhos realizada, a não utilização do sentido do verbo na tarefa de APS e a existência de somente um trabalho que tenha retirado os

Trabalhos	Língua	Modelo	Córpus
Palmer et al. (2005)	Inglês	Supervisionado	PropBank
Toutanova et al. (2008)	Inglês	Supervisionado	PropBank
Clercq et al. (2012)	Holandês	Supervisionado	SoNaR
Titov and Klementiev (2012b)	Alemão, Inglês	Não-Supervisionado	Europarl v6, WSJ, SALSA
Titov and Klementiev (2012c)	Inglês	Semissupervisionado	CoNLL-2009
Alva-Manchego (2013)	Português	Supervisionado	PropBank.Br
Fonseca (2013)	Português	Supervisionado	Wikipédia, PLN-BR
Sequeira et al. (2012)	Português	Supervisionado	Wikipédia, BosqueUE, CoNLL-2004
Kozhevnikov and Titov (2013)	Alemão, Chinês, Espanhol, Inglês, Tcheco	Semissupervisionado	Europarl v6, MultiUN

Tabela 16: Resumo das características dos trabalhos relacionados apresentados no Capítulo 3.

verbos auxiliares da anotação automática via AM, a fim de melhorar a qualidade da APS, fortalecendo a relevância da Hipótese 2 e Hipótese 3 desse trabalho de mestrado. Também não foi verificado grande esforço dos trabalhos atuais em investigar a APS em árvores sintáticas não revisadas, ainda assim, o trabalho de [Lluís et al. \(2013\)](#) avaliou seu sistema de APS sobre árvores sintáticas não revisadas de textos em holandês. Pesquisas em APS para o português brasileiro, no entanto, ainda não investigaram árvores sintáticas não revisadas. Por fim, apesar de perceber a tendência dos trabalhos atuais em utilizar diversos gêneros para o treinamento de sistemas de APS, nenhum deles tem como alvo o gênero de opinião sobre produtos na web, foco de trabalhos da Hipótese 5.

O Capítulo 4 apresenta o projeto *Brazilis*, desenvolvido durante esse trabalho de mestrado. Esse trabalho pode contribuir com o estado da arte não apenas da anotação de papéis semânticos para a língua portuguesa mas também provendo resultados que possam servir de motivação para pesquisas em demais línguas ao: utilizar *null elements* para explicitar a ocorrência de sujeitos ocultos, o que propicia a APS sobre uma maior quantidade de sujeitos; testar um gênero distinto dos avaliados pela literatura; averi-

guar o ganho de informação do uso do sentido do verbo como *feature* de AM; separar os verbos auxiliares do conjunto de anotação via AM e; pesquisar a tarefa de APS sobre árvores sintáticas não revisadas de textos em português brasileiro.

Projeto Brazilis

Este trabalho de mestrado, chamado Projeto Brazilis, está inserido em um projeto maior denominado PROS@ - “Processamento Semântico de Textos em Português Brasileiro”¹, do convênio entre o ICMC-USP e a empresa SAMSUNG Eletrônica da Amazônia, cujo objetivo geral é avançar o estado da arte em processamento semântico de textos/documentos escritos em português brasileiro via realização de pesquisa de base e desenvolvimento de ferramentas e produtos de PLN. Mais especificamente, os objetivos da pesquisa de base do projeto maior são a anotação manual de papéis semânticos em textos dos gêneros jornalístico e de opinião de produtos na web. Também é de interesse a criação de ferramentas automáticas para anotação de papéis semânticos (APS) em textos jornalísticos e a exploração da APS em textos de opiniões sobre produtos na web. Com base nesses recursos e ferramentas, o projeto maior tem o objetivo final de construir aplicações de mineração e sumarização de textos, com especial enfoque em opiniões sobre produtos encontrados na web.

Os objetivos deste trabalho de mestrado foram avançar o estado da arte na APS de textos do gênero jornalístico em português brasileiro e avaliar um anotador de papéis semânticos do português brasileiro para o gênero de textos de opinião sobre produtos

¹<http://nilc.icmc.usp.br/semanticnlp/>

na web.

Neste capítulo, são apresentados os métodos desenvolvidos ao longo do projeto de mestrado, suas avaliações e também os recursos utilizados. Alguns recursos como o *cópus* PropBank.Br e as regras para identificação de verbos auxiliares foram desenvolvidos por outros pesquisadores do projeto maior, mas a compilação da parcela do *cópus* PLN-Br, anotada com papéis semânticos e a seleção da recolha do *website* Buscapé, também anotado com papéis semânticos, tiveram participação ativa deste mestrando.

Visando avançar o estado da arte da APS para o português brasileiro, foi desenvolvido um sistema de APS para anotar verbos auxiliares como modificadores do verbo principal. Essa atividade nunca foi abordada na área de APS. Também foi utilizado um *cópus* jornalístico maior para o treinamento do classificador de papéis semânticos do que os usados nos trabalhos de [Alva-Manchego \(2013\)](#) e [Fonseca \(2013\)](#).

Visando anotar papéis semânticos em textos da web contendo opiniões sobre produtos, foi feita uma seleção de sentenças das *reviews* recolhidas do *website* Buscapé. Esse *cópus* foi anotado manualmente com papéis semânticos com a finalidade de verificar a qualidade da anotação automática de um sistema treinado com dados de um *cópus* do gênero jornalístico e testado no gênero de opinião sobre produtos da web. Por fim, foi desenvolvida uma ferramenta de inserção automática de sujeitos não expressos. Sabemos que a omissão de sujeitos é permitida no português, ocorrendo frequentemente em contextos informais. Logo, ao inserir sujeitos não expressos no gênero de opinião, garantimos a anotação do sujeito dos verbos em primeira pessoa. Esses sujeitos são fortes candidatos a receberem um rótulo de papel semântico referente ao agente da ação do verbo. A ferramenta também foi utilizada para o *cópus* jornalístico.

Nas próximas seções são detalhados o procedimento de seleção e compilação de recursos (Seção [4.1](#)) e o desenvolvimento de métodos de uma arquitetura híbrida entre AM e regras para APS (Seção [4.2](#)).

4.1 Recursos

Nesta seção, são detalhados os *cópus* utilizados ao longo do trabalho, apresentando dados quantitativos de cada um. Também é descrita a metodologia de seleção de ins-

tâncias para APS dos corpus PLN-Br e Buscapé, que considerou a duração deste projeto de mestrado. Além disso, são apresentados o procedimento e os resultados da anotação manual de papéis semânticos sobre as seleções realizadas. Por fim, as regras para identificação de verbos auxiliares, empregadas em um módulo do sistema de APS proposto, são apresentadas.

4.1.1 Corpus e Processos de Anotação Manual

Três grandes corpus foram utilizados neste trabalho: o corpus PropBank.Br (versão 1.1), o corpus PLN-Br e o corpus Buscapé.

PropBank.Br (versão 1.1)

O **PropBank.Br (versão 1.1)**, composto por sentenças do gênero jornalístico anotadas com papéis semânticos, difere de sua primeira versão em três pontos: (1) anotação explícita do sentido do verbo conforme os *frames* do repositório Verbo-Brasil, (2) número maior de instâncias (5.942 em sua versão atual e 5.831 em sua versão anterior) e (3) maior número de verbos alvos distintos (1.025 em sua versão atual e 1.023 em sua versão anterior). Um exemplo da anotação do sentido do verbo por ser visto na Figura 6, em que a primeira e a segunda versões do corpus estão alinhadas (esquerda e direita, respectivamente). O item “*wordtag*”, anotado à direita, contém o sentido do verbo alvo (“existir.01”), e um *link* para o *id* do verbo na árvore sintática (“*idref*”). Assim, é possível extrair essa informação para utilizá-la como *feature* no sistema de APS baseado em AM.

Seleção do PLN-Br

O projeto Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil (**PLN-Br**) (Bruckschen et al., 2008), desenvolvido entre 2006 e 2008 por um grupo de universidades brasileiras, teve por objetivo geral a construção de um espaço interinstitucional de interação e intercâmbio de práticas de análise e investigação linguístico-computacional sobre a representação e a recuperação de informação de natureza semântica e pragmático-discursiva veiculada por enunciados

<pre> <frames> <frame name="Argumentos" id="bosC.s2608_f4"> <target> <fenode idref="bosC.s2608_22"/> </target> <fe name="argm-tmp" id="bosC.s2608_f4_e1"> <fenode idref="bosC.s2608_521"/> </fe> <fe name="argm-mnr" id="bosC.s2608_f4_e2"> <fenode idref="bosC.s2608_520"/> </fe> <fe name="arg1" id="bosC.s2608_f4_e3"> <fenode idref="bosC.s2608_518"/> </fe> </frame> </frames> <usp> <uspframes> </uspframes> <uspfes> </uspfes> </usp> <wordtags> </wordtags> </sem> </pre>	<pre> <frame name="Argumentos" id="bosC.s2608_f4"> <target> <fenode idref="bosC.s2608_22"/> </target> <fe name="argm-tmp" id="bosC.s2608_f4_e1"> <fenode idref="bosC.s2608_521"/> </fe> <fe name="argm-mnr" id="bosC.s2608_f4_e2"> <fenode idref="bosC.s2608_520"/> </fe> <fe name="arg1" id="bosC.s2608_f4_e3"> <fenode idref="bosC.s2608_518"/> </fe> </frame> </frames> <usp> <uspframes> </uspframes> <uspfes> </uspfes> </usp> <wordtags> <wordtag idref="bosC.s2608_22" name="sentido">existir.01</wordtag> </wordtags> </sem> </pre>
---	---

Figura 6: Exemplo de diferença entre os corpúscos PropBank.Br versão 1 (à esquerda) e PropBank.Br versão 1.1 (à direita). Circulada em vermelho, a anotação que demarca o sentido do verbo.

produzidos em português brasileiro. O projeto contou com a criação de um corpúscos do gênero informativo (subgênero do jornalístico) (Bruckschen et al., 2008), composto por textos da Folha de São Paulo, entre os anos 1994 e 2005. As estatísticas do corpúscos PLN-Br, que conta com 34.110.689 *tokens* publicados, é apresentada na Tabela 17.

PLN-Br	Número de formas	Números de tipos
Total de palavras	28.972.916	582.669
Palavras em minúscula	25.211.486	217.474
Palavras com inicial maiúscula	2.013.789	66.110
Palavras todas em maiúsculas	797.397	8.167
Números	558.231	23.294
Pontuação	3.710.469	8

Tabela 17: Estatísticas sobre os 34.110.689 *tokens* do corpúscos PLN-Br.

Durante a anotação da primeira versão do PropBank.Br, observou-se que muitos verbos tinham apenas uma instância de anotação. Essa esparsidade de dados, como sabemos, é indesejável para técnicas de aprendizagem de máquina. Segundo Duran et al. (2014), três variáveis devem ser consideradas ao se planejar uma anotação automática de papéis semânticos:

- Quais verbos estão representados no corpúscos;

- Quais sentidos dos verbos estão representados no *cópus*;
- Quais alternâncias sintáticas estão representadas no *cópus*.

Sabemos que um verbo pode admitir vários sentidos e que a distinção entre um sentido e outro depende do nível de detalhamento da análise realizada, ou seja, a distinção de sentidos deve ser feita levando-se em conta a finalidade da tarefa. Para a anotação de papéis semânticos, a mudança de sentido dos verbos está relacionada à mudanças na estrutura sintática das orações, ou seja, existe correlação entre sintaxe e semântica. O detalhamento ideal, nesse caso, é aquele que está associado à marcas na superfície textual, como o uso de diferentes preposições, por exemplo (Duran et al., 2014).

Quanto às alternâncias sintáticas, em geral, a quantidade de sentidos de um verbo está associada à quantidade de alternâncias sintáticas que ele admite, e podem existir verbos que admitem grande alternância para um mesmo sentido. Dessa forma, é necessário levar em conta a quantidade de alternâncias sintáticas de um verbo durante a seleção de sentenças. Assim, o *cópus* deve conter mais **instâncias** de verbos que possuem uma maior diversidade de alternâncias sintáticas. Como o *cópus* PLN-Br contém 4.159.046 ocorrências de verbos, ele pode gerar esse mesmo número de instâncias de anotação. Esse número, contudo, exigiria um altíssimo esforço de anotação para um projeto de dois anos de duração e com um número pequeno de pesquisadores do projeto maior (uma pós-doutoranda, um mestrando e dois alunos de Iniciação Científica, embora o grupo tenha recebido ajuda de outros dois anotadores por um período curto) e, por isso, foi necessário selecionar uma amostra representativa desse universo, que contivesse os mesmos sentidos de verbos contidos no *cópus* original.

Dentro do projeto PROS@, Duran et al. (2014) fizeram uma recolha de sentenças do *cópus* PLN-Br com o objetivo de melhorar o repositório de verbos Verbo-Brasil, criado para apoiar a APS do português do Brasil, por exemplo, a anotação manual de papéis semânticos da ampliação do *cópus* PropBank.Br. Esse novo conjunto de dados serviria de subsídio para um sistema de APS treinado com dados passados pelo *parser* PALAVRAS, sem revisão humana das árvores sintáticas, diferentemente dos sistemas de Alva-Manchego (2013) e Fonseca (2013), que utilizaram o antigo PropBank.Br, cujas árvores sintáticas foram revisadas. A abordagem utilizada limitou a quantidade de exemplos

dos sentidos altamente frequentes e aumentou a amostra de instâncias de anotação para capturar sentidos pouco frequentes e que não são identificáveis sintaticamente.

Para limitar os sentidos altamente frequentes, foram selecionadas quantidade fixa de sentenças de todos os verbos seguidos da conjunção integrante “que”. Esses verbos são altamente frequentes no corpus (composto por textos jornalísticos) pois, em sua maioria, são verbos de elocução (dizer que, falar que, afirmar que, contar que, etc.) (Duran et al., 2014). Também foram selecionadas uma quantidade fixa de sentenças de verbos seguidos de outros verbos na forma infinitiva que, no geral, possuem apenas um sentido. Ainda, foi selecionado um número fixo de instâncias de verbos de cópula (ser, estar, ficar, permanecer, parecer). Os verbos de cópula, embora sejam pouco polissêmicos, admitem uma grande variedade de formas de argumentos (Duran et al., 2014). Por fim, uma quantidade fixa de verbos pronominais foi selecionada. Como o uso dos pronomes reflexivos junto a esses verbos pode assumir diferentes funções, o trabalho de desambiguação constitui grande desafio para a APS (Duran et al., 2014). Assim, como observou-se um alto nível de dificuldade de anotação desse conjunto, nós não o incluímos para APS neste trabalho.

Para o restante do corpus, em que nós não identificamos um padrão sintático para seleção de instâncias, selecionamos todos os verbos com frequência maior do que 1.000. O número de sentenças selecionadas para cada verbo respeitou a Equação 1,

$$seleção (verbo) = \frac{número\ de\ trigramas\ diferentes\ do\ verbo}{frequência\ do\ verbo} \quad (1)$$

considerando trigramas iniciados pelo verbo em questão. Os trigramas foram extraídos de um modelo de língua treinado no corpus PLN-Br. Essa recolha permitiu maior coleta das alternâncias menos frequentes (ou menos previsíveis), que são mais difíceis de serem identificadas via APS automática. Nós não selecionamos instâncias em que um verbo auxiliar foi selecionado como alvo da anotação.

A coleta resultou em 8.345 instâncias, sendo 386 instâncias de verbos seguidos pela preposição “que”; 216 instâncias de verbos seguidos por outro verbo no infinitivo; 71 instâncias de verbos de cópula e 7.672 instâncias de 510 verbos cujo padrão sintático não foi identificado. A distribuição das instâncias selecionadas é apresentada na Tabela 18. Separamos os verbos sem padrão sintático pelas suas letras iniciais para organizar

a distribuição de instâncias de anotação. Esse mestrando colaborou com o processo de seleção de instâncias do corpus PLN-Br.

Categoria	Instâncias	Porcentagem
Letra A	1.193	14,3%
Letra B	121	1,4%
Letra C	1.114	13,3%
Letra D	691	8,3%
Letra E	684	8,2%
Letra F	246	2,9%
Letra G	119	1,4%
Letra H	5	0,1%
Letra I	432	5,2%
Letra J	75	0,9%
Letra L	176	2,1%
Letra M	246	2,9%
Letra N	73	0,9%
Letra O	199	2,4%
Letra P	699	8,4%
Letra Q	50	0,6%
Letra R	535	6,4%
Letra S	375	4,5%
Letra T	387	4,6%
Letra U	65	0,8%
Letra V	187	2,2%
Cópula	71	0,9%
Verbo+QUE	386	4,6%
Verbo+INF	216	2,6%
Total:	8.345	

Tabela 18: Distribuição das instâncias coletadas do PLN-Br.

Anotação Manual de Papéis Semânticos sobre a Seleção do PLN-Br

Após a coleta direcionada à diversidade verbal e sintática de um conjunto de sentenças do corpus PLN-Br, a pós-doutoranda Magali Duran planejou a anotação manual dos conjuntos ArgN e ArgM de papéis semânticos, seguindo a anotação do projeto PropBank (Palmer et al., 2005), porém com manual adaptado para a língua portuguesa. Para tanto, nós geramos as árvores sintáticas de cada instância por meio do *parser* sintático PALAVRAS (Bick, 2000). A saída do *parser* são árvores no formato TigerXML, mesmo

formato do corpus PropBank.Br. Além disso, árvores sintáticas no formato TigerXML podem ser processadas com a ferramenta SALTO (Burchardt et al., 2006).

SALTO² é uma plataforma gráfica para anotação semântica em corpus (um exemplo de sua interface pode ser visto na Figura 7). Além de proporcionar uma anotação fácil e intuitiva, ela possui o recurso de auxílio à adjudicação³ de instâncias de anotação feita por vários anotadores. O papel do adjudicador é resolver as discordâncias entre os anotadores, garantindo um corpus com uniformidade e aderência ao manual de anotação.

Figura 7: Exemplo de interface para anotação na ferramenta SALTO.

Um grupo de 7 anotadores e 1 adjudicador participou do processo de anotação. Os anotadores receberam treinamento do adjudicador e um manual sobre como proceder com a tarefa de anotação de papéis semânticos e também receberam *feedback* inicial sobre sua anotação. O conjunto de papéis semânticos a serem anotados pode ser visto na Tabela 42 do Apêndice B. Após terminado o treinamento, os anotadores consultaram o adjudicador apenas para sanar dúvidas referentes ao manual de anotação. Além disso, o repositório de verbos e seus sentidos (Verbo-Brasil), construído pela pós-doutoranda do projeto PROS@, foi utilizado para consulta, já que primeiramente

²Disponível em <http://www.coli.uni-saarland.de/projects/salsa/page.php?id=software>.

³A **adjudicação** é o processo em que um juiz (geralmente uma pessoa altamente especializada) recolhe a anotação de dois ou mais anotadores, sobre uma mesma instância, e resolve as suas discordâncias.

é necessário identificar o sentido do verbo na sentença para então anotar seus argumentos. Alguns anotadores foram substituídos ao longo da anotação. Porém, o projeto maior sempre manteve 4 anotadores ativos. Esse mestrando participou da equipe de anotação.

O regime de anotação seguiu o padrão **duplo-cego**, em que dois anotadores anotam uma mesma parcela de instâncias e, por fim, o adjudicador resolve as discordâncias. Nós distribuimos a anotação das instâncias por blocos, iniciando por verbos que começam com a letra A e terminando com os que começam com a letra V. Paralelamente, foram distribuídos os conjuntos de verbos de cópula, verbos seguidos pela preposição “que” e verbos seguidos por outro verbo no infinitivo. A distribuição e recuperação das instâncias de anotação foi feita via ferramenta SALTO. Uma instância de anotação nunca foi delegada à mais de uma dupla de anotadores. Como a SALTO não possui um mecanismo rápido para pós-processar as anotações, separando o que teve plena concordância do que apresentou discordância e necessita ir para adjudicação, esse mestrando desenvolveu um *script* para automatizar o *merging* das instâncias.

Finalizada a anotação, calculamos os valores da estatística Kappa ([Carletta, 1996](#)). Como a tarefa de APS tradicionalmente consiste na identificação de argumentos e classificação dos papéis semânticos, calculamos o Kappa individual para cada uma dessas tarefas. Também calculamos o Kappa da identificação de sentido do verbo, uma vez que essa também foi uma tarefa realizada pelos anotadores. Os valores obtidos são apresentados na Tabela 19.

	Identificação	Classificação	Sentido
Verbos de Cópula	0,96	0,90	1,0
Verbo + INF e Verbo + QUE	0,96	0,95	1,0
Verbos sem padrão sintático	0,79	0,75	0,92
Total	0,80	0,76	0,93

Tabela 19: Valores de Kappa para a Identificação de Argumentos, Classificação de Argumentos e escolha de sentido da anotação manual da seleção do PLN-Br.

Percebemos que os conjuntos Verbo + INF (verbos seguidos de outro verbo no infinitivo) e verbo + QUE (verbos seguidos da preposição “que”) obtiveram resultados

<i>Kappa Statistic</i>	<i>Strength of Agreement</i>
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost Perfect

Figura 8: Dados para interpretação da estatística Kappa segundo [Landis and Koch \(1977\)](#).

Fonte: [Landis and Koch \(1977\)](#).

relevantes de Kappa, considerados “quase perfeitos” pela escala Kappa de [Landis and Koch \(1977\)](#) (ver Figura 8). Como a anotação desses conjuntos é muito previsível e possui pouca diversidade sintática, a identificação e a classificação obtiveram valores altíssimos (valor de Kappa igual a 1). A anotação dos verbos de cópula, também obteve resultados “quase perfeitos”. Nesse cenário, porém, existe uma maior diversidade sintática, o que fez com que houvesse uma queda na concordância da classificação de argumentos em relação a anotação dos verbos seguidos de “que” e verbos seguidos de verbo no infinitivo. Isso significa que os anotadores concordaram em identificar que um determinado nó sintático é argumento do verbo, porém discordaram em selecionar o papel semântico a ser atribuído. Por fim, o conjunto dos verbos cujo padrão sintático não foi identificado obteve um Kappa “substancial”. Para esse conjunto, verificamos que houve discordância na escolha do sentido dos verbos, o que desencadeou discordâncias sobre a detecção dos argumentos e a seleção dos papéis semânticos. Além disso, esse é o conjunto dos verbos cuja anotação não é previsível, ou seja, que possui uma alta variabilidade de uso na língua. Logo, compreende-se a alta discordância na anotação.

Ainda analisando a anotação, separamos a parcela do corpus em que não houve discordância entre os anotadores, ou seja, que não havia necessidade de revisão. A parcela restante foi enviada ao adjudicador, que tratou as discordâncias e gerou uma versão final da anotação. A Tabela 30a apresenta a proporção de instâncias de anotação em que houve plena concordância entre as duplas de anotadores e a porção que teve que ser revisada. Os valores de concordância e discordância dos conjuntos Verbo + QUE e Verbo + INF não constam na tabela pois foram enviados diretamente para adjudicação,

como forma de exercício da tarefa. Percebemos que, apesar do uso do manual de anotação, houve discordância em mais de 50% das instâncias. Isso significa que, apesar dos altos valores de Kappa, a maioria das instâncias teve que ser revisada pelo adjudicador. Por exemplo, uma instância é enviada para adjudicação mesmo que, de 10 argumentos, os anotadores tenham concordado em 9 e discordado em apenas 1.

É importante ser ressaltado que muitas árvores sintáticas contêm erros gerados pelo *parser* automático. Apesar da disponibilidade do manual de anotação, a livre interpretação sobre como anotar essas árvores inconsistentes pode ter contribuído para o número de discordâncias obtido.

Seleção do Buscapé

O *website* Buscapé⁴ é especializado em sugerir mercadorias aos usuários, utilizando métricas como preço em diversas lojas *online* e a opinião dos usuários sobre os produtos. Hartmann et al. (2014) fizeram uma coleta de opiniões do *website* via *crawling*, que resultou em um total de 85.910 *reviews* contendo 17.000.000 *tokens*. A Figura 9 mostra três textos do *corpus* para ilustrar alguns problemas para um *parser* dedicado a analisar textos bem escritos. Os textos estão anotados com os problemas que eles apresentam e uma possível correção entre colchetes. Como o *parser* é desenvolvido para lidar com textos bem escritos, é necessário corrigir os erros mais frequentes nesse novo gênero a fim de aproximá-lo da norma culta de escrita da língua e, assim, possibilitar uma melhor performance do *parser* sobre esses dados.

Nós geramos estatísticas para o total de palavras contidas, palavras com inicial em maiúsculo, palavras em caixa baixa, palavras em caixa alta, números e também pontuação. As considerações acerca da definição dessas categorias é apresentada a seguir:

- **Palavra:** Consideramos que uma palavra é um *token* composto ao menos por uma letra ou um número. Consideramos diversos tipos de acentuação (mesmo incorretos) para tentar capturar o maior conjunto de palavras possíveis (lembrando que estamos trabalhando com conteúdo gerado por usuários na web que traz os problemas apontados na Figura 9).

⁴Disponível em <http://buscape.com.br>.

	Concordância Plena	Alguma Discordância	Total
Letra A	501	692	1.193
Letra B	62	59	121
Letra C	477	637	1.114
Letra D	272	419	691
Letra E	279	405	684
Letra F	100	146	246
Letra G	38	81	119
Letra H	0	5	5
Letra I	177	255	432
Letra J	25	50	75
Letra L	74	102	176
Letra M	97	149	246
Letra N	27	46	73
Letra O	74	125	199
Letra P	268	431	699
Letra Q	18	32	50
Letra R	1.200	335	535
Letra S	136	239	375
Letra T	186	201	387
Letra U	33	32	65
Letra V	86	101	187
Cópula	46	25	71
Verbo + QUE	—	—	386
Verbo + INF	—	—	216
Total	3.176	4.567	8.345

Tabela 20: Proporção da anotação do PLN-Br em que houve plena concordância entre os anotadores.

- **Número:** Foram considerados como sendo números aquelas cadeias unicamente numéricas, ou com ocorrências dos símbolos: ‘,’, ‘.’, ‘%’ e ‘\$’, ou seja, números decimais, valores monetários e porcentagens.
- **Pontuação:** As pontuações aceitas foram os símbolos: ‘,’, ‘.’, ‘;’, ‘!’, ‘?’ e ‘-’.

Os resultados obtidos em termos de formas e tipos são apresentados na Tabela 21. Os dados do Buscapé foram comparados com os de um corpus jornalístico, o Corpus NILC (ou Corpo São Carlos)⁵. Os valores da coluna “Percentual” do Total de palavras

⁵Informações sobre o Corpo São Carlos estão disponíveis em <http://www.linguateca.pt/>

#2, internetês (it), nome próprio (ner), falta de acento (ac), erro de ortografia (co), siglas (si), palavra em inglês (in)
ela <u>e</u> [ac: é] muito escura quando <u>vc</u> [it: você] esta deitado se <u>vc</u> [it: você] <u>tiver</u> [it: estiver] sentado ela <u>e</u> [ac: é] boa mas deitada <u>nao</u> [ac: não] <u>e</u> [ac: é] muito <u>nao</u> [ac: não]. A <u>Samsung</u> [ner] inova o mercado de <u>tv's</u> [si: televisões] com uma grande obra de arte que se adequa à qualquer ambiente. Esta <u>tv</u> [si: televisão] possui excelente imagem quando ligada a uma fonte de <u>dvd</u> [si: Digital Versatile Disc] com <u>hdmi</u> [si: High-Definition Multimedia Interface] e na <u>tv</u> [si: televisão] a cabo (Digital). O som é perfeito quando é personalizado pelo usuário. Ou seja, MENU, Soud [co: SOUND], EQUALIZAR E ENTER [in].
#5, internetês (it), nome próprio (ner), erro de ortografia (co)
Gostei <u>dimais</u> [co: demais] dessa câmera, comprei outra! Além de uma excelente e reconhecida marca, essa câmera tem um <u>design</u> [in] super inovador e <u>mtu</u> [it: muito] atraente... uma resolução <u>mtu</u> [it: muito] boa e <u>td</u> [it: tudo] o que uma boa câmera SONY tem que ter!! Até <u>hj</u> [it: hoje] nunca me deixou na mão... recomendo!!!
#8, erro de ortografia (co), internetês (it), falta de acento (ac)
<u>muto</u> [co: muito] bom para <u>manuziar</u> [co: manusear] quando <u>vc</u> [it: você] <u>ta</u> [it: está] <u>trabalhano</u> [co: trabalhando] com este produto , <u>nao</u> [ac: não] tenho que <u>recramar</u> [co: reclamar] gostei mesmo parabéns. RECOMENDO O PRODUTO, FÁCIL DE USAR, ADOREI !

Figura 9: Exemplos de textos do corpus Buscapé com seus problemas característicos para um *parser* dedicado à analisar textos bem escritos.

Corpo São Carlos			
	Número de formas	Números de tipos	Percentual
Total de palavras	32.462.201	399.808	1,23%
Palavras em minúscula	22.979.950	157.039	73,92%
Palavras com inicial maiúscula	4.794.306	129.635	15,42%
Palavras todas em maiúsculas	307.083	10.460	0,98%
Números	547.592	3.384	1,76%
Pontuação	2.456.825	20.564	7,90%
Buscapé			
	Número de formas	Números de tipos	Percentual
Total de palavras	14.556.810	125.772	0,86%
Palavras em minúscula	12.426.706	66.318	74,93%
Palavras com inicial maiúscula	1.062.395	10.986	6,40%
Palavras todas em maiúsculas	771.902	12.771	4,65%
Números	132.428	1.976	0,79%
Pontuação	2.190.640	7	13,20%

Tabela 21: Estatísticas do corpus Buscapé, em comparação com o Corpo São Carlos.

representam o *type/token ratio*, que avalia a riqueza lexical (variedade do léxico em-

acesso/corpus.php?corpus=SAOCARLOS. Resultados desse corpus estão disponíveis em http://www.linguateca.pt/acesso/desc_corpus.php?corpus=SAOCARLOS#quant.

pregado), e é de 1,23% no texto jornalístico e de 0,86% no gênero de opiniões sobre produtos. Isso significa que, nos textos de opiniões, a diversidade lexical é 30% menor do que no gênero jornalístico, ou seja, mais pobre. Essa medida pode indicar que o gênero é próximo da fala, ou que os revisores (aqueles que escreveram sobre o produto na internet) não são muito letrados. As demais porcentagens apresentam a proporção das categorias em relação ao total de elementos.

No gênero de opiniões, o uso de palavras em minúsculas é de 74,93%, enquanto que no gênero jornalístico é de 73,92%. A diferença de uso de inicial maiúscula é mais expressiva: no corpus jornalístico é de 15,42%, enquanto no gênero de opinião é de apenas 6,40%. Análises preliminares mostram que isso não significa ausência de nomes próprios no gênero de opiniões, mas sim a sua grafia totalmente em minúscula. O corpus jornalístico estudado contém cerca de 70% mais números que o corpus de opinião e uma das prováveis causas disso é o fato de, no corpus de opinião, os números serem combinados aos nomes de produtos, sendo caracterizados como palavras e não como números (e.g. iPhone6). O corpus Buscapé usa proporcionalmente o dobro da pontuação usada no corpus jornalístico, porém, devido a diferenças na definição⁶ de pontuação utilizada nos corpus, não se pode tirar conclusões acerca do uso da pontuação.

Conhecendo as características dos dados obtidos, foi feita uma seleção de instâncias do corpus Buscapé para a tarefa de APS. Além de sermos pioneiros na exploração de APS para esse gênero no PB, um dos objetivos desta pesquisa de mestrado foi investigar o desempenho de um sistema treinado no gênero jornalístico para textos do gênero de opinião de produtos da web. Para tanto, primeiramente nós realizamos um pré-processamento das sentenças do corpus Buscapé a fim de normalizá-las. O *pipeline* de execução, composto pelo sistema UGCNormal (Duran et al., 2015) e por um *speller* contextual (ainda não divulgado), pode ser visto na Figura 10.

O *pipeline* do UGCNormal é composto por um sistema sentenciador, baseado em AM, que identifica a falta de pontuação do usuário e segmenta um texto (López and Pardo,

⁶Nota do site da Linguateca: Para o Corpo São Carlos, a categoria “pontuação” inclui todos as unidades que contenham sinais de pontuação e que não tenham sido considerados como abreviaturas (incluindo portanto números do tipo “3.876.210”, abreviaturas com mais de um ponto como “S.O.S.” ou “E.U.A”, assim como erros de codificação ainda não depurados, como “primeiro-ministro.”).

Figura 10: Pipeline de pré-processamento do corpus Buscapé.

2015) – a falta de pontuação é uma característica do gênero dos textos em questão. Após pontuar e sentenciar o texto, ele é tokenizado. Em sequência, é aplicado o *speller* de Avanço et al. (2014), desenvolvido para lidar com os desafios de textos web no PB, contando com um algoritmo fonético desenvolvido especialmente para lidar com esses dados. Em termos de desempenho, esse *speller* corrige 16% mais palavras do corpus Buscapé do que o *speller* Aspell⁷ (Atkinson, 2006) (resultado obtido em um experimento que restringiu o conjunto de teste para aquelas sentenças com erros de *spelling* que independem de contexto). Por fim, as siglas são capitalizadas, os termos da internet (internetês) são mapeados para seus correspondentes na língua padrão e os nomes próprios têm suas primeiras letras capitalizadas.

Após a aplicação do UGCNormal, foi utilizado um segundo *speller* (ainda não publicado) que utiliza uma abordagem de aprendizagem de máquina, capturando erros por meio do contexto em que ocorrem. Em termos de desempenho, esse *speller* corrige 33% mais palavras do corpus Buscapé do que o *speller* Aspell (resultado obtido em um experimento do autor que restringiu o conjunto de teste para aquelas sentenças que possuem erros conhecidos de *spelling*). O uso de contexto possibilita diferenciar, por exemplo, uma letra “e” de um “é”, não acentuado, e também “esta” de “está”, não acentuado. O uso dos dois sistemas possibilitou, por meio da correção de diversos erros de escrita, a obtenção de um corpus mais próximo do gênero jornalístico.

Uma vez processado o corpus Buscapé, nós contabilizamos os verbos mais frequentes do corpus e selecionamos os com frequência acima de 1.000, que representam 78% dos verbos do corpus (a lista de verbos e a quantidade de instâncias selecionadas para cada um podem ser vistas na Tabela 41 do Apêndice A). Então, foram selecionadas 11 sentenças para cada um dos 80 verbos com frequência acima de 1.000. Das 880

⁷Disponível em <http://aspell.net>.

instâncias resultantes, foram descartadas aquelas com árvore sintática desconexa e árvore sintática sem um nó VP como pai do verbo alvo de anotação. Isso resultou em 873 instâncias de anotação. Para todos os verbos, foram selecionadas 11 instâncias de anotação, exceto: “caber” com 10, “conseguir” com 10 e “dever” com 6. Nós fizemos um experimento aumentando a seleção de instâncias do verbo “dever”, porém, mesmo selecionando 100 instâncias deste verbo, a recolha ainda era insuficiente para suprir as 11 instâncias desejadas (erros na árvore sintática). Logo, manteve-se o critério original. Finalmente, foram inseridos 703 sujeitos ocultos (Eu e Nós) e também anotados papéis semânticos em verbos auxiliares: 49 ArgM-Asp, 28 ArgM-Mod, 9 ArgM-Pas e 18 ArgM-Tml. O significado de cada um desses papéis pode ser visto na Tabela 42 do Apêndice B.

O corpus gerado foi anotado em duplas por três anotadores, seguindo os moldes da anotação da seleção do PLN-Br.

Anotação Manual de Papéis Semânticos sobre a Seleção do Buscapé

A anotação das instâncias selecionadas do corpus Buscapé seguiu os mesmos moldes da anotação da seleção do PLN-Br. Dessa vez, no entanto, nós selecionamos apenas 3 anotadores para a tarefa, sendo que esse mestrando também participou do processo de anotação.

Após ser realizada a anotação, primeiramente avaliamos a concordância da anotação via estatística Kappa. Os valores apresentados na Tabela 22 são “quase perfeitos” para a identificação de argumentos e escolha de sentidos e “substancial” para a classificação de argumentos (ver Figura 8). Os valores de Kappa mostram um resultado pouco superior ao obtido para o conjunto de verbos com frequência acima de 1.000 do corpus PLN-Br. Isso mostra que a anotação do Buscapé foi mais simples do que a do PLN-Br, porém os anotadores continuaram discordando. Acreditamos que o gênero impactou fortemente a tarefa de anotação. Por mais que tenha sido realizada uma etapa de pré-processamento do corpus, muitos erros se mantiveram. Alguns deles, como a falta de pontuação ou de coesão, impactam no resultado do *parser* e, portanto, no resultado da anotação. Vale lembrar que a árvore sintática não é perfeita nas sentenças desse corpus e isso abre margem para a interpretação do anotador, o que pode gerar discordância.

	Identificação	Classificação	Sentido
Total	0,82	0,79	0,90

Tabela 22: Valores de Kappa para a Identificação de Argumentos, Classificação de Argumentos e Escolha de Sentido da anotação manual da seleção do Buscapé.

Também calculamos a taxa de instâncias em que houve plena concordância entre os anotadores. Pelos valores apresentados na Tabela 23, verificamos que 45,47% das instâncias anotadas obtiveram 100% de concordância entre os anotadores, enquanto aproximadamente 54,53% foram enviadas para adjudicação. Assim como na anotação do PLN-Br, verificamos pelo Kappa que a anotação do Buscapé foi muito uniforme, mas, ainda assim, mais da metade das instâncias tiveram de ser adjudicadas, gerando retrabalho.

	Concordância Plena	Alguma discordância	Total
Total	397	454	873

Tabela 23: Proporção da anotação do Buscapé em que houve plena concordância entre os anotadores.

4.1.2 Regras para Identificação de Verbos Auxiliares

Verbos auxiliares são aqueles verbos que adicionam informação gramatical à outro verbo, chamado de verbo principal (ou pleno), de quatro maneiras (Duran and Aluísio, 2010):

- **Aspecto** – Indica se a ação do verbo principal está começando (inceptiva), terminando (terminativa) ou se ainda não se realizou (imperfectiva); se tem uma duração (durativa) ou se repete ao longo do tempo (frequentativa). Exemplo: A oração “Eu acabei de terminar a lição de casa.” é um exemplo de ação terminativa.
- **Modo** – Indica se a ação do verbo principal pertence ao universo da possibilidade, obrigatoriedade ou proibição. Seguem três exemplos ilustrando possibili-

dade, obrigação e proibição, respectivamente: “Eu posso realizar a tarefa”, “Eu devo realizar a tarefa” e “Eu não devo realizar a tarefa”.

- **Tempo** – Em português, alguns tempos se expressam por meio de flexão verbal. Exemplo: “fará” (futuro do verbo “fazer”). Outros tempos verbais, chamados de tempos compostos, são construídos com a ajuda de um verbo auxiliar, que assume a flexão de tempo e evita que o verbo principal tenha que ser flexionado, por exemplo, “teria feito” (pretérito mais-que-perfeito composto do subjuntivo). Além dos verbos “ter” e “haver” seguidos do particípio passado, tradicionalmente conhecidos como auxiliares de tempo, o verbo “ir”, seguido de infinitivo, também forma tempos compostos, por exemplo, “vai fazer” (futuro composto do indicativo). As formas compostas são muito utilizadas pelo fato de representarem uma economia na língua, já que o falante só precisa saber flexionar os auxiliares e não os verbos principais.
- **Voz passiva** – A mudança de voz do verbo em português, em sua forma desenvolvida (analítica), é feita com o auxílio do verbo “ser” seguido do verbo principal no particípio passado. A voz passiva construída pelo verbo “ser” coloca o paciente na posição de sujeito e desloca o agente, que é sujeito na voz ativa, para uma posição de menor destaque, introduzido pela preposição “por”. Na voz passiva, o agente pode, inclusive, não ser expresso, por exemplo, “O acordo foi feito com sucesso [pelos dois países].” A voz passiva construída com o verbo “estar” não é reconhecida por todos os gramáticos, pois o particípio pode ser interpretado como um adjetivo, por exemplo, “A cidade está cercada por favelas.”. Para a finalidade deste trabalho, contudo, ambos os verbos, “ser” e “estar” são considerados auxiliares de voz passiva.

O PropBank inglês não distingue a forma de anotação dos verbos auxiliares daquela utilizada para os verbos plenos, ou seja, um verbo auxiliar possui argumentos semânticos. Como os verbos auxiliares possuem argumentos, na maioria das vezes previsíveis (salvo exceções em que a estrutura gramatical da sentença é complexa), a anotação de papéis semânticos para esses verbos é trivial, isso é, o verbo auxiliado sempre estará diretamente à direita do verbo auxiliar. Essa decisão do projeto PropBank faz com que

a performance de sistemas de APS baseados em aprendizagem de máquina seja mais elevada do que a dos sistemas que desconsideram a anotação dos auxiliares.

Entendemos que o uso de métodos de aprendizagem de máquina é adequado para casos em que há algum tipo de diversidade nos dados que não pode ser facilmente capturado por regras. Portanto, ao desconsiderar a anotação de verbos auxiliares, temos a performance do classificador mais condizente com as dificuldades reais atribuídas a ele.

[Duran and Aluísio \(2010\)](#) desenvolveram uma lista de padrões para identificar verbos auxiliares em uma sentença e a maneira na qual atuam sobre o verbo principal (ver Tabela 24). Os padrões consistem em um verbo auxiliar, seguido ou não por uma preposição, terminando com o verbo auxiliado. Dependendo da regra, o verbo auxiliado pode estar no gerúndio (v-ger), no infinitivo (v-inf) ou no particípio (v-pcp). A última coluna da Tabela 24 lista a etiqueta referente ao papel semântico relacionado à forma como o verbo auxiliar está modificando o verbo principal. As etiquetas são:

- **ArgM-Asp** – Etiqueta para verbos auxiliares de aspecto.
- **ArgM-Mod** – Etiqueta para verbos auxiliares de modo.
- **ArgM-Pas** – Etiqueta para verbos auxiliares de voz passiva.
- **ArgM-Tml** – Etiqueta para verbos auxiliares de tempo.

Ao utilizar as regras apresentadas para identificação de verbos auxiliares, desconsideramos a proposta utilizada no PropBank e anotamos esses verbos, via regras, como modificadores do verbo auxiliado (verbo alvo de uma instância de anotação). Um exemplo de como a anotação ficaria se utilizássemos a proposta do PropBank pode ser visto na Figura 11. Nesse exemplo, o verbo auxiliar “ser” é alvo de anotação e possui como argumento o verbo “abandonar”. Um exemplo de como a anotação é realizada utilizando a proposta de [Duran and Aluísio \(2010\)](#) pode ser visto na Figura 12. Nesse cenário, o verbo auxiliar “ser” não é alvo de anotação e sim argumento do verbo “abandonar”. A diferença entre as duas abordagens é que no PropBank.Br perde-se uma instância de anotação (a do verbo auxiliar) e ganha-se um novo papel semântico, que marca a forma

Auxiliar	Preposição	Auxiliado	Etiqueta
acabar	-	v-ger	argm-asp
acabar	de	v-inf	argm-asp
acabar	por	v-inf	argm-asp
andar	-	v-ger	argm-asp
cessar	de	v-inf	argm-asp
chegar	a	v-inf	argm-asp
começar	a	v-inf	argm-asp
começar	por	v-inf	argm-asp
continuar	-	v-ger	argm-asp
continuar	a	v-inf	argm-asp
correr	a	v-inf	argm-asp
costumar	-	v-inf	argm-asp
dar	de	v-inf	argm-asp
deixar	de	v-inf	argm-asp
desatar	a	v-inf	argm-asp
dever	-	v-inf	argm-mod
estar	-	v-ger	argm-asp
estar	para	v-inf	argm-asp
estar	por	v-inf	argm-asp
ficar	-	v-ger	argm-asp
ficar	de	v-inf	argm-mod
haver	de	v-inf	argm-mod
haver	que	v-inf	argm-mod

Auxiliar	Preposição	Auxiliado	Etiqueta
haver	-	v-pcp	argm-tml
hesitar	-	v-inf	argm-asp
hesitar	em	v-inf	argm-asp
ir	-	v-ger	argm-asp
ir	-	v-inf	argm-tml
ousar	-	v-inf	argm-asp
parar	de	v-inf	argm-asp
passar	a	v-inf	argm-asp
permanecer	-	v-ger	argm-asp
poder	-	v-inf	argm-mod
recomeçar	a	v-inf	argm-asp
sair	-	v-ger	argm-asp
seguir	-	v-ger	argm-asp
ser	-	v-pcp	argm-pas
ter	de	v-inf	argm-mod
ter	que	v-inf	argm-mod
ter	-	v-pcp	argm-tml
terminar	-	v-ger	argm-asp
tornar	a	v-inf	argm-asp
vir	-	v-ger	argm-asp
vir	a	v-inf	argm-asp
viver	-	v-ger	argm-asp
voltar	a	v-inf	argm-asp

Tabela 24: Regras para identificação de verbos auxiliares e etiqueta atribuída à função exercida pelo verbo auxiliar.

como o verbo auxiliar está auxiliando. O método de anotação de verbos auxiliares com papéis semânticos modificadores é apresentado na Seção 4.2.2.

Figura 11: Exemplo de anotação de verbos auxiliares, conforme o PropBank inglês.

Figura 12: Exemplo de anotação de verbos auxiliares, conforme proposto por [Duran and Aluísio \(2010\)](#).

4.2 Uma Arquitetura Híbrida para APS Automática

Como visto no Capítulo 3, a maioria dos trabalhos atuais emprega técnicas de aprendizagem de máquina para a anotação de papéis semânticos. Porém, essa nem sempre é a melhor solução. Em cenários em que verificamos a existência de um padrão muito recorrente, pode ser mais vantajoso elaborar regras para detectar esses padrões, uma vez que, dado o caráter estatístico da aprendizagem de máquina, essa abordagem é sujeita a erros. Também já foi discutido na Seção 4.1 que a diferença na forma de anotação dos verbos auxiliares para os verbos plenos é importante para não inflar a performance do sistema de APS com um conjunto de dados previsível de ser anotado.

Além disso, é relevante ressaltar outros detalhes do projeto PropBank. A seção financeira do córpus Penn Treebank ([Marcus et al., 1993](#)) foi anotada com papéis semânticos para ser usada como córpus de treinamento, sendo o uso de *null elements*, um constituinte artificial inserido para representar a omissão de um elemento, um dos pontos fortes da sua anotação sintática. O córpus PropBank.Br, em contraste com o córpus PropBank, não possui a explicitação de sujeitos não expressos. Essa é uma das razões da diferença de performance dos classificadores de papéis semânticos treinados no PropBank.Br, em relação aos treinados no PropBank. Dada a diferença de características entre o gênero jornalístico e de opinião de produtos, a principal contribuição da explicitação dos sujeitos ocultos é para a APS do córpus Buscapé (gênero de opinião), pois nesse gênero se relatam opiniões pessoais, fazendo largo uso de pronomes na primeira pessoa do singular. Como apresentado na Seção 4.1, nós inserimos 703 sujeitos

ocultos nas 873 instâncias selecionadas (0,80 inserções por instância), que é uma quantidade significativa de sujeitos – que podem vir a ser anotados com papéis semânticos que demarcam o sujeito de um verbo.

Dada a premissa de que o uso de regras é adequado para cenários em que um padrão bem definido deve ser detectado, este trabalho de mestrado propõe um sistema híbrido de APS, em que parte atua com o uso de regras e parte com aprendizagem de máquina. O sistema proposto possui um módulo de pré-processamento que insere sujeitos ocultos (Eu e Nós), a fim de aumentar o conjunto de dados de anotação. Para o módulo de AM, utilizamos o sistema anotador de papéis semânticos de [Alva-Manchego \(2013\)](#) mas, diferente desse trabalho, o foco aqui é na anotação sobre árvores sintáticas sem revisão.

A Figura 13 ilustra a arquitetura em mais alto nível do sistema proposto neste mestrado. As etapas de execução são descritas abaixo e os três módulos principais são detalhados nas próximas seções:

Figura 13: Diagrama do sistema híbrido proposto para anotação de papéis semânticos.

1. A entrada do usuário é um texto com várias sentenças.
2. Os textos do usuário são processados pelo *parser* PALAVRAS e as árvores sintáticas

são salvas em um arquivo no formato TigerXML.

3. O módulo de inserção de sujeitos ocultos é executado sobre as árvores sintáticas das sentenças do usuário, inserindo nós sintáticos NP, contendo o respectivo sujeito Eu/Nós, para os casos de sujeitos não expressos.
4. (a) O módulo baseado em regras anota os verbos auxiliares como modificadores do verbo principal, com base nos padrões identificados por [Duran and Aluísio \(2010\)](#).
(b) O módulo baseado em aprendizagem de máquina anota os conjuntos ArgN e ArgM clássicos de papéis semânticos, descritos por [Palmer et al. \(2010\)](#) e utilizados por [Alva-Manchego \(2013\)](#), [Fonseca \(2013\)](#) e pelo projeto de mestrado atual (veja o conjunto de papéis semânticos utilizado nesse trabalho de mestrado na Tabela 42 do Apêndice B).
5. A saída do sistema é um arquivo no formato CoNLL anotado com papéis semânticos.

4.2.1 Módulo de Inserção de Sujeitos Não Expressos

Diferente da língua inglesa, a língua portuguesa possibilita inferir a pessoa gramatical a partir da flexão do verbo. Assim, ao omitir um sujeito nós não perdemos a flexão do verbo. Apesar dessa omissão do sujeito dessa omissão não impactar muito a compreensão humana, ela representa uma dificuldade extra para sistemas de PLN, como os sistemas anotadores de papéis semânticos, que perdem dados de treinamento e/ou teste. Três fenômenos de ocorrência de sujeitos não expressos no texto são permitidos:

- **Sujeitos elípticos de orações subordinadas e coordenadas.** Por exemplo, “Nós queremos (Nós) dar uma festa.” e “Eles lutaram e (Eles) venceram a batalha.”.
- **Sujeitos ocultos**, que podem ser inferidos pela flexão do verbo. Por exemplo, “(Eu) Acho que isso é possível.”.
- **Sujeitos indeterminados.** Por exemplo, “(Eles/Elas) Dizem que isso não é possível.”.

Segundo o guia de anotação do PropBank (Palmer et al., 2005), existe uma ordem de importância nos papéis semânticos e os mais importantes devem ser anotados primeiro. Uma metodologia que utiliza isso é a *re-ranking*, que foi largamente difundida após o trabalho de Toutanova et al. (2008). A abordagem de *re-ranking* advoga que: a anotação do Arg1 depende da tentativa de anotação do Arg0, a anotação do Arg2 depende da tentativa de anotação do Arg1 e assim por diante. Desse modo, se o sujeito não está expresso no texto, a falta de sua anotação compromete a sequência de anotação dos demais papéis.

A seção financeira do corpus Penn Treebank (Marcus et al., 1993) foi anotada com papéis semânticos pelo projeto PropBank. Marcus et al. (1993) afirmam que “o mecanismo mais simples para incluir informação sobre a estrutura predicado-argumento, mesmo que de forma indireta, é fazendo com que a árvore sintática contenha, explicitamente, null elements.”. A etiqueta *PRO* do Penn Treebank, utilizada quando o sujeito de um verbo não está especificado ou realizado, também pode ser utilizada para abordar sujeitos ocultos no português brasileiro. Por exemplo, na sentença:

“Nós esperamos vencer o jogo.”

deve ser inserido o sujeito do verbo “vencer”, que não está realizado na sentença. Nesse caso é inserida a etiqueta *PRO* como apresentado a seguir:

“Nós-1 esperamos *PRO*-1 vencer o jogo.”

Claro que adaptações para inserir a pessoa verbal (Eu/Nós/Ele/Eles/Ela/Elas) podem ser feitas.

Neste trabalho de mestrado foi realizado um estudo exploratório para o levantamento de regras para identificação e inserção automática de sujeitos ocultos (Eu e Nós) e de indeterminados (elemento artificial “Suj” demarcando um sujeito indeterminado). Como o nosso desejo é atuar sobre dados reais, rodamos o *parser* PALAVRAS sobre as sentenças do corpus PropBank.Br, gerando uma versão não *gold-standard* do corpus. Iniciamos com uma regra simples: “Inserir um nó NP, contendo um sujeito, toda vez que não houver um nó NP à esquerda do verbo.”. A análise da regra nos mostrou a necessidade de inserir novas restrições no sistema. O procedimento de identificação e implementação de restrições foi realizado até ser sanada a maioria dos casos.

Determinamos que o sistema não deve inserir sujeitos faltantes quando:

- Um VP contém apenas um verbo e ele está no infinitivo, gerúndio ou no particípio.
- Um PP contém apenas um verbo e ele é um dos seguintes verbos impessoais: chover, existir, haver, nevar ou ventar.

Para os casos canônicos, em que o sujeito deveria estar à esquerda do verbo, o sistema insere sujeito quando:

1. Um VP contém uma ocorrência do verbo “ser” e esse verbo ou está iniciando uma sentença ou assume o papel de verbo auxiliar.
2. Um VP inicia uma sentença.
3. Não há um NP imediatamente antes de um PP entre vírgulas. O PP deve preceder uma vírgula que precede imediatamente um VP.
4. Um VP sucede imediatamente uma pontuação, essa pontuação não pode casar com a regra 3 e não há NP imediatamente antes da pontuação.
5. Não há um NP ou ADVP precedendo imediatamente um VP.
6. Há um ADVP precedendo diretamente um VP e não há NP logo antes do ADVP.
7. Há um pronome oblíquo diretamente antes de um VP e não há NP logo antes do pronome.
8. Um elemento com o rótulo “ACC” (rótulo para objeto direto dado pelo *parser* PALAVRA) imediatamente precede um VP e não há NP imediatamente antes do ACC. Uma variação válida é quando um ADVP precede diretamente um ACC mas não há NP logo antes do ADVP.

No Português é possível inverter a ordem canônica da relação sujeito-verbo, posicionando o sujeito à direita do verbo. Em casos em que não há NP à esquerda do verbo, pode ser que um NP à direita seja o seu sujeito ao invés de um complemento direto. Também há casos ambíguos em que a detecção do sujeito se torna ainda mais complexa, como:

“Antes do nascer do sol, acordou o trabalhador.”.

A possibilidade de inversão na posição sujeito-verbo representa uma dificuldade extra para a identificação de sujeitos não expressos. Verbos de elocução⁸, que são encontrados nos discursos diretos e indiretos, são um caso comum de ocorrência desse tipo de inversão. Para sabermos se estamos lidando com um cenário de verbo de elocução, ou seja, se à esquerda é esperada a fala e, à direita, quem é o falante, definimos:

1. Verbos de elocução são os verbos a seguir: acentuar, acrescentar, afirmar, alegar, argumentar, berrar, bradar, clamar, contar, dizer, exclamar, falar, gritar, indagar, informar, perguntar, responder e sustentar.
2. Se um VP é diretamente precedido por uma pontuação⁹ e seguido por um NP ou pela cadeia ADVP e NP, o verbo alvo é um verbo de elocução.

Por fim, elaboramos regras para o posicionamento do sujeito inserido. O padrão é que o sujeito seja posicionado exatamente antes do verbo, mas, dependendo da estrutura sintática da sentença, o sujeito pode ser deslocado algumas posições à esquerda ou à direita do verbo:

- O rótulo *ks* (*subordinating conjunction*), anotado pelo *parser* PALAVRAS, é utilizado para identificar quais advérbios devem ser mantidos entre o sujeito e o verbo e quais devem ser posicionados antes do sujeito. Porém, como o *parser* comete muitos erros ao categorizar os advérbios, também fixamos os advérbios que mais frequentemente são posicionados entre o sujeito e o verbo: não, nunca, sempre, quase, só, já e também. Sempre que for necessário inserir um sujeito e houver um dos advérbios listados ou um advérbio com o rótulo “ks”, o sujeito é inserido antes do advérbio. Do contrário, o sujeito é mantido junto ao verbo.
- Quando houver um pronome reflexivo diretamente antes do verbo, o sujeito deve ser inserido antes do pronome.
- Quando houver um pronome reflexivo diretamente antes do verbo e houver um dos pronomes identificados na regra 1 de posicionamento do sujeito, o sujeito

⁸Verbos de elocução são utilizados para introduzir ou anunciar uma fala.

⁹Essa pontuação é a marca que delimita o discurso indireto.

deve ser deslocado duas posições à esquerda, antes do pronome reflexivo e do advérbio.

- Na ocorrência de um verbo de elocução, o sujeito deve ser posicionado imediatamente após o verbo.

Uma análise das regras obtidas (Hartmann et al., 2014) em um conjunto de 200 sentenças do PropBank.Br apresentou a acurácia de 88,8% para a inserção de sujeito para as primeiras pessoas do singular e plural (Eu e Nós) e de apenas 55,5% para a inserção de sujeito para as terceiras pessoas do singular e plural (ver Tabela 25). A Tabela 26 ainda apresenta a fonte dos erros do sistema, apontando muitos erros do *parser* nas inserções de sujeito para as terceiras pessoas, além de falta de regras. Pelo resultado positivo na inserção de sujeitos não expressos para as primeiras pessoas e pelos erros nessa inserção não serem dominados por problemas do *parser*, optou-se neste trabalho de mestrado por inserir automaticamente apenas os sujeitos não expressos das primeiras pessoas nas seleções do PLN-Br e do Buscapé.

	Esperados	Inserções Automáticas	Inserções Corretas	Inserções Incorretas	Inserções Faltantes	Acurácia
Suj	106	120	67	43	10	55,8%
Eu/Nós	26	27	24	2	1	88,8%
Total	132	147	91	45	11	61,9%

Tabela 25: Resultados de uma análise preliminar sobre a inserção de sujeitos não expressos em uma seleção de 200 sentenças do corpus PropBank.Br.

	Total de Erros	Erros do <i>parser</i>	Erros por falta de regras
Suj	53	29	24
Eu/Nós	3	1	2
Total	56	30	26

Tabela 26: Fonte de erros da análise preliminar sobre a inserção de sujeitos não expressos em uma seleção de 200 sentenças do corpus PropBank.Br.

4.2.2 Anotação de Verbos Auxiliares com Papéis Semânticos Modificadores

Existe mais de uma visão acerca dos papéis semânticos para verbos auxiliares. O PropBank inglês trata essa categoria de verbos como verbos plenos, e não como verbos que de fato carregam informação semântica à outro verbo. Assim, o PropBank trata verbos auxiliares como verbos comuns, anotando papéis semânticos tradicionais via métodos de aprendizagem de máquina. Contextualizando para o cenário do português brasileiro, a anotação do verbo “começar” na sentença “Eu comecei a gostar do seriado.”, por exemplo, atribuiria o papel Arg1 à “a gostar”. [Duran and Aluísio \(2010\)](#), no entanto, vêem essa categoria de verbos como modificadores de verbos plenos, ou seja, esses verbos não esperam argumentos semânticos, mas suprem o verbo a ser modificado com um papel semântico modificador auxiliar de: aspecto (ArgM-Asp), modo (ArgM-Mod), tempo (ArgM-Tml) ou voz passiva (ArgM-Pas). Para o exemplo dado, não seriam atribuídos argumentos ao verbo “começar”, mas ele próprio seria rotulado como “ArgM-Asp” do verbo “gostar”.

A abordagem de [Duran and Aluísio \(2010\)](#) nunca foi implementada em APS. Logo, a literatura não apresenta um conjunto de *features* para classificar, via AM, os papéis semânticos propostos. Além disso, a teoria de papéis semânticos de [Palmer et al. \(2005\)](#) declara que os papéis semânticos estão localizados fora do sintagma verbal do verbo alvo de anotação, o que implica na necessidade de uma reformulação na maneira como os classificadores de papéis semânticos são implementados para identificar essa nova classe de papéis semânticos.

Como existe um conjunto de regras bem definido para identificação de verbos auxiliares, nós desenvolvemos neste trabalho de mestrado um anotador de papéis semânticos para verbos auxiliares utilizando uma abordagem baseada em regras. Apesar de saber que não existe um consenso sobre o conjunto que contemple os verbos auxiliares para o Português ([Baptista et al., 2010](#); [Bick, 2000](#); [Duran and Aluísio, 2010](#)), optou-se neste trabalho de mestrado por adotar o conjunto sugerido por [Duran and Aluísio \(2010\)](#). O método desenvolvido implementa as regras propostas sobre o esquema genérico da Equação 2, em que os elementos entre colchetes são opcionais, ou seja, podem não ser realizados em uma sentença.

verbo auxiliar [*intervenientes*] [**preposição**] [*intervenientes*] **verbo auxiliado**. (2)

Os elementos *intervenientes* apresentados na Equação 2, apesar de não estarem descritos nas regras da Tabela 24, podem ser encontrados dentro das cadeias. Esses elementos são: advérbios, preposições e pronomes reflexivos. Por exemplo, “Eu acabei de me ferir.”, o pronome reflexivo “me” aparece como objeto interveniente na cadeia do verbo auxiliar “acabar”. Também é possível detectar cadeias com mais de um verbo auxiliar modificando um mesmo verbo alvo, como na sentença “Eu finalmente passei a deixar de dever dinheiro.”, em que ambos os verbos “passar” e “deixar” são auxiliares de “dever”.

A seguir, é apresentado o módulo de anotação de papéis semânticos, baseado em AM, que selecionamos para lidar com os papéis semânticos clássicos de [Palmer et al. \(2005\)](#), os ArgNs e os ArgMs.

4.2.3 Módulo de APS baseado em AM

O módulo anotador de papéis semânticos via AM, utilizado neste trabalho, é responsável pela anotação dos papéis semânticos clássicos (ArgN e ArgM) propostos pelo projeto PropBank ([Palmer et al., 2005](#)). Como não é o objetivo deste trabalho desenvolver um novo sistema de APS que utilize AM, optamos por fazer uso do sistema de [Alva-Manchego \(2013\)](#), que possui os melhores resultados para APS no PB (79,6 de F1 em *10-fold cross-validation* sobre o corpus PropBank.Br versão 1.0).

Sabe-se que muitos verbos podem ser empregados em mais de um contexto. Por exemplo, o verbo “jogar” pode assumir o papel de “arremessar” ou de “praticar” algo. O que define a sua função é a estrutura sintática em que o verbo está inserido e um pouco de conhecimento de mundo para distinguir qual a função mais adequada. [Toutanova et al. \(2008\)](#) comentam, citando o trabalho de [Roland \(2002\)](#), que se o sentido do verbo em uma sentença for desambiguizado, a anotação dos papéis semânticos torna-se mais fácil, melhorando a qualidade da anotação.

O PropBank.Br versão 1.1 possui cerca de 1,36 sentidos por verbo e 80% dos verbos presentes no corpus apresentam apenas um sentido; 12,8% apresentam dois sentidos;

3,4% apresentam três sentidos e 3,7% apresentam quatro ou mais sentidos. Como a taxa de sentidos é maior que a taxa de verbos, temos um indicativo de que há necessidade de conhecer o sentido para melhor anotar o verbo. Porém, como a variedade de sentidos é maior que a variedade de verbos (1.396 e 1.025, respectivamente) e o *cópus* conta com apenas 5.942 instâncias, acreditamos que o sentido é muito esparsos para ser utilizado como *feature* de AM. Assim, há a necessidade de avaliar essa questão e ter um indicativo da eficácia do uso de uma *feature* que explicita o sentido do verbo alvo em um sistema de APS baseado em AM. A avaliação sobre o ganho de informação do uso do sentido do verbo alvo, como *feature* de AM para APS, será apresentada na Seção 5.1.

4.3 Considerações Finais

Neste capítulo foram apresentados os recursos utilizados e desenvolvidos ao longo do trabalho de mestrado e detalhados os procedimentos de seleção de *cópus* e anotação. Também foram apresentados os métodos desenvolvidos neste trabalho de mestrado. A avaliação dos métodos desenvolvidos é apresentada no Capítulo 5.

A recolha do *cópus* PLN-Br possibilita a este trabalho de mestrado o estudo da APS para árvores sintáticas não revisadas em português brasileiro. Tal estudo, que implica na análise da APS em um cenário real de aplicação, nunca foi realizado para a língua portuguesa.

A recolha do *cópus* Buscapé possibilita a este trabalho de mestrado o estudo da APS no gênero de textos de produtos recolhidos da web. Percebemos na revisão apresentada na Seção 3 que os trabalhos atuais vêm analisando a tarefa de APS em diferentes gêneros. No entanto, nenhum deles apresentou resultados sobre o gênero de revisões de produtos. Para realizar tal estudo, fizemos uso de um *pipeline* de pré-processamento do gênero, a fim de aproximar seus textos da normal culta da língua. Também desenvolvemos um sistema de inserção de sujeitos omissos pois, como o gênero web se aproxima da fala, há maior incidência de omissão de sujeitos, potenciais candidatos a receberem rótulos de papéis semânticos.

Por fim, discordando da forma que Palmer et al. (2005) anota papéis semânticos em verbos auxiliares, desenvolvemos um sistema baseado em regras para lidar separada-

mente com esses verbos. Ao contrário do PropBank, que trata verbos auxiliares como verbos plenos (alvos de anotação), tratamos esses verbos como modificadores dos verbos auxiliares, ou seja, como argumentos. Assim, o sistema de APS baseado em AM não deve mais lidar com a anotação previsível dessa classe de verbos, que acarretaria em uma inflação no desempenho do sistema.

Avaliações: Resultados e Discussão

Neste capítulo serão apresentados os experimentos realizados ao longo do projeto de mestrado e os resultados obtidos. Discussões, causas e justificativas para os resultados também são levantadas. Cada uma das cinco seções deste capítulo trata da avaliação de cada uma das hipóteses dessa pesquisa, relacionadas à: avaliação do módulo de inserção de sujeitos ocultos na Seção 5.1, avaliação do módulo de anotação de verbos auxiliares na Seção 5.2, avaliação do uso do sentido do verbo alvo da APS como *feature* de AM na Seção 5.3, análise dos desafios da APS sobre árvores sintáticas não revisadas por humanos na Seção 5.4 e investigação da performance de um sistema de APS treinado no gênero jornalístico e testado no gênero de opiniões web é apresentada na Seção 5.5.

Alva-Manchego (2013), em seu sistema de APS, adotou o modelo de avaliações conjuntas da CoNLL (*Conference on Computational Natural Language Learning*) como estratégia de avaliação e, dessa forma, a entrada e a saída do sistema são mapeadas em um formato plano de colunas. Um exemplo desse formato pode ser visto na Figura 14.

A CoNLL é uma conferência com o objetivo de, dado um tema, comparar os diversos trabalhos submetidos, tentando avançar o estado da arte no tema proposto. Para tanto, *scripts* avaliadores são disponibilizados, retornando a nota da saída do sistema com as

1	Agora	agora	ADV	-	(S*	(FCL*	(FCL(ADV*P)	-	(AM-TMP*)	*
2	,	,	PU	-	*	*	*	-	*	*
3	os	o	ART	MIP	*	*	(CU(NP*	-	(A0*	*
4	soldados	soldado	N	MIP	*	*	*)	-	*	*
5	e	e	CONJ-C	-	*	*	*	-	*	*
6	a	o	ART	FIS	*	*	(NP*	-	*	*
7	polícia	polícia	N	FIS	*	*	*))	-	*)	*
8	estão	estar	V-FIN	PR13PIIND	*	*	(VP*	-	*	*
9	trabalhando	trabalhar	V-GER	-	*	*	*)	trabalhar	(V*	*
10	juntos	junto	ADV	MIP	*	*	(ADV*P)	-	(AM-MNR*)	*
11	para	para	PRP	-	*	*	(PP*	-	(AM-PNC*	*
12	prender	prender	V-INF	-	(S*	(ICL*	(ICL(VP*)	prender	*	(V*)
13	os	o	ART	MIP	*	*	(NP*	-	*	(A1*
14	traficantes	traficante	N	MIP	*)	*)	*))	-	*)	*)
15	.	.	PU	-	*)	*)	*)	-	*	*

Figura 14: Sentença “Agora, os soldados e a polícia estão trabalhando juntos para prender os traficantes.” no formato plano de colunas da CoNLL.

Fonte: Alva-Manchego (2013).

métricas utilizadas na área. No caso de APS supervisionada, as medidas são Precisão, Acurácia, e F1. Alva-Manchego utilizou o padrão proposto pela CoNLL a fim de utilizar o *script* avaliador disponível e para ter um ambiente de igualdade de comparação com os demais trabalhos da área. Assim, os resultados relacionados ao método de AM desse trabalho de mestrado foram obtidos pelo mesmo *script* avaliador.

É importante esclarecer que o conjunto de papéis semânticos anotados no PLN-Br e no Buscapé é maior do que o anotado no PropBank.Br. Para uniformizar os testes, os papéis semânticos foram restritos aos anotados no PropBank.Br. Para tanto, foram removidas: 31 anotações de ArgM-Com, 38 anotações de ArgM-Gol, 39 anotações de ArgM-Nse e 20 anotações de ArgM-Pin.

Outro ponto interessante a ser esclarecido é que as recolha dos córpis PLN-Br e Buscapé foram feitas por verbos, não por sentenças. Apesar da proporção média de 1,06 instâncias por sentença, obtivemos casos em que a mesma sentença foi selecionada para diferentes verbos. Esse é um contraste em relação ao córpis PropBank.Br, que teve todos os verbos plenos das sentenças anotados, resultando em 1,75 verbos anotados por sentença. Essa questão não afeta os métodos de AM, cuja unidade de anotação é o verbo e seus argumentos e não a sentença.

Nós utilizamos o Teste de Wilcoxon para os testes de significância estatística de sistemas desenvolvidos. Para termos 95% de confiabilidade nos resultados (valor comumente utilizado), trabalhamos com o valor de corte de 0,05. A escolha do teste de

Wilcoxon se deve ao fato de que os valores das métricas para se comparar sistemas de PLN (que usam dados como textos, sentenças, cláusulas, entre outros) não seguem uma distribuição normal (Søgaard et al., 2014).

Por fim, algumas instâncias foram descartadas durante o mapeamento dos *corpus* do formato Tiger-XML para o formato CoNLL por conterem erros sintáticos impossíveis de serem mapeados, por exemplo, um nó sintático com dois pais. Foram perdidas 268 instâncias da seleção do PLN-Br (aproximadamente 8% dos dados) durante esse processo de conversão. Esse mapeamento resultou em 3.908 instâncias de anotação, distribuídas em 2.830 sentenças. Também foram perdidas 119 instâncias de anotação do Buscapé (aproximadamente 13,6% dos dados), resultando em 754 instâncias de anotação distribuídas em 712 sentenças. Essa quantidade significativa de dados perdidos se deve a problemas de desempenho do *parser* PALAVRAS.

5.1 Avaliação do Módulo de Inserção de Sujeitos Não Expressos

Nós avaliamos manualmente a inserção de sujeitos ocultos (“Eu” e “Nós”) nas 8.345 instâncias de anotação de papéis semânticos selecionadas do *corpus* PLN-Br. A **precisão obtida foi de 87,8%** na inserção de sujeitos no *corpus* PLN-Br, que corresponde à 506 das 576 inserções realizadas. Esse resultado, similar ao apresentado em Hartmann et al. (2014), mostra que o sistema desenvolvido é capaz de inserir sujeitos não expressos das primeiras pessoas do singular e plural com uma boa precisão em textos jornalísticos. Ainda é possível verificar que o valor de 576 inserções corresponde à 33% das 1.730 ocorrências de “Eu” e “Nós” no *corpus*. Esse valor nos mostra a porção representativa da lacuna preenchida para APS.

Devido ao fato de termos selecionados um número fixo ocorrências de verbos para compor a nossa seleção do PLN-Br, parcela dos verbos presentes nas sentenças selecionadas ficaram sem anotação. Isso significa que o ganho real da inserção de sujeitos ocultos para APS não é 33%. Das 576 inserções, apenas 142 (correspondendo a 24% das inserções e 8,2% das ocorrências de “Eu” e “Nós” no *corpus*) foram realizadas sobre verbos alvo de anotação, resultando em 131 anotações de Arg0 e 11 anotações de Arg1. As anotações de Arg0 sobre sujeitos artificialmente inseridos representam 4% das

anotações de Arg0 do corpus e, as de Arg1, apenas 0,17%.

Para verificar se há significância estatística entre as diferenças de performance do sistema treinado sobre o PLN-Br, utilizando as anotações sobre os sujeitos artificialmente inseridos, e a performance do sistema treinado sobre o PLN-Br, sem utilizar as anotações sobre os sujeitos artificialmente inseridos, realizamos o **Teste de Wilcoxon** sobre os resultados de F1 do *10-fold cross-validation* dos sistemas em questão. As hipóteses de pesquisa são:

H0: A hipótese nula é de que não há diferença significativa entre a performance do sistema treinado sobre o PLN-Br, utilizando as anotações sobre os sujeitos artificialmente inseridos, e a performance do sistema treinado sobre o PLN-Br, sem utilizar as anotações sobre os sujeitos artificialmente inseridos.

H1: A hipótese alternativa é de que há diferença significativa entre as performances, sendo que o que usa as anotações sobre os sujeitos artificialmente inseridos tem maior medida F1.

Podemos refutar H0 se o *p-value* do teste for menor que 0,05. Isso nos dá 95% de significância estatística. O teste retornou um *p-value* de 0,6842. Logo, **não há evidências para rejeitar a hipótese nula no teste sobre o PLN-Br**. Uma avaliação mais ampla precisa ser realizada.

Também foi verificada a qualidade da inserção de sujeitos não expressos na seleção do corpus Buscapé. Das 664 inserções realizadas, apenas 36 estavam erradas, o que resulta em **94,5% de precisão na inserção de “Eu” e “Nós” na seleção do corpus Buscapé**. As inserções correspondem à 33% do total de 1.623 ocorrências de sujeitos nas primeiras pessoas do singular e plural. Interessante observar que a ocorrência de sujeitos nas primeiras pessoas é 10 vezes mais frequente nesse gênero do que no jornalístico¹. Quando consideramos apenas as instâncias de anotação, obtemos o valor de 151 inserções (corresponde à 22,7% das inserções e 9,3% do total de “Eu” e “Nós” do corpus) de sujeitos não expressos, sendo 145 de Arg0 e 6 de Arg1. As anotações de

¹A seleção do corpus PLN-Br é 10 vezes maior que a do Buscapé. No entanto, há 1.623 sujeitos nas primeiras pessoas do singular e plural no corpus de opinião e apenas 1.730 no corpus jornalístico.

Arg0 sobre os sujeitos artificialmente inseridos correspondem à 48,8% dos argumentos Arg0, enquanto as inserções de Arg1 correspondem apenas a 1,1%.

Realizamos o **Teste de Wilcoxon** para averiguar se há ganho significativo na inserção de sujeitos não expressos para a APS sobre o corpus Buscapé. Análogo ao experimento com o PLN-Br, comparamos o resultado de *10-fold cross-validation* de uma versão contendo os sujeitos ocultos e outra sem contê-los. Com um *p-value* de 0,01953, rejeitamos H_0 e **temos evidências de que há diferença estatisticamente significativa entre a inserção de sujeitos não expressos e a não inserção para a APS sobre o Buscapé**. Podemos verificar na Tabela 27 que houve uma queda de 22,43 pontos de F1 na anotação do Arg0 (sujeito prototípico) e uma queda de 4,95 pontos de F1 na nota geral do sistema ao desconsiderar os sujeitos não expressos inseridos. A não utilização desses sujeitos fez com que sobrassem apenas 152 anotações de Arg0 que, em contraste às 500 ocorrências de Arg1, é um valor restrito.

	corr.	excess	missed	prec.	rec.	F1
Overall	850	476	448	64,17	65,54	64,79
A0	232	65	35	77,67	86,3	81,72
A1	326	163	138	66,88	70,48	68,52
A2	83	55	57	61,50	61,31	59,73
A3	1	3	5	3,33	10,00	5,00
A4	2	4	8	15,00	13,33	14,00
AM-ADV	17	23	22	45,83	47,17	41,95
AM-CAU	7	15	24	39,83	25,00	24,49
AM-DIS	11	14	28	28,00	23,43	24,16
AM-EXT	9	5	10	56,67	54,17	50,33
AM-LOC	25	22	16	57,92	62,14	56,85
AM-MNR	34	43	31	46,77	53,45	46,76
AM-NEG	66	7	16	90,74	79,56	84,38
AM-PRD	0	5	9	0,00	0,00	0,00
AM-PRP	11	15	11	36,5	43,67	38,85
AM-TMP	26	37	38	43,55	41,81	40,48

	corr.	excess	missed	prec.	rec.	F1
Overall	695	470	467	59,79	59,91	59,84
A0	82	57	54	60,57	60,12	59,29
A1	319	147	140	68,52	69,58	68,95
A2	86	50	54	62,43	60,82	61,20
A3	1	3	5	5,00	10,00	6,67
A4	3	4	7	17,50	23,33	19,52
AM-ADV	19	22	20	53,21	52,50	45,37
AM-CAU	8	14	23	35,83	36,25	35,01
AM-DIS	8	19	31	27,00	24,83	20,69
AM-EXT	9	4	9	40,00	40,00	37,38
AM-LOC	24	21	17	56,00	58,45	55,32
AM-MNR	37	50	29	43,22	60,96	48,99
AM-NEG	63	10	19	85,51	77,40	80,30
AM-PRD	0	5	9	0,00	0,00	0,00
AM-PRP	11	16	11	33,33	41,67	36,28
AM-TMP	25	48	39	34,07	35,14	32,80

(a) Resultados considerando os sujeitos não expressos inseridos e suas anotações.

(b) Resultados desconsiderando os sujeitos não expressos inseridos e suas anotações.

Tabela 27: Resultados de um experimento *10-fold cross-validation* com a seleção do corpus Buscapé, na análise do uso de sujeitos ocultos.

Assim, percebemos que a inserção de sujeitos ocultos cobriu uma lacuna significativa na APS sobre textos opinativos do corpus Buscapé. Entendemos ainda que, além dos casos de omissão de sujeitos de primeira pessoa, há também casos de omissão dos sujei-

tos das terceiras pessoas do singular e plural. Logo, se o módulo de inserção de sujeitos cobrir as terceiras pessoas do singular e plural, acreditamos que melhores resultados na APS em textos de opinião sobre produtos coletados da web podem ser obtidos.

No cópulus PLN-Br, devido à frequência de sujeitos das primeiras pessoas não ser significativa, não há evidências de que o módulo de inserção de sujeitos ocultos melhorou a performance de APS. Já no cópulus Buscapé, a restrição de ocorrência de casos de Arg0 fez com que o sistema não fosse capaz de generalizar a sua anotação tão bem e, portanto, o módulo de inserção de sujeitos fez diferença na performance de APS. Assim, entendemos que os sujeitos inseridos contribuíram para a tarefa de APS no cenário em que a sua frequência de inserção foi significativa (cópulus Buscapé). Portanto, **validamos a Hipótese 1 desse trabalho.**

5.2 Avaliação do Sistema de APS em Verbos Auxiliares

Para a análise da robustez do sistema que implementou as regras de identificação de verbos auxiliares, analisamos manualmente a inserção automática de papéis semânticos para verbos auxiliares nas 8.345 instâncias selecionadas do cópulus PLN-Br. Foram contabilizadas todas as inserções automaticamente realizadas, as inserções excedentes (indevidas) e também as faltantes. As métricas utilizadas foram a **Suporte** (Equação 3) e a **Confiança** (Equação 4). Suporte indica a proporção de inserções de uma regra em relação às inserções de todas as regras. Confiança indica a taxa de acerto das inserções realizadas para uma determinada regra, considerando as inserções corretas, indevidas e faltantes.

$$\text{suporte}(x_i) = \frac{\text{inserido}(x_i)}{\sum_{k=1}^n \text{inserido}(x_k)}, \quad 1 \leq i \leq n \quad (3)$$

$$\text{confiança}(x_i) = \frac{\text{inserido}(x_i)}{\sum_{k=1}^n \text{inserido}(x_k) + \text{excedente}(x_k) + \text{faltante}(x_k)}, \quad 1 \leq i \leq n \quad (4)$$

Nº	Regra	Inseriu	Excedente	Faltante	Suporte	Confiança
1	acabar + v-ger	8	-	-	0,47%	100,00%
2	acabar + de + v-inf	4	-	-	0,24%	100,00%
3	acabar + por + v-inf	6	-	-	0,35%	100,00%

4	andar + v-ger	1	-	-	0,06%	100,00%
5	cessar + de + v-inf	-	-	-	0,00%	0,00%
6	chegar + a + v-inf	22	-	-	1,30%	100,00%
7	começar + a + v-inf	39	-	5	2,30%	88,64%
8	começar + por + v-inf	-	-	-	0,00%	0,00%
9	continuar + v-ger	7	-	-	0,41%	100,00%
10	continuar + a + v-inf	5	-	-	0,30%	100,00%
11	correr + a + v-inf	-	-	-	0,00%	0,00%
12	costumar + v-inf	7	-	-	0,41%	100,00%
13	dar + de + v-inf	-	-	-	0,00%	0,00%
14	deixar + de + v-inf	13	-	1	0,77%	92,86%
15	desatar + a + v-inf	-	-	-	0,00%	0,00%
16	dever + v-inf	149	2	1	8,80%	98,03%
17	estar + para + v-inf	-	-	-	0,00%	0,00%
18	estar + v-ger	99	4	-	5,84%	96,12%
19	estar + por + v-inf	-	-	-	0,00%	0,00%
20	ficar + v-ger	2	-	-	0,12%	100,00%
21	ficar + sem + v-inf	1	-	-	0,06%	100,00%
22	ficar + de + v-inf	-	-	-	0,00%	0,00%
23	haver + de + v-inf	-	-	-	0,00%	0,00%
24	haver + que + v-inf	-	-	-	0,00%	0,00%
25	haver + v-pcp	18	-	-	1,06%	100,00%
26	hesitar + em	-	-	-	0,00%	0,00%
27	ir + v-ger	5	-	-	0,30%	100,00%
28	ir + v-inf	236	2	6	13,93%	96,72%
29	ousar + v-inf	4	-	-	0,24%	100,00%
30	parar + de + v-inf	1	-	-	0,06%	100,00%
31	passar + a + v-inf	21	-	-	1,24%	100,00%
32	permanecer + v-ger	-	-	-	0,00%	0,00%
33	poder + v-inf	263	9	8	15,53%	93,93%
34	recomeçar + a + v-inf	-	-	-	0,00%	0,00%
35	sair + v-ger	-	-	-	0,00%	0,00%
36	seguir + v-ger	-	-	-	0,00%	0,00%
37	ser + v-pcp	587	-	3	34,65%	99,49%
38	ter + de + v-inf	19	-	-	1,12%	100,00%
39	ter + que + v-inf	40	1	1	2,36%	95,24%
40	ter + v-pcp	96	1	2	5,67%	96,97%
41	terminar + v-ger	-	-	-	0,00%	0,00%
42	tornar + a + v-inf	1	-	-	0,06%	100,00%
43	vir + v-ger	13	1	-	0,77%	92,86%
44	vir + a + v-inf	4	-	2	0,24%	66,67%
45	viver + v-ger	1	-	-	0,06%	100,00%
46	voltar + 'a' + 'v-inf'	13	-	-	0,77%	100,00%

Tabela 28: Resultados da inserção automática de papéis semânticos para verbos auxiliares.

Ao analisar os dados da Tabela 28, percebe-se uma disparidade no número de inserções de cada regra. A Figura 15 ilustra a distribuição da métrica Suporte das regras examinadas nas 8.345 sentenças selecionadas do corpus PLN-Br. Percebemos a predominância do verbo “ser” (regra 37) nas ocorrências de auxiliares, totalizando 34,65% das inserções. Outros verbos vêm em sequência, sendo eles: “poder”, “ir” (seguido de verbo no infinitivo), “dever”, “estar” (seguido de verbo no gerúndio) e “ter” (seguido de verbo no particípio). Todos os outros casos possuem menos de 2,5% de ocorrências cada e, representam, juntos, apenas 15% das inserções.

Figura 15: Valores para a métrica Suporte das regras utilizadas na identificação de verbos auxiliares no corpus PLN-Br.

A esparsidade do uso de regras também nos indica que devemos ser cautelosos ao analisar a Confiança das regras empregadas. Uma vez que uma regra tenha sido pouco utilizada no corpus, não podemos considerar seus resultados como sendo representativos (apesar do grande conjunto de dados analisado). Por exemplo, apesar da regra “andar seguido de verbo no gerúndio” possuir 100% de Confiança, ela possui apenas uma ocorrência no corpus (0,06% de Suporte), o que não é significativo. Portanto,

ao invés de tirar a média da Confiança das regras, fizemos uma soma ponderada da Confiança pelo Suporte (ver Equação 5), o que dá um maior peso para as regras mais utilizadas (mais frequentes no corpus). Com isso, temos uma **Confiança Ponderada de 96,76%** no sistema que implementou as regras desenvolvidas para identificação de verbos auxiliares e rotulação de seus papéis semânticos.

$$\text{confiança ponderada}(x) = \sum_{k=1}^n \text{confiança}(x_k) * \text{suporte}(x_k) \quad (5)$$

Por fim, levantamos as causas dos erros de inserção de papéis semânticos para verbos auxiliares. Alguns erros são oriundos de falhas no sistema e outros do *parser* utilizado. A seguir serão listados os erros com alguns exemplos:

- **Confusão entre os verbos “ser” e “ir”** – Clássico problema de etiquetação morfossintática rotulada pelo *parser*, por exemplo, em “...a questão foi parar na delegacia.”, a palavra “foi” foi rotulada como verbo “ser” ao invés de “ir”.
- **Caractere “aspas” não sendo considerado como objeto interveniente na cadeia de auxiliares** – Esse caractere não foi considerado como um elemento válido na cadeia entre o verbo auxiliar e o verbo alvo, o que acarreta a não aceitação da cadeia e, logo, a não identificação do verbo auxiliar, por exemplo, “...ir ‘lutar’...” e “pode ‘até nomear...’”.
- **Problema na etiqueta morfossintática do verbo** – Verbo auxiliar ou verbo alvo não foi identificado como verbo pelo *parser*, por exemplo, em “...passe a jogar...”, a palavra “passe” (verbo passar) foi identificado como substantivo e em “...vão quebra tudo.”, devido à um erro de grafia, o *parser* rotulou a palavra “quebra” (verbo quebrar) como substantivo.
- **Problemas algorítmicos** – O verbo é auxiliar do verbo alvo mas também de um outro verbo dentro da cadeia que não é auxiliar do verbo alvo, por exemplo, “...um acordo deve ser anunciado para levantar o veto...” e “...pode ser sequestrada para pagar...”.
- **Problemas com a partícula “se” não sendo considerada em algumas cadeias** – A partícula “se” é uma grande fonte de erros na anotação do *parser* PALAVRAS.

Logo, ao tentar tratar todas as possíveis exceções de anotação, erros podem ser inseridos no sistema. Também é necessário melhorar a cadeia de elementos intervenientes aceitos em algumas regras, por exemplo, “...começou a se mexer...” e “...começará a se refletir...”.

Para finalizar a análise, alteramos o algoritmo de *prunning* de árvores sintáticas, utilizado por [Alva-Manchego \(2013\)](#), para considerar os demais elementos dentro do VP do verbo principal como candidatos à anotação. Assim, podemos anotar os verbos auxiliares via método de AM e comparar o resultado com o do sistema baseado em regras. A Tabela 29 mostra o quanto é previsível anotar os verbos auxiliares no corpus PLN-Br, cujas performances estão acima da obtida na anotação do papel semântico de negação (AM-NEG), conhecido por obter alto desempenho de anotação, devido a sua baixa variedade sintática e lexical. O desempenho na anotação dos verbos auxiliares foi de 94,82 de F1 para ArgM-Asp, 94,24 de F1 para ArgM-Mod, 97,45 de F1 para ArgM-Pas e 98,08 de F1 para ArgM-Tml. Isso resulta em uma média de 96,14 de F1 para a anotação dos verbos auxiliares. Essa alta taxa de performance infla a nota geral do sistema, como observado na Seção 4.1.2. Esse resultado nos mostra que, além da tarefa poder ser realizada via regras, com um bom desempenho, excluí-la do método de AM faz com que não infleamos a performance do sistema anotando um conjunto tão trivial como o apresentado.

Acreditamos que, por mais que o sistema baseado em regras necessite de aperfeiçoamento, sua Confiança Ponderada de 96,76% indica que ele funciona muito bem para textos do gênero jornalístico. Isso nos mostra que a anotação de papéis semânticos para verbos auxiliares pode ser feita via utilização de regras, **validando a Hipótese 2 desse trabalho.**

	corr.	excess	missed	prec.	rec.	F1
Overall	12.532	5.416	4.671	69,81	72,83	71,29
A0	2.708	678	494	79,99	84,54	82,19
A1	4.606	1.720	1.647	72,81	73,65	73,22
A2	869	961	678	47,52	56,31	51,47
A3	22	68	127	26,99	15,92	18,62
A4	18	8	30	62,50	37,87	44,88
AM-ADV	164	148	134	52,73	55,49	53,74
AM-ASP	247	4	23	98,50	91,45	94,82
AM-CAU	50	72	89	41,31	38,08	38,69
AM-DIR	0	4	13	0,00	0,00	0,00
AM-DIS	183	126	160	59,22	53,03	55,86
AM-EXT	74	40	51	65,59	59,20	61,42
AM-LOC	526	465	202	53,21	72,15	61,15
AM-MNR	302	344	253	46,84	54,38	50,19
AM-MOD	409	6	44	98,51	90,43	94,24
AM-NEG	361	42	54	89,40	87,22	88,19
AM-PAS	545	7	22	98,74	96,22	97,45
AM-PRD	32	126	137	19,42	17,88	17,95
AM-PRP	145	111	76	56,60	66,38	60,17
AM-REC	0	0	1	0,00	0,00	0,00
AM-TML	314	6	7	98,27	97,91	98,08
AM-TMP	929	444	340	67,73	73,04	70,19

Tabela 29: Resultado da anotação de verbos auxiliares no corpus PLN-Br, via AM.

5.3 Experimentos com o Uso do Sentido do Verbo como *Feature* de AM

Para verificar se o sentido do verbo aumenta a performance na classificação de papéis semânticos via métodos de AM, 3 *features* utilizadas por [Alva-Manchego \(2013\)](#) foram alteradas para incorporarem o sentido do verbo:

- **Lema do Predicado + Caminho** → Sentido do Predicado + Caminho.
- **Lema do Predicado + Núcleo** → Sentido do Predicado + Núcleo.
- **Lema do Predicado + Tipo de Sintagma** → Sentido do Predicado + Tipo de Sintagma.

A primeira *feature* é composta pelo sentido do verbo e o caminho até o constituinte em análise, a segunda *feature* é composta pelo sentido do verbo e o núcleo do constituinte em análise e a terceira *feature* é composta pelo sentido do verbo e o tipo de sintagma do constituinte em análise. Realizamos experimentos com a F1 do *10-fold cross-validation* de diferentes *corpus* para comparar os desempenhos obtidos na APS ao utilizar, ou não, as *features* de sentido do verbo. As seguintes hipóteses são verificadas:

H0: A hipótese nula é de que não há diferença entre a performance do sistema treinado utilizando as *features* de sentido do verbo e do sistema que não utiliza essas *features*.

H1: A hipótese alternativa é de que há diferença na performance, sendo que o que usa as *features* de sentido do verbo tem maior medida F1.

Aplicamos o **Teste de Wilcoxon** para comparar os desempenhos obtidos ao anotar papéis semânticos no *corpus* PropBank.Br, na seleção realizada sobre o *corpus* PLN-Br e também na junção dos dois *corpus*. Em suma, há divergência entre alguns sentidos utilizados nos *corpus*. Como a anotação humana se baseia nos *framefiles* do repositório Verbo-Brasil e esse repositório foi sendo aprimorado à medida que os *corpus* foram sendo anotados, alguns sentidos foram divididos ou aglomerados. Ainda assim, optamos por realizar um experimento com ambos os *corpus* para aproveitar a maioria dos sentidos que se manteve inalterada.

Para o PLN-Br, obtivemos 69,12 de F1 com as *features* clássicas e 69,3 de F1 com as *features* que incorporam o sentido do verbo alvo. Com um *p-value* de 0,6953, **não há evidências para rejeitar a hipótese nula para o *corpus* PLN-Br**. Uma avaliação mais ampla precisa ser realizada.

Para o PropBank.Br, obtivemos 79,6 de F1 com as *features* clássicas e 79,75 de F1 com as *features* que incorporam o sentido do verbo alvo. Com um *p-value* de 0,9219, **não há evidências para rejeitar a hipótese nula para o *corpus* PropBank.Br**. Uma avaliação mais ampla precisa ser realizada.

Para a mescla dos *corpus* PropBank.Br e seleção do PLN-Br, obtivemos 73,63 de F1 com as *features* clássicas e 73,71 de F1 com as *features* que incorporam o sentido do verbo alvo. Com um *p-value* de 0,9218, **não há evidências para rejeitar a hipótese**

nula para a mescla dos **córpus PropBank.Br e seleção do PLN-Br**. Uma avaliação mais ampla precisa ser realizada.

Entendemos que a alta granularidade de verbos em ambos os **córpus PLN-Br e PropBank.Br** foi determinante para os resultados obtidos. Apesar do sentido do verbo determinar o conjunto de papéis semânticos válidos para anotação, a baixa frequência de verbos acarretou uma esparsidade ainda maior de sentidos e, como sabemos, a frequência dos dados é determinante para o bom desempenho de métodos de AM. Assim, **não podemos validar a Hipótese 3, nem rejeitá-la**. Mais estudos são necessários.

5.4 Experimentos de APS no Gênero Jornalístico: PropBank.Br e PLN-Br

Nesta seção, apresentamos os experimentos desenvolvidos para contrastar a performance de sistemas de APS treinados em árvores sintáticas revisadas por humanos (*tree-banks* ou dados *gold-standard*) e árvores sintáticas sem revisão. Os **córpus** utilizados são o PropBank.Br versão 1.1 e a seleção anotada do PLN-Br. O primeiro contém 6.399 instâncias anotadas em 3.316 sentenças e o segundo contém 7.979 instâncias anotadas em 7.172 sentenças. Há uma disparidade na quantidade de instâncias de anotação, mas, como o **córpus** não é o mesmo, não é possível manter um ambiente de total igualdade. Utilizamos o sistema de [Alva-Manchego \(2013\)](#) na maioria dos experimentos por se tratar do sistema estado da arte na APS em PB. Também utilizamos o sistema de [Fonseca \(2013\)](#) em um experimento final para contrastar a sua performance com a do sistema de [Alva-Manchego \(2013\)](#). O sistema de [Fonseca \(2013\)](#) não utiliza árvores sintáticas e, portanto, não sofre com os erros sintáticos gerados pelo *parser*. No cenário do **córpus** PLN-Br, é interessante observar se os erros gerados pelo *parser* são significativos para fazer com que o sistema de [Fonseca \(2013\)](#) desempenhe melhor APS do que o sistema de [Alva-Manchego \(2013\)](#).

Os experimentos relatados a seguir avaliam o sistema de [Alva-Manchego \(2013\)](#) sobre diferentes conjuntos do **córpus** PLN-Br: o conjunto Concordância, composto pelas instâncias em que os anotadores concordaram plenamente e, portanto, não foram enviadas para adjudicação; o conjunto Adjudicados, composto pelas instâncias que fo-

ram enviadas à adjudicação; e o conjunto *Full*, composto por Concordância e Adjudicados. Primeiramente, realizamos a avaliação *10-fold cross-validation* para o conjunto em que houve concordância entre os anotadores do PLN-Br. O resultado apresentado na Tabela 30a nos mostra um indicativo razoável de qualidade na APS, uma vez que o conjunto Concordância possui aproximadamente 50% da quantidade de instâncias do PropBank.Br utilizado por Alva-Manchego (2013). Para contrastar com o conjunto Concordância, também foi verificada a qualidade da APS sobre o conjunto que foi enviado para adjudicação. O resultado apresentado na Tabela 30b nos mostra que o conjunto Adjudicados, que possui aproximadamente 80% da quantidade de instâncias do PropBank.Br e é aproximadamente 40% maior do que o conjunto Concordância, não possui uma anotação fácil (ou simples) de papéis semânticos. Podemos interpretar que, se o humano possui dificuldade na tarefa de anotação, a máquina também terá – justificando a diferença de 6,98 pontos de F1 nas performances do conjunto Concordância (72,71 de F1) e Adjudicados (65,73 de F1). Também podemos especular que os erros sintáticos, que causaram diferentes interpretações de anotação ao olhar humano, dificultaram o aprendizado automático da tarefa por aumentar a esparsidade dos dados.

Ao realizar o experimento de *10-fold cross-validation* sobre o conjunto *Full* do PLN-Br, ou seja, sobre ambos os conjuntos Concordância e Adjudicados, obtivemos o valor 69,12 de F1 (ver Tabela 31), que está entre os valores 65,73 (conjunto Adjudicado) e 72,71 (conjunto Concordância). Podemos especular que, apesar do conjunto Adjudicado possuir mais instâncias e ter resultado inferior ao Concordância, o Concordância supre o sistema com árvores sintáticas sem erros e também com árvores que possuem erros frequentes e que, portanto, são identificáveis por AM. Também é interessante comentar que a diferença de F1 entre o sistema treinado sobre os dados *gold-standard* do PropBank.Br e o do sistema treinado sobre as árvores sintáticas não revisadas do PLN-Br, é de 10,48 pontos de F1. A diferença em torno de 10 pontos de F1 entre um sistema treinado sobre árvores revisadas e árvores não revisadas já foi averiguada por Palmer et al. (2005) e Toutanova et al. (2008).

O contraste entre os valores de F1 para os corpúsculos PLN-Br e PropBank.Br é preocupante. Apesar da seleção feita sobre o corpúsculo PLN-Br respeitar uma distribuição que deveria favorecer a APS e esse corpúsculo ser aproximadamente 24% maior que o Prop-

	corr.	excess	missed	prec.	rec.	F1
Overall	3.640	1.401	1.332	72,23	73,22	72,71
A0	807	230	146	77,98	84,64	81,09
A1	1.853	491	600	79,13	75,56	77,28
A2	367	294	223	55,48	62,14	58,42
A3	10	17	43	45,67	26,82	27,83
A4	2	1	15	20,00	8,33	11,67
AM-ADV	9	12	13	28,00	26,67	24,56
AM-CAU	2	4	16	13,33	8,33	9,00
AM-DIR	0	0	0	–	–	–
AM-DIS	6	8	16	21,67	16,67	18,16
AM-EXT	16	4	11	75,00	60,83	64,33
AM-LOC	103	113	50	48,69	67,53	55,54
AM-MNR	80	75	57	52,57	60,55	54,88
AM-NEG	134	14	19	90,30	87,80	88,83
AM-PRD	0	0	1	0,00	0,00	0,00
AM-PRP	39	28	17	58,89	71,23	62,89
AM-REC	0	0	0	0,00	–	–
AM-TMP	212	110	105	65,75	67,07	66,03

(a) Resultados de *10-fold cross-validation* para o conjunto Concordância do PLN-Br.

	corr.	excess	missed	prec.	rec.	F1
Overall	7.164	4.001	3.469	64,16	67,37	65,73
A0	1.873	491	377	79,26	83,24	81,18
A1	2.703	1.208	1.142	69,12	70,29	69,68
A2	489	708	517	40,82	48,76	44,32
A3	14	36	82	28,39	14,14	18,00
A4	8	5	23	36,00	19,05	23,00
AM-ADV	157	134	119	54,37	57,50	55,49
AM-CAU	47	67	75	42,80	38,17	39,43
AM-DIR	0	5	13	0,00	0,00	0,00
AM-DIS	150	76	172	66,76	46,85	54,69
AM-EXT	56	35	43	62,70	58,19	58,55
AM-LOC	413	374	161	52,45	72,60	60,79
AM-MNR	200	281	219	41,81	47,78	44,48
AM-NEG	226	22	39	90,92	86,16	88,10
AM-PRD	46	125	124	27,44	27,53	26,91
AM-PRP	97	86	69	53,52	59,89	56,06
AM-REC	0	0	1	0,00	0,00	0,00
AM-TMP	684	338	271	67,24	71,64	69,26

(b) Resultados de *10-fold cross-validation* para o conjunto Adjudicados do PLN-Br.

Tabela 30: Resultados de *10-fold cross-validation* para os conjuntos Concordância e Adjudicados do PLN-Br, ao utilizar o sistema de [Alva-Manchego \(2013\)](#).

Bank.Br, os resultados nos mostram um sistema com 10,48 pontos de F1 inferior ao resultado obtido do sistema treinado por [Alva-Manchego \(2013\)](#) sobre o cópurs Prop-Bank.Br (69,12 de F1 contra 79,6). Os indícios de que o uso de *parser* automático afeta negativamente a performance de sistemas de APS se tornam mais evidentes ao contrastarmos os resultados da Tabela 30a e da Tabela 30b. A questão, no entanto, é sabermos se essa “queda na qualidade” de fato representa um sistema com pior desempenho, ou se a anotação nesse cenário com *parser* automático é tão difícil a ponto do valor de F1 do sistema treinado no PLN-Br (apresentado na Tabela 31) ser considerado bom. Para tanto, realizamos 2 experimentos: um anotando o cópurs PropBank.Br com o sistema treinado sobre o conjunto *Full* do cópurs PLN-Br e outro fazendo o caminho inverso – anotando o conjunto *Full* do cópurs PLN-Br com o sistema treinado sobre o cópurs PropBank.Br.

Podemos ver na Tabela 32a que a APS em árvores sintáticas revisadas, utilizando o sistema de [Alva-Manchego \(2013\)](#) treinado sobre árvores sintáticas não revisadas, é factível, apesar de sua qualidade não ser comparável ao de um sistema treinado sobre

	corr.	excess	missed	prec.	rec.	F1
Overall	11.047	5.335	4.537	67,45	70,89	69,12
A0	2.703	671	500	80,10	84,37	82,15
A1	4.630	1.713	1.664	73,01	73,57	73,28
A2	898	974	702	48,07	56,09	51,67
A3	25	64	124	28,62	16,72	20,85
A4	18	10	30	68,00	36,07	45,21
AM-ADV	159	150	139	50,00	52,37	50,64
AM-CAU	54	69	86	43,71	39,87	40,99
AM-DIR	0	4	13	0,00	0,00	0,00
AM-DIS	183	116	161	60,95	52,57	56,20
AM-EXT	72	41	54	63,05	57,06	59,51
AM-LOC	529	460	198	53,73	72,96	61,75
AM-MNR	302	352	254	46,42	54,34	49,84
AM-NEG	360	42	58	88,91	85,90	87,23
AM-PRD	37	123	134	24,33	20,79	21,96
AM-PRP	147	99	75	59,75	66,54	62,48
AM-REC	0	0	1	0,00	0,00	0,00
AM-TMP	930	447	342	67,53	73,14	70,12

Tabela 31: Resultados de *10-fold cross-validation* para o conjunto *Full* da seleção do PLN-Br, composto pelos conjuntos Concordância e Adjudicados, ao utilizar o sistema de [Alva-Manchego \(2013\)](#).

árvores revisadas (72,62 de F1 contra 79,6 de F1 de [Alva-Manchego \(2013\)](#)). Verificamos inclusive que o valor 72,62 obtido é maior que o 69,12 do *10-fold cross-validation* do sistema treinado na seleção do PLN-Br. Logo, percebemos que, além de um sistema treinado sobre árvores sintáticas não revisadas conseguir anotar árvores sintáticas que contenham erros, ele consegue anotar árvores perfeitas com uma performance superior ao que ele anota em seu próprio cenário. Isso significa que **treinar um sistema de APS sobre árvores sintáticas ruins faz com que o sistema capture padrões de árvores boas e também aprenda a lidar com o ruído contido em árvores ruins**.

Fazendo o caminho inverso de anotação, podemos verificar na Tabela [32b](#) que a qualidade da APS das árvores problemáticas do PLN-Br, utilizando o sistema treinado sobre o PropBank.Br, é 14,36 pontos de F1 inferior ao resultado do *10-fold cross-validation* do conjunto *Full* do PLN-Br (69,12 de F1 contra 54,76). Desejamos saber se essa diferença é estatisticamente significativa e, para isso, verificamos as seguintes hipóteses via Teste de Wilcoxon:

	corr.	excess	missed	prec.	rec.	F1
Overall	9.622	2.958	4.298	76,49	69,12	72,62
A0	2.281	463	778	83,13	74,57	78,61
A1	4.187	863	1142	82,91	78,57	80,68
A2	772	526	613	59,48	55,74	57,55
A3	20	60	127	25,00	13,61	17,62
A4	35	26	79	57,38	30,70	40,00
A5	0	0	1	0,00	0,00	0,00
AM-ADV	162	95	212	63,04	43,32	51,35
AM-CAU	75	55	77	57,69	49,34	53,19
AM-DIR	0	0	13	0,00	0,00	0,00
AM-DIS	180	87	129	67,42	58,25	62,50
AM-EXT	32	20	46	61,54	41,03	49,23
AM-LOC	523	257	190	67,05	73,35	70,06
AM-MNR	168	194	239	46,41	41,28	43,69
AM-NEG	315	47	28	87,02	91,84	89,36
AM-PRD	14	62	169	18,42	7,65	10,81
AM-PRP	97	52	56	65,10	63,40	64,24
AM-REC	0	0	8	0,00	0,00	0,00
AM-TMP	761	151	388	83,44	66,23	73,85

	corr.	excess	missed	prec.	rec.	F1
Overall	7.972	5.559	7.612	58,92	51,16	54,76
A0	2.113	1.040	1.090	67,02	65,97	66,49
A1	3.426	1.543	2.868	68,95	54,43	60,84
A2	478	526	1.122	47,61	29,88	36,71
A3	11	24	138	31,43	7,38	11,96
A4	8	13	40	38,10	16,67	23,19
A5	0	0	0	–	–	–
AM-ADV	100	179	198	35,84	33,56	34,66
AM-CAU	36	106	104	25,35	25,71	25,53
AM-DIR	0	2	13	0,00	0,00	0,00
AM-DIS	106	74	238	58,89	30,81	40,46
AM-EXT	63	45	63	58,33	50,00	53,85
AM-LOC	412	520	315	44,21	56,67	49,67
AM-MNR	242	472	314	33,89	43,53	38,11
AM-NEG	204	49	214	80,63	48,80	60,80
AM-PRD	8	245	163	3,16	4,68	3,77
AM-PRP	101	126	121	44,49	45,50	44,99
AM-REC	0	25	1	0,00	0,00	0,00
AM-TMP	664	570	608	53,81	52,20	52,99

(a) Resultados da anotação do PropBank.Br utilizando o sistema de APS treinado sobre o conjunto *Full* do PLN-Br.

(b) Resultados da anotação do conjunto *Full* do PLN-Br, utilizando o sistema de APS treinado sobre o PropBank.Br.

Tabela 32: Resultados dos experimentos de APS verificando a troca de contexto entre as fases de treinamento e teste para árvores sintáticas revisadas e não revisadas, ao utilizar o sistema de [Alva-Manchego \(2013\)](#).

H0: A hipótese nula é de que não há diferença significativa entre a performance dos sistemas.

H1: A hipótese alternativa é de que há diferença significativa na performance, sendo que o sistema treinado sobre o PLN-Br possui maior valor de F1.

Com um *p-value* de 0,001953, temos **evidências de que as diferenças de performances são significativas**. O valor de 54,76 de F1 na anotação de árvores sintáticas não revisadas por um sistema treinado em dados *gold-standard* é inferior ao valor de 69,12 do *10-fold cross-validation* da seleção do PLN-Br e essa diferença é estatisticamente significativa. Esse resultado nos mostra que um sistema treinado sobre árvores sintáticas perfeitas possui maiores dificuldades ao anotar árvores sintáticas defeituosas. Como o corpus *gold-standard* do PropBank.Br não contém exemplos ruidosos, um mínimo desvio do padrão correto de árvore sintática é o suficiente para fazer com que o

sistema erre na anotação.

Para reforçar a análise dos resultados, anotamos os conjuntos Concordância e Adjudicados do PLN-Br com o sistema treinado sobre o PropBank.Br. A Tabela 33a apresenta o resultado de 58,96 de F1 na anotação do conjunto Concordância do PLN-Br pelo sistema treinado no PropBank.Br e a Tabela 33b apresenta 52,66 de F1 para a anotação do conjunto enviado à adjudicação. A diferença na performance da anotação dos conjuntos foi de 6,3 pontos de F1, enquanto a diferença do *10-fold cross-validation* realizado nesses conjuntos foi de 6,98 pontos de F1. Como a diferença na performance de anotação dos conjuntos é próxima, entendemos que, de fato, **existe um limiar de complexidade sintática que distingue os conjuntos Concordância e Adjudicados do PLN-Br.**

	corr.	excess	missed	prec.	rec.	F1
Overall	2.803	1.718	2.184	62,00	56,21	58,96
A0	631	372	322	62,91	66,21	64,52
A1	1.455	491	998	74,77	59,32	66,15
A2	210	144	380	59,32	35,59	44,49
A3	4	10	49	28,57	7,55	11,94
A4	4	4	13	50,00	23,53	32,00
AM-ADV	11	58	11	15,94	50,00	24,18
AM-CAU	5	24	13	17,24	27,78	21,28
AM-DIR	0	1	0	0,00	0,00	0,00
AM-DIS	11	27	11	28,95	50,00	36,67
AM-EXT	14	14	13	50,00	51,85	50,91
AM-LOC	92	144	61	38,98	60,13	47,30
AM-MNR	73	130	64	35,96	53,28	42,94
AM-NEG	81	14	72	85,26	52,94	65,32
AM-PRD	0	58	1	0,00	0,00	0,00
AM-PRP	27	44	29	38,03	48,21	42,52
AM-REC	0	6	0	0,00	0,00	0,00
AM-TMP	185	177	132	51,10	58,36	54,49

	corr.	excess	missed	prec.	rec.	F1
Overall	5.166	3.844	5.444	57,34	48,69	52,66
A0	1.482	668	768	68,93	65,87	67,36
A1	1.972	1.051	1.873	65,23	51,29	57,43
A2	264	386	742	40,62	26,24	31,88
A3	7	14	89	33,33	7,29	11,97
A4	4	9	27	30,77	12,90	18,18
AM-ADV	89	121	187	42,38	32,25	36,63
AM-CAU	31	82	91	27,43	25,41	26,38
AM-DIR	0	1	13	0,00	0,00	0,00
AM-DIS	95	47	227	66,90	29,50	40,95
AM-EXT	49	31	50	61,25	49,49	54,75
AM-LOC	320	376	254	45,98	55,75	50,39
AM-MNR	169	342	250	33,07	40,33	36,34
AM-NEG	123	35	142	77,85	46,42	58,16
AM-PRD	8	187	162	4,10	4,71	4,38
AM-PRP	74	82	92	47,44	44,58	45,96
AM-REC	0	19	1	0,00	0,00	0,00
AM-TMP	479	393	476	54,93	50,16	52,44

(a) Resultados da anotação no conjunto Concordância do PLN-Br.

(b) Resultados da anotação no conjunto Adjudicados do PLN-Br.

Tabela 33: Resultados da anotação dos conjuntos Concordância e Adjudicados do PLN-Br utilizando o sistema de APS treinado sobre o PropBank.Br, ao utilizar o sistema de [Alva-Manchego \(2013\)](#)

Por fim, realizamos um experimento *10-fold cross-validation* com o sistema de [Fonseca \(2013\)](#) e o conjunto *Full* do corpus PLN-Br. O valor de 53,58 de F1, apresentado na Tabela 34, mostra que o desempenho obtido pelo sistema de [Fonseca \(2013\)](#) é significativamente inferior ao valor de 69,12 obtido pelo sistema de [Alva-Manchego \(2013\)](#).

	corr.	excess	missed	prec.	rec.	F1
Overall	8.038	6.475	7.464	55,44	51,87	53,58
A0	2.318	1.058	911	68,72	71,82	70,18
A1	3.497	2.979	2.755	54,13	55,93	55,00
A2	440	938	11.16	31,97	28,26	29,94
A3	0	6	150	0,00	0,00	0,00
A4	0	0	48	0,00	0,00	0,00
AM-ADV	90	63	208	62,09	30,84	39,95
AM-CAU	5	23	134	10,19	4,36	6,04
AM-DIR	0	0	13	0,00	0,00	0,00
AM-DIS	121	94	222	57,06	35,73	42,93
AM-EXT	53	34	72	61,13	43,78	49,81
AM-LOC	282	469	444	37,57	38,93	37,98
AM-MNR	143	180	412	45,09	26,13	32,86
AM-NEG	355	40	59	89,81	86,34	87,89
AM-PRD	6	46	163	16,00	3,33	5,02
AM-PRP	64	100	154	38,15	29,89	33,18
AM-REC	0	0	1	0,00	0,00	0,00
AM-TMP	664	445	602	59,92	52,51	55,83

Tabela 34: Resultados de *10-fold cross-validation* para o conjunto *Full* da seleção do PLN-Br, composto pelos conjuntos Concordância e Adjudicados, utilizando o sistema de [Fonseca \(2013\)](#).

Os resultados obtidos nessa seção mostram que um sistema de APS sensível a erros sintáticos, treinado sobre dados ruidosos (árvores sintáticas com erros), desempenha uma melhor APS sobre dados ruidosos do que ao ser treinado sobre árvores sem erros. Assim, **validamos a Hipótese 4 desse trabalho**. Percebemos, ainda, que o sistema de [Fonseca \(2013\)](#) desempenha APS inferior ao de [Alva-Manchego \(2013\)](#), mesmo em um cenário desfavorável para o sistema do segundo autor. Acreditamos que, para o PB, o uso de árvores sintáticas para a tarefa de APS ainda é o meio mais promissor e, portanto, esforços são bem vindos para o desenvolvimento de *parsers* sintáticos. Entendemos ainda que, em um cenário real de aplicação (processamento *on the fly*), os dados são passados diretamente do *parser* sintático para o sistema de APS, sem intervenção humana para corrigir as árvores. Assim, podemos afirmar que, em um cenário real de aplicação, o sistema de [Alva-Manchego \(2013\)](#) treinado sobre a seleção do PLN-Br desempenha uma melhor anotação de textos jornalísticos do que o sistema treinado sobre o PropBank.Br.

5.5 Experimentos de APS no Gênero de Opiniões sobre Produtos da Web: Buscapé

Nesta seção apresentamos os experimentos desenvolvidos para verificar o desafio da tarefa de anotar papéis semânticos em textos extraídos da web contendo opiniões sobre produtos. Esse gênero, distinto do jornalístico, contém diversos erros que afetam o desempenho do *parser* sintático e, portanto, espera-se que também afete o desempenho de sistemas de APS que dependem de árvores sintáticas. Assim, realizamos experimentos com o sistema de [Alva-Manchego \(2013\)](#), dependente de árvores sintáticas, e com o sistema de [Fonseca \(2013\)](#). Os experimentos relatados a seguir agem sobre diferentes conjuntos do *corpus* Buscapé: o conjunto *Concordância*, composto pelas instâncias em que os anotadores concordaram plenamente e, portanto, não foram enviadas para adjudicação; o conjunto *Adjudicados*, composto pelas instâncias que foram enviadas à adjudicação; e o conjunto *Full*, composto por *Concordância* e *Adjudicados*. A maioria dos experimentos avaliam o desempenho do sistema de [Alva-Manchego \(2013\)](#), para avaliar a dependência do desempenho do seu sistema sobre árvores sintáticas revisadas e não revisadas. Também realizamos dois experimentos para contrastar o desempenho dos sistemas de [Alva-Manchego \(2013\)](#) e [Fonseca \(2013\)](#), ao serem treinados sobre os *corpos* PLN-Br e PropBank.Br e avaliados sobre o *corpus* Buscapé.

Realizamos um experimento *10-fold cross-validation* com o sistema de [Alva-Manchego \(2013\)](#), ao utilizar o conjunto em que houve concordância entre os anotadores do Buscapé. O resultado apresentado na Tabela 35a nos mostra um valor próximo ao obtido para o conjunto *Concordância* do PLN-Br (68,72 de F1 para o Buscapé e 72,71 de F1 para o PLNBr). As anotações de Arg0 e Arg1 (os papéis mais frequentes) são razoáveis. No entanto, percebemos que os demais papéis obtiveram resultados muito aquém do desejado. Isso se dá pela falta de instâncias de anotação (por exemplo, pode-se perceber na Tabela 35a que há 3 acertos de ArgM-Tmp e 11 falhas, totalizando 14 casos de anotação. Para o mesmo experimento no PLN-Br, tínhamos 317 casos de anotação para o conjunto *Concordância*.). Assim, temos indicativos de que, como o gênero apresenta sentenças mais curtas e simples, um sistema de APS treinado sobre ele não obterá bom

desempenho.

Para contrastar com o conjunto Concordância, também foi verificada a qualidade da APS do sistema de [Alva-Manchego \(2013\)](#) sobre o conjunto que foi enviado para adjudicação. O resultado apresentado na Tabela 35b nos mostra que o conjunto Adjudicados, que é aproximadamente 15% maior do que o conjunto Concordância, não possui anotação fácil (ou simples) de papéis semânticos, obtendo um resultado muito inferior ao experimento similar realizado no PLN-Br (56,07 de F1 para o Buscapé e 65,73 de F1 para o PLN-Br). As anotações de Arg0 e Arg1 foram razoáveis, porém a anotação do ArgM-Tmp, que aqui contém 50 casos de anotação, também foi muito baixa. Vale lembrar que o ArgM-Tmp é um dos papéis semânticos mais fáceis de serem anotados, devido à sua falta de variedade lexical e sintática. Interpretamos que o *parser* impactou fortemente nos resultados obtidos aqui nesse gênero. Assim como no PLN-Br, se os humanos discordam na anotação, a máquina também não obtém bons resultados ao anotar o conjunto de dados.

	corr.	excess	missed	prec.	rec.	F1
Overall	349	168	152	67,39	69,51	68,41
A0	103	29	14	79,02	88,22	83,04
A1	157	76	58	67,29	72,97	69,85
A2	32	20	22	64,76	57,86	58,08
A3	0	0	2	0,00	0,00	0,00
A4	0	0	3	0,00	0,00	0,00
AM-ADV	3	1	1	30,00	30,00	30,00
AM-CAU	2	0	2	20,00	20,00	20,00
AM-DIS	0	0	2	0,00	0,00	0,00
AM-EXT	3	4	5	20,00	20,00	20,00
AM-LOC	5	9	9	26,67	20,00	21,00
AM-MNR	10	13	10	43,33	49,17	43,81
AM-NEG	28	2	4	93,75	89,67	89,56
AM-PRD	0	0	0	–	–	–
AM-PRP	3	5	4	20,00	25,00	21,67
AM-TMP	3	9	11	16,67	20,00	18,00

	corr.	excess	missed	prec.	rec.	F1
Overall	445	345	357	56,68	55,7	56,07
A0	121	47	29	73,26	80,20	76,04
A1	153	98	96	61,77	61,25	61,19
A2	45	41	41	50,37	51,85	50,66
A3	0	2	4	0,00	0,00	0,00
A4	1	4	6	10,00	3,33	5,00
AM-ADV	20	18	15	41,67	46,50	42,51
AM-CAU	5	18	22	20,83	17,33	17,64
AM-DIS	12	14	25	39,33	33,33	33,19
AM-EXT	4	3	7	33,33	35,00	31,67
AM-LOC	14	19	13	42,33	53,33	43,28
AM-MNR	21	35	24	41,50	44,72	39,51
AM-NEG	35	7	15	83,14	73,39	76,74
AM-PRD	1	3	8	10,00	3,33	5,00
AM-PRP	5	12	10	32,50	29,17	30,50
AM-TMP	8	24	42	35,00	14,18	18,29

(a) Resultados de *10-fold cross-validation* para o conjunto Concordância da seleção do Buscapé.

(b) Resultados de *10-fold cross-validation* para o conjunto Adjudicados da seleção do Buscapé.

Tabela 35: Resultados de *10-fold cross-validation* para os conjuntos Concordância e Adjudicados da seleção do Buscapé, ao utilizar o sistema de [Alva-Manchego \(2013\)](#).

Ao realizar o experimento *10-fold cross-validation* com o sistema de [Alva-Manchego](#)

(2013) sobre o conjunto *Full* do Buscapé, ou seja, sobre ambos os conjuntos Concor-
dância e Adjudicados, obtivemos o valor 64,79 de F1 (ver Tabela 36), que está entre os
valores 56,07 (conjunto Adjudicado) e 68,41 (conjunto Concordância). Esse resultado,
próximo ao obtido nos experimentos com a seleção realizada sobre o corpus PLN-Br,
não é significativo em relação ao estado da arte e mostra que deve ser investido esforço
na tarefa de APS sobre esse gênero. Uma solução seria treinar sistemas de APS sobre
uma quantidade exaustiva de dados, a fim que o sistema aprenda a lidar com os ruídos
existentes. Outra solução seria melhorar o pré-processamento do gênero para tornar os
textos mais próximos do gênero jornalístico e ainda melhorar o desempenho do *parser*.
Porém, a literatura nos mostra que, em cenários como esse, esforços são realizados a fim
de treinar um sistema de APS em um gênero genérico (como o jornalístico) e testá-lo
no gênero alvo, o qual não se tem muito conhecimento e nem dados para treinamento.
Assim, realizamos alguns experimentos de anotação dos conjuntos Concordância, Ad-
judicados e *Full*, do Buscapé, com os sistemas treinados sobre os corpus PropBank.Br e
PLN-Br.

	corr.	excess	missed	prec.	rec.	F1
Overall	850	476	448	64,17	65,54	64,79
A0	232	65	35	77,67	86,30	81,72
A1	326	163	138	66,88	70,48	68,52
A2	83	55	57	61,50	61,31	59,73
A3	1	3	5	3,33	10,00	5,00
A4	2	4	8	15,00	13,33	14,00
AM-ADV	17	23	22	45,83	47,17	41,95
AM-CAU	7	15	24	39,83	25,00	24,49
AM-DIS	11	14	28	28,00	23,43	24,16
AM-EXT	9	5	10	56,67	54,17	50,33
AM-LOC	25	22	16	57,92	62,14	56,85
AM-MNR	34	43	31	46,77	53,45	46,76
AM-NEG	66	7	16	90,74	79,56	84,38
AM-PRD	0	5	9	0,00	0,00	0,00
AM-PRP	11	15	11	36,50	43,67	38,85
AM-TMP	26	37	38	43,55	41,81	40,48

Tabela 36: Resultados de *10-fold cross-validation* para o conjunto *Full* da seleção do
Buscapé, composto pelos conjuntos Concordância e Adjudicados, ao utilizar o sistema
de Alva-Manchego (2013).

Desejamos saber qual sistema possui melhor desempenho em anotar o *corpus* Buscapé: o sistema treinado sobre o PLN-Br ou o treinado sobre o PropBank.Br. Para tanto, desejamos verificar as seguintes hipóteses via Teste de Wilcoxon:

H0: A hipótese nula é de que não há diferença significativa entre a performance dos sistemas treinados sobre o PLN-Br e o PropBank.Br e avaliados no Buscapé.

H1: A hipótese alternativa é de que há diferença significativa na performance, sendo que o sistema treinado sobre o PLN-Br possui maior valor de F1.

Com o *p-value* de 0,003906, **há evidências para rejeitar H0 no experimento sobre a anotação do conjunto Concordância do Buscapé** e que, portanto, os valores apresentados na Tabela 37a e na Tabela 37b indicam que **o sistema treinado sobre os dados do PLN-Br obteve melhores resultados na anotação do conjunto Concordância do Buscapé**. No entanto, com *p-value* de 0,06445, não temos evidência para refutar H0 e concluir se o sistema treinado sobre o PLN-Br possui resultados melhores que os obtidos pelo *10-fold cross-validation* do conjunto Concordância do Buscapé.

Com o *p-value* de 0,003906, **há evidências para rejeitar H0 no experimento sobre a anotação do conjunto Adjudicados do Buscapé** e, portanto, os valores apresentados na Tabela 38a e na Tabela 38b indicam que **o sistema treinado sobre os dados do PLN-Br obteve melhor desempenho na anotação do conjunto Adjudicados do Buscapé**. Ainda, com um *p-value* de 0.01367 podemos rejeitar H0 e ter evidências de que o sistema treinado sobre o PLN-Br anota melhor o conjunto Adjudicados do Buscapé que o próprio sistema treinado sobre o conjunto Adjudicados do Buscapé. Os resultados também mostram que, assim como nos experimentos da Seção 5.4, há uma maior dificuldade na anotação dos dados enviados à adjudicação e o sistema treinado sobre dados ruidosos (PLN-Br) obteve melhor desempenho do que o treinado sobre dados sem ruídos (PropBank.Br).

Por fim, com o *p-value* de 0,001953, **há evidências para rejeitar H0 no experimento sobre a anotação do conjunto Full do Buscapé** e, portanto, os valores apresentados na Tabela 40a e na Tabela 39b indicam que o sistema treinado sobre o PLN-Br obteve melhores resultados do que o treinado sobre o PropBank.Br na anotação do conjunto *Full* do Buscapé. Isso reforça o resultado obtido na Seção 5.4, em que o sistema

	corr.	excess	missed	prec.	rec.	F1
Overall	326	212	175	60,59	65,07	62,75
A0	107	59	10	64,46	91,45	75,62
A1	136	58	79	70,10	63,26	66,50
A2	18	12	36	60,00	33,33	42,86
A3	0	2	2	0,00	0,00	0,00
A4	1	1	2	50,00	33,33	40,00
AM-ADV	4	6	0	40,00	100,00	57,14
AM-CAU	3	2	1	60,00	75,00	66,67
AM-DIS	0	1	2	0,00	0,00	0,00
AM-EXT	3	2	5	60,00	37,50	46,15
AM-LOC	8	8	6	50,00	57,14	53,33
AM-MNR	9	25	11	26,47	45,00	33,33
AM-NEG	23	2	9	92,00	71,88	80,70
AM-PRD	0	13	0	0,00	0,00	0,00
AM-PRP	5	8	2	38,46	71,43	50,00
AM-TMP	9	13	5	40,91	64,29	50,00

(a) Resultados da anotação no conjunto Concordância do Buscapé com o uso do sistema treinado sobre o PropBank.Br.

	corr.	excess	missed	prec.	rec.	F1
Overall	378	177	123	68,11	75,45	71,59
A0	112	38	5	74,67	95,73	83,90
A1	154	58	61	72,64	71,63	72,13
A2	35	32	19	52,24	64,81	57,85
A3	1	5	1	16,67	50,00	25,00
A4	1	0	2	100,00	33,33	50,00
AM-ADV	4	6	0	40,00	100,00	57,14
AM-CAU	2	1	2	66,67	50,00	57,14
AM-DIS	0	2	2	0,00	0,00	0,00
AM-EXT	4	3	4	57,14	50,00	53,33
AM-LOC	9	5	5	64,29	64,29	64,29
AM-MNR	12	15	8	44,44	60,00	51,06
AM-NEG	30	2	2	93,75	93,75	93,75
AM-PRD	0	4	0	0,00	0,00	0,00
AM-PRP	5	3	2	62,50	71,43	66,67
AM-TMP	9	3	5	75,00	64,29	69,23

(b) Resultados da anotação no conjunto Concordância do Buscapé com o uso do sistema treinado sobre o conjunto *Full* do PLN-Br

Tabela 37: Resultados da anotação do conjunto Concordância do Buscapé utilizando os sistema de APS treinados sobre o PropBank.Br e sobre o PLN-Br, ao utilizar o sistema de [Alva-Manchego \(2013\)](#).

treinado sobre o PropBank.Br também não obteve boa performance ao anotar as árvores sintáticas não revisadas do PLN-Br.

Também realizamos experimentos com o sistema de [Fonseca \(2013\)](#). Já verificamos que esse sistema teve um desempenho pior na APS sobre o córpus PLN-Br do que o sistema de [Alva-Manchego \(2013\)](#). Como o cenário do córpus Buscapé conta com uma maior incidência de erros sintáticos, oriundos de uma linguagem que foge das normas cultas de escrita, é interessante verificar o desempenho do sistema de [Fonseca \(2013\)](#) sobre o córpus Buscapé. A avaliação do sistema sobre o córpus Buscapé resultou em uma performance de 46,23 de F1 ao ser treinado sobre o córpus PropBank.Br e 48,40 ao ser treinado sobre o córpus PLN-Br (valores podem ser conferidos na Tabela 40). Ambos os valores são inferiores ao desempenho do sistema de [Alva-Manchego \(2013\)](#) ao anotar o córpus Buscapé quando treinado sobre o córpus PropBank.Br (57,72 de F1) e PLN-Br (65,34 de F1). Isso nos mostra que, mesmo com um córpus mais propício a gerar erros sintáticos, o sistema de [Alva-Manchego \(2013\)](#) desempenhou melhor na

	corr.	excess	missed	prec.	rec.	F1
Overall	422	329	380	56,19	52,62	54,35
A0	107	64	43	62,57	71,33	66,67
A1	148	99	101	59,92	59,44	59,68
A2	25	23	61	52,08	29,07	37,31
A3	1	2	3	33,33	25,00	28,57
A4	1	0	6	100,00	14,29	25,00
AM-ADV	20	12	15	62,50	57,14	59,70
AM-CAU	6	12	21	33,33	22,22	26,67
AM-DIS	9	5	28	64,29	24,32	35,29
AM-EXT	5	4	6	55,56	45,45	50,00
AM-LOC	15	16	12	48,39	55,56	51,72
AM-MNR	19	24	26	44,19	42,22	43,18
AM-NEG	34	6	16	85,00	68,00	75,56
AM-PRD	2	18	7	10,00	22,22	13,79
AM-PRP	6	14	9	30,00	40,00	34,29
AM-TMP	24	30	26	44,44	48,00	46,15

	corr.	excess	missed	prec.	rec.	F1
Overall	489	313	307	60,97	61,43	61,20
A0	127	23	48	84,67	72,57	78,15
A1	169	80	101	67,87	62,59	65,13
A2	38	48	43	44,19	46,91	45,51
A3	2	2	10	50,00	16,67	25,00
A4	1	6	2	14,29	33,33	20,00
AM-ADV	14	21	14	40,00	50,00	44,44
AM-CAU	7	20	8	25,93	46,67	33,33
AM-DIS	10	27	6	27,03	62,50	37,74
AM-EXT	8	3	9	72,73	47,06	57,14
AM-LOC	17	10	14	62,96	54,84	58,62
AM-MNR	22	23	22	48,89	50,00	49,44
AM-NEG	38	12	3	76,00	92,68	83,52
AM-PRD	2	7	5	22,22	28,57	25,00
AM-PRP	9	6	12	60,00	42,86	50,00
AM-TMP	25	25	10	50,00	71,43	58,82

(a) Resultados da anotação no conjunto Adjudicados do Buscapé com o uso do sistema treinado sobre o PropBank.Br.

(b) Resultados da anotação no conjunto Adjudicados do Buscapé com o uso do sistema treinado sobre o conjunto *Full* do PLN-Br

Tabela 38: Resultados da anotação do conjunto Adjudicados do Buscapé utilizando os sistema de APS treinados sobre o PropBank.Br e sobre o PLN-Br, ao utilizar o sistema de [Alva-Manchego \(2013\)](#).

APS do que o sistema de [Fonseca \(2013\)](#).

Verificamos que, independentemente do gênero jornalístico ou de opinião, **sistemas de APS treinados sobre dados ruidosos desempenham melhor APS quando aplicados sobre dados ruidosos e que, sistemas treinados sobre dados que não contém ruídos, não conseguem lidar com os desafios gerados pelos erros de *parsing***. Também verificamos que, independentemente do gênero testado, o conjunto em que houve plena concordância entre os anotadores também obteve melhores resultados quando avaliado por um sistema de APS, enquanto o conjunto em que houve discordância entre os anotadores obteve resultados significativamente inferiores.

Finalmente, sabendo que o *10-fold cross-validation* do sistema de [Alva-Manchego \(2013\)](#), treinado sobre o PLN-Br, obteve 69,12 de F1 (ver Tabela 31) e, ao trocar de gênero, anotou o Buscapé com 65,34 de F1 (ver Tabela 39b), tivemos uma queda de 3,78 ponto de F1. Já para o sistema treinado sobre o PropBank.Br, cujo valor de *10-fold cross-validation* foi de 79,6 de F1 e a anotação sobre o Buscapé obteve 57,72 pontos de

	corr.	excess	missed	prec.	rec.	F1
Overall	748	555	541	57,41	58,03	57,72
A0	214	53	123	80,15	63,50	70,86
A1	284	180	157	61,21	64,40	62,76
A2	43	97	35	30,71	55,13	39,45
A3	1	5	4	16,67	20,00	18,18
A4	2	8	1	20,00	66,67	30,77
AM-ADV	24	15	18	61,54	57,14	59,26
AM-CAU	9	22	14	29,03	39,13	33,33
AM-DIS	9	30	6	23,08	60,00	33,33
AM-EXT	8	11	6	42,11	57,14	48,48
AM-LOC	23	18	24	56,10	48,94	52,27
AM-MNR	28	37	49	43,08	36,36	39,44
AM-NEG	57	25	8	69,51	87,69	77,55
AM-PRD	2	7	31	22,22	6,06	9,52
AM-PRP	11	11	22	50,00	33,33	40,00
AM-TMP	33	31	43	51,56	43,42	47,14

	corr.	excess	missed	prec.	rec.	F1
Overall	867	436	484	66,54	64,17	65,34
A0	239	28	86	89,51	73,54	80,74
A1	323	141	159	69,61	67,01	68,29
A2	73	67	75	52,14	49,32	50,69
A3	3	3	15	50,00	16,67	25,00
A4	2	8	2	20,00	50,00	28,57
AM-ADV	18	21	20	46,15	47,37	46,75
AM-CAU	9	22	9	29,03	50,00	36,73
AM-DIS	10	29	8	25,64	55,56	35,09
AM-EXT	12	7	12	63,16	50,00	55,81
AM-LOC	26	15	19	63,41	57,78	60,47
AM-MNR	34	31	37	52,31	47,89	50,00
AM-NEG	68	14	5	82,93	93,15	87,74
AM-PRD	2	7	9	22,22	18,18	20,00
AM-PRP	14	8	15	63,64	48,28	54,90
AM-TMP	34	30	13	53,12	72,34	61,26

(a) Resultados da anotação no conjunto *Full* do Buscapé com o uso do sistema treinado sobre o PropBank.Br.

(b) Resultados da anotação no conjunto *Full* do Buscapé com o uso do sistema treinado sobre o conjunto *Full* do PLN-Br

Tabela 39: Resultados da anotação do conjunto *Full* do Buscapé utilizando os sistema de APS treinados sobre o o PropBank.Br. e sobre o PLN-Br, ao utilizar o sistema de [Alva-Manchego \(2013\)](#).

F1 (ver Tabela 40a), a queda de performance ao trocar de gênero foi de 21,88 pontos de F1. Sabemos que o sistema treinado no PropBank.Br possui grandes dificuldades para anotar árvores sintáticas não revisadas, enquanto o sistema treinado no corpus PLN-Br é mais adequado para tal tarefa. Assim, para validar a Hipótese 5, devemos considerar apenas o resultado do PLN-Br, pois assim descartamos a variável “sistema não treinado sobre árvores problemáticas”. Assim, verificamos que a queda de desempenho ao trocar de gênero entre as fases de treinamento e teste foi inferior ao esperado na Hipótese 5 e, portanto, **não validamos a Hipótese 5 desse trabalho**. A perda de 3,78 pontos de F1 obtida com o sistema treinado sobre o PLN-Br e avaliado no Buscapé é inferior aos 10% de perda reportados por [Pradhan et al. \(2008\)](#). Isso mostra que nossos resultados são promissores.

	corr.	excess	missed	prec.	rec.	F1
Overall	634	806	669	4403	48,66	46,23
A0	217	187	50	53,71	81,27	64,68
A1	246	363	218	40,39	53,02	45,85
A2	17	37	123	31,48	12,14	17,53
A3	0	1	6	0,00	0,00	0,00
A4	0	0	10	0,00	0,00	0,00
AM-ADV	19	35	20	35,19	48,72	40,86
AM-CAU	3	7	28	30,00	9,68	14,63
AM-DIS	9	11	30	45,00	23,08	30,51
AM-EXT	5	8	14	38,46	26,32	31,25
AM-LOC	15	37	26	28,85	36,59	32,26
AM-MNR	20	47	45	29,85	30,77	30,30
AM-NEG	68	15	14	81,93	82,93	82,42
AM-PRD	0	1	9	0,00	0,00	0,00
AM-PRP	0	0	22	0,00	0,00	0,00
AM-TMP	15	26	49	36,59	23,44	28,57

	corr.	excess	missed	prec.	rec.	F1
Overall	626	658	677	48,75	48,04	48,40
A0	215	129	52	62,50	80,52	70,38
A1	216	309	248	41,14	46,55	43,68
A2	28	83	112	25,23	20,00	22,31
A3	0	0	6	0,00	0,00	0,00
A4	0	0	10	0,00	0,00	0,00
AM-ADV	19	5	20	79,17	48,72	60,32
AM-CAU	2	4	29	33,33	6,45	10,81
AM-DIS	19	16	20	54,29	48,72	51,35
AM-EXT	6	13	13	31,58	31,58	31,58
AM-LOC	16	30	25	34,78	39,02	36,78
AM-MNR	14	29	51	32,56	21,54	25,93
AM-NEG	74	8	8	90,24	90,24	90,24
AM-PRD	0	5	9	0,00	0,00	0,00
AM-PRP	0	10	22	0,00	0,00	0,00
AM-TMP	17	17	47	50,00	26,56	34,69

(a) Resultados da anotação no conjunto *Full* do Buscapé com o uso do sistema treinado sobre o PropBank.Br.

(b) Resultados da anotação no conjunto *Full* do Buscapé com o uso do sistema treinado sobre o conjunto *Full* do PLN-Br

Tabela 40: Resultados da anotação do conjunto *Full* do Buscapé utilizando os sistema de APS treinados sobre o o PropBank.Br. e sobre o PLN-Br, ao utilizar o sistema de [Fonseca \(2013\)](#).

5.6 Considerações Finais

Nesse capítulo, realizamos os experimentos referentes aos métodos desenvolvidos ao longo do projeto de mestrado e também a verificação das hipóteses de pesquisa levantadas no projeto. Foi possível investigar as hipóteses levantadas para esse trabalho de mestrado e resultados interessantes foram encontrados. A discussão sobre os resultados obtidos é apresentada no Capítulo 6.

Conclusões

Nesse trabalho de mestrado foram utilizados os *córpus* PropBank.Br versão 1.1, PLN-Br e Buscapé para aprendizagem e avaliação da tarefa de APS, com vistas a melhorar o estado da arte dessa tarefa para o português do Brasil.

O primeiro deles é o *córpus* utilizado por [Alva-Manchego \(2013\)](#) em seu trabalho sobre APS para a língua portuguesa. Esse *córpus* possui árvores sintáticas revisadas, ou seja, sem erros. Já nesse projeto de mestrado, que se insere no projeto PROS@do convênio entre a empresa SAMSUNG Eletrônica da Amazônia e o ICMC/USP, foi selecionado o PLN-Br para ampliar o PropBank.Br por meio de uma seleção de sentenças focada na diversidade e frequência dos verbos na língua. Foi apresentado o processo de seleção de instâncias de anotação de papéis semânticos do *córpus* PLN-Br, bem como a distribuição do *córpus* em valores quantitativos. Também foi descrito o procedimento de anotação do *córpus*, bem como os valores de Kappa para a concordância entre os pares de anotadores. Diferentemente do PropBank.Br versão 1.1, as árvores sintáticas da seleção do PLN-Br não sofreram revisão humana, o que indica a existência de erros de *parsing*. Acreditamos que esse cenário é favorável para a APS em um cenário real de aplicação (dados passados por *parser* automático, sem revisão humana), pois sistemas automáticos de APS estarão lidando com dados de aprendizagem que também contém

inconsistências.

Apresentamos também o procedimento de seleção de instâncias do *córpus Buscapé*, o processo de anotação humana de papéis semânticos e os resultados quantitativos da anotação. A seleção do *córpus Buscapé* nos possibilitou a avaliação da qualidade dos sistemas de anotação de papéis semânticos treinados em dados do gênero jornalístico e testados em um gênero distinto, nunca avaliado para a língua portuguesa nessa tarefa. É importante salientar que o *córpus Buscapé* possui muitos erros de *spelling* e de *parsing*.

Foi apresentado o conjunto de regras desenvolvido por [Duran and Aluísio \(2010\)](#) para identificação de verbos auxiliares. Essas regras nos permitiram automatizar a anotação de verbos auxiliares, com papéis semânticos modificadores de verbos plenos. Tal sistema até então nunca havia sido desenvolvido na área de APS.

Foi apresentada a arquitetura híbrida proposta e desenvolvida nesse trabalho para anotar parte dos papéis semânticos com uso de regras e parte com um método de AM. Foi apresentado o módulo de inserção de sujeitos não expressos, desenvolvido para aumentar a frequência de papéis semânticos anotados nos sujeitos dos verbos. Foi apresentado o sistema baseado em regras para anotação de papéis semânticos em verbos auxiliares, bem como a motivação para tal tomada de decisão. Foi apresentado o sistema de APS baseado em AM utilizado nesse trabalho de mestrado. Não foi implementado um novo método de anotação de papéis semânticos com a abordagem de AM, mas foi utilizado o método que apresentou os melhores resultados para o português brasileiro, para avançarmos em outras questões de pesquisa.

Foi mostrado que o módulo de inserção de sujeitos ocultos desenvolvido possui uma boa performance mas que, no entanto, contribui significativamente apenas para a APS do *córpus Buscapé*. O *córpus* de opiniões possui uma linguagem mais pessoal, o que acarreta um maior volume de sujeitos nas primeiras pessoas, enquanto o *córpus* jornalístico possui uma linguagem mais impessoal, possuindo uma frequência menor dos sujeitos nas primeiras pessoas do singular e plural.

Verificamos que a APS de verbos auxiliares, via AM, possui alto valor de precisão e medida F1 e que, portanto, é interessante remover essa classe de papéis semânticos da anotação via AM para não inflar os resultados do sistema de APS. Também foi mostrado que o sistema baseado em regras desenvolvido para anotar os verbos auxiliares

com papéis semânticos modificadores possui bons resultados e que, portanto, é possível utilizá-lo para tal tarefa.

Analizamos o desempenho do sistema de APS de [Alva-Manchego \(2013\)](#), dependente de árvores sintáticas, e o de [Fonseca \(2013\)](#), independente de árvores sintáticas, treinando ambos os sistemas sobre dados ruidosos e também sobre árvores sintáticas revisadas. Em todos os experimentos realizados, o sistema de [Alva-Manchego \(2013\)](#) obteve performance superior, o que significa que os erros sintáticos não afetaram negativamente o sistema de [Alva-Manchego \(2013\)](#) a ponto do sistema de [Fonseca \(2013\)](#) obter melhores resultados. Quanto ao sistema de [Alva-Manchego \(2013\)](#), verificamos que, para anotar automaticamente dados ruidosos, ou seja, árvores sintáticas não revisadas, é necessário treinar o sistema nesse cenário. Dessa forma, o anotador de papéis semânticos captura padrões de erros sintáticos e aprende a lidar com eles. Verificamos também o sistema de [Alva-Manchego \(2013\)](#) obteve resultados razoáveis ao ser treinado sobre árvores sintáticas não revisadas e aplicado sobre árvores sintáticas revisadas.

Por fim, analisamos a performance de APS sobre o gênero de opiniões de produtos na web. Verificamos que, devido às características do gênero, o *pipeline* de pré-processamento não foi capaz de corrigir todos os erros de *spelling* do *corpus*, o que acarretou maiores desafios ao *parser*. Ainda assim, o sistema de [Alva-Manchego \(2013\)](#) obteve resultados superiores ao sistema de [Fonseca \(2013\)](#) em todos os experimentos, o que nos mostra que o sistema de APS consegue aprender a lidar com os erros sintáticos e ortográficos. Por fim, obtivemos como melhor performance de anotação do Buscapé, o sistema de [Alva-Manchego \(2013\)](#) treinado sobre o PLN-Br. Por estarmos lidando com dados gerados por *parser* automático, o sistema treinado sobre o PLN.Br obteve os melhores resultados na anotação desse *corpus*.

6.1 Contribuições

Pontualmente, esse trabalho de mestrado apresenta as seguintes contribuições:

- Seleção de instâncias para APS do *corpus* PLN.Br e anotação delas. Este aluno de mestrado teve participação nas duas tarefas.

- Recolha e compilação do *cópus* Buscapé. Este aluno de mestrado, juntamente com outros bolsistas do projeto PROS@, teve participação na tarefa.
- Seleção de instâncias para APS do *cópus* Buscapé e anotação delas. A anotação foi realizada por este aluno e mais dois anotadores, a seleção foi realizada integralmente por este aluno.
- Desenvolvimento de um sistema de inserção de sujeitos não expressos das primeiras pessoas do singular e plural com precisão boa.
- Desenvolvimento de um sistema de inserção de sujeitos não expressos das terceiras pessoas do singular e plural com baixa precisão e, portanto, com utilização não recomendada.
- Avaliação do ganho da inserção de sujeitos das primeiras pessoas do singular e plural para a tarefa de APS.
- Desenvolvimento de um sistema baseado em regras para anotação de papéis semânticos em verbos auxiliares.
- Avaliação da performance da anotação de papéis semânticos em verbos auxiliares, via AM, e utilização do sistema baseado em regras, desenvolvido, para não inflar as medidas de precisão e F1 dos sistemas de APS baseados em AM.
- Avaliação da anotação de árvores sintáticas não revisadas na tarefa de APS. Neste cenário recomendamos utilizar um sistema de APS sensível a árvores sintáticas, como o de [Alva-Manchego \(2013\)](#), treinando-o com árvores sintáticas não revisadas.
- Avaliação da performance de APS sobre o gênero de opiniões de produtos na web. Os resultados mostraram que o sistema de APS treinado no gênero jornalístico, porém com dados não revisados, obteve a melhor performance no gênero de opiniões de produtos na web.
- Disponibilização de DEMOS para os sistemas desenvolvidos ao longo do trabalho de mestrado: sistema de APS (Todo o *pipeline* apresentado na Seção [4.2](#)), sistema de inserção de sujeitos ocultos e sistema protótipo de normalização lexical.

O endereço de acesso às DEMOS é: <http://nilc.icmc.usp.br/semanticnlp/brazilis>.

6.2 Limitações

As limitações no trabalho de mestrado foram relacionadas com a performance do *parser* PALAVRAS, utilizado no projeto, e com o gênero de opiniões de produtos na web, cujos textos tiveram que ser pré-processados antes do uso do *parser*. Essas limitações são detalhadas a seguir.

A performance do *parser* PALAVRAS impactou diretamente a qualidade do sistema de inserção de sujeitos. Para a inserção dos sujeitos não expressos das primeiras pessoas do singular e plural, não foram identificados muitos erros de *parsing*, tanto é que a performance obtida na tarefa foi muito boa. Na inserção dos sujeitos das terceiras pessoas do singular e plural, no entanto, o *parser* impactou diretamente na qualidade dos resultados, não considerados bons. Errar a pessoa gramatical do verbo ou rotular um sujeito como verbo foram alguns dos erros comumente encontrados. Árvores sintáticas mal construídas também inviabilizaram a inserção dessa classe de sujeitos não expressos.

No sistema anotador de verbos auxiliares, baseado em regras, identificou-se um conjunto de casos não cobertos pelas regras do sistema e que, portanto, devem ser tratados futuramente. Também foram identificados erros do *parser* PALAVRAS que inviabilizaram o correto processamento das regras. Um exemplo desses erros é a confusão entre os verbos “ir” e “ser”.

Quanto ao uso do sentido do verbo alvo como *feature* para AM, concluímos que o seu uso é inviabilizado a menos que o *corpus* de treinamento tenha uma alta frequência de verbos para cada sentido. Isso significa que o *corpus* de treinamento do sistema de APS deve ser bem maior do que o que temos hoje. Além disso, para utilizar as *features* de sentido elaboradas, em um cenário real de produção, é necessário incorporar um módulo de desambiguação lexical de sentidos. Esse módulo é o responsável por rotular o sentido do verbo alvo em uma sentença para então poder utilizá-lo como *feature*. Atualmente, não existe um sistema desambiguizador lexical de sentidos para o

português brasileiro com bom desempenho, muito menos um desenvolvido para lidar com os sentidos do repositório Verbo-Brasil.

Quanto à anotação de papéis semânticos sobre árvores sintáticas não revisadas, percebemos uma grande influência do *parser* PALAVRAS nos resultados obtidos. Uma vez que as árvores não são revisadas, diversos erros de *parsing* são carregados para o sistema de APS que, se não conseguir tratá-los, comete erros de anotação.

Quanto à anotação de papéis semânticos sobre o gênero de opiniões de produtos na web, percebemos a necessidade de mais pesquisa sobre o pré-processamento do gênero, a fim de corrigir erros como os apresentados em [Hartmann et al. \(2014\)](#). Essas características do gênero tornam a tarefa do *parser* mais desafiadora e, como ele já comete muitos erros com textos do gênero jornalístico, comete ainda mais no de opinião.

6.3 Trabalhos Futuros

Como trabalhos futuros, apontamos a seguinte lista de atividades:

- Melhorar as regras de inserção de sujeitos não expressos das primeiras pessoas do singular e plural. Com isso, os erros obtidos dependeriam apenas do *parser*.
- Desenvolver um sistema de inserção de sujeitos não expressos, baseado em AM, para as terceiras pessoas do singular e plural. Como visto que o uso de regras não foi suficiente para tratar essa classe de sujeitos, uma abordagem via AM deve ser mais eficiente.
- Compilar um *corpus*, para APS, com maior frequência de verbos, de forma que haja uma quantidade significativa de cada sentido dos verbos. Assim, as *features* de sentido do verbo devem agregar performance ao sistema de APS.
- Sugerir melhorias aos mantenedores do *parser* PALAVRAS, de modo à obter melhores resultados do *parsing*.
- Desenvolver ou utilizar um sistema com plena normalização do gênero de opiniões da web, de forma a aproximá-lo do gênero jornalístico para o qual se investiu mais esforço para a tarefa de APS.

Referências Bibliográficas

- Alva-Manchego, F. (2013). Anotação automática semissupervisionada de papéis semânticos para o português do Brasil. Master's thesis, University of São Paulo, Brazil.
- Atkinson, K. (2006). Gnu aspell 0.60. 4. <http://aspell.net>. Accessed on 2015-04-27.
- Avanço, L. V., M. S. Duran, and M. G. V. Nunes (2014). Towards a phonetic brazilian portuguese spell checker. In *ToRPorEsp - Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish*, pp. 24–31.
- Baker, C. F., C. J. Fillmore, and J. B. Lowe (1998). The berkeley framenet project. In *Proceedings of the 17th international conference on Computational Linguistics*, pp. 86–90.
- Baker, K., M. Bloodgood, B. J. Dorr, C. Callison-Burch, N. W. Filardo, C. Piatko, L. Levin, and S. Miller (2012). Use of modality and negation in semantically-informed syntactic mt. Volume 38, pp. 411–438. MIT Press.
- Baptista, J., N. Mamede, and F. Gomes (2010). Auxiliary verbs and verbal chains in european portuguese. In *Proceedings of Computational Processing of the Portuguese Language (PROPOR-2010)*, Volume 6001 of *Lecture Notes in Computer Science*, pp. 110–119. Springer Berlin Heidelberg.
- Basili, R., D. De Cao, D. Croce, B. Coppola, and A. Moschitti (2009). Cross-language frame semantics transfer in bilingual corpora. In *Proceedings of the 10th International*

- Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2009)*, pp. 332–345.
- Bick, E. (2000). *The Parsing System “Palavras”: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press Aarhus.
- Bruckschen, M., F. Muniz, J. Souza, J. Fuchs, K. Infante, M. Muniz, P. Gonçalves, R. Vieira, and S. Aluísio (2008). Anotação Lingüística em XML do Corpus PLN-BR. NILC–TR–09–08. Technical report, University of São Paulo, Brazil.
- Burchardt, A., K. Erk, A. Frank, A. Kowalski, and S. Pado (2006). Salto – a versatile multi-level annotation tool. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, pp. 517–520.
- Burchardt, A., K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal (2006). The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*.
- Cançado, M. (2005). Posições argumentais e propriedades semânticas. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada* 21, 23–56.
- Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22(2), 249–254.
- Chambers, N. and D. Jurafsky (2011). Template-based information extraction without the templates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pp. 976–986. Association for Computational Linguistics.
- Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference (NAACL-2000)*, pp. 132–139. Association for Computational Linguistics.
- Clercq, O. D., V. Hoste, and P. Monachesi (2012). Evaluating automatic cross-domain dutch semantic role annotation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, pp. 88–93. European Language Resources Association (ELRA).

- Cohen, W. W. (2004). Minorthird: Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data. <http://minorthird.sourceforge.net>.
- Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. Ph. D. thesis, Univesity of Pennsylvania, USA.
- Collins, M. and T. Koo (2000). Discriminative reranking for natural language parsing. In *Proceedings of the International Conference on Machine Learning (ICML-2000)*, pp. 25–70. MIT Press.
- Collobert, R. and J. Weston (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML-2008)*, pp. 160–167. ACM.
- Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12, 2493–2537.
- Dligach, D. and M. Palmer (2009). Using language modeling to select useful annotation data. In *Proceedings of the Student Research Workshop, NAACL-HLT*, pp. 25–30. Association for Computational Linguistics.
- Duffield, C. J., J. D. Hwang, S. W. Brown, D. Dligach, S. E. Vieweg, J. Davis, and M. Palmer (2007). Criteria for the manual grouping of verb senses. In *Proceedings of the Linguistic Annotation Workshop*, pp. 49–52. Association for Computational Linguistics.
- Duran, M. S. and S. M. Aluísio (2010). Verbos Auxiliares no Português do Brasil. NILC-TR-10-05. Technical report, University of São Paulo, Brazil.
- Duran, M. S. and S. M. Aluísio (2011). Propbank-br: A brazilian portuguese corpus annotated with semantic role labels. In *Proceedings of The 8th Brazilian Symposium in Information and Human Language Technology (STIL-2011)*.
- Duran, M. S. and S. M. Aluísio (2012). Propbank-br: a brazilian treebank annotated with semantic role labels. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC-2012)*, pp. 1862–1867.

- Duran, M. S., M. A. Amâncio, and S. M. Aluísio (2010). Assigning wh-questions to verbal arguments: Annotation tools evaluation and corpus building. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, pp. 1445–1551. European Language Resources Association (ELRA).
- Duran, M. S., L. V. Avanço, and M. G. V. Nunes (2015). UGCNormal: A Normalizer for UGC in Brazilian Portuguese. In *(To appear) Proceedings of ACL 2015, Workshop on Noisy User-generated Text (W-NUT), Beijing, China*.
- Duran, M. S., J. P. Martins, and S. M. Aluísio (2013). Um repositório de verbos para a anotação de papéis semânticos disponível na web. In *Proceedings of the Brazilian Symposium in Information and Human Language Technology (STIL-2013)*, pp. 168–172.
- Duran, M. S., L. Sepúlveda-Torres, M. C. Viviani, N. S. Hartmann, and S. M. Aluísio (2014). Seleção de sentenças do corpus PLN-Br para compor o corpus de anotação de papéis semânticos Propbank-Br.v2. NILC-TR-14-07. Technical report, University of São Paulo, Brazil.
- Fan, R.-E., K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin (2008). Liblinear: A library for large linear classification. *Journal of Machine Learning Research* 9, 1871–1874.
- Feng, M., W. Sun, and H. Ney (2012). Semantic cohesion model for phrase-based SMT. In *Proceedings of the International Conference on Computational Linguistics (COLING-2012)*, pp. 867–878. The COLING 2012 Organizing Committee.
- Ferneda, E. (2003). *Recuperação de Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação*. Ph. D. thesis, University of São Paulo, Brazil.
- Fillmore, C. J. (1967). The case for case. In *Texas Symposium on Language Universals*. ERIC.
- Fleischman, M., N. Kwon, and E. Hovy (2003). Maximum entropy models for frame-net classification. In *Proceedings of the Conference on Empirical Methods in Natural*

- Language Processing (EMNLP-2003)*, pp. 49–56. Association for Computational Linguistics.
- Fonseca, E. R. (2013). Uma abordagem conexionista para anotação de papéis semânticos. Master’s thesis, University of São Paulo, Brazil.
- Fürstenau, H. and M. Lapata (2009). Semi-supervised semantic role labeling. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, pp. 220–228. Association for Computational Linguistics.
- Gao, Q. and S. Vogel (2011). Corpus expansion for statistical machine translation with semantic role label substitution rules. In *ACL:HLT-2011*, pp. 294–298.
- Gildea, D. and J. Hockenmaier (2003). Identifying semantic roles using combinatorial categorial grammar. In *Proceedings of the Conference on Empirical methods in Natural Language Processing (EMNLP-2003)*, pp. 57–64. Association for Computational Linguistics.
- Gildea, D. and D. Jurafsky (2001). Identifying semantic roles in text. In *17th International Joint Conference on Artificial Intelligence (IJCAI-01)*.
- Gildea, D. and D. Jurafsky (2002). Automatic labeling of semantic roles. *Computational Linguistics* 28(3), 245–288.
- Gildea, D. and M. Palmer (2002). The necessity of syntactic parsing for predicate argument recognition. In *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (Coling-2002)*, pp. 239–246.
- Giuglea, A.-M. and A. Moschitti (2006). Semantic role labeling via framenet, verbnet and propbank. In *Proceedings of the 21st International Conference on Computational Linguistics (ACL-2006)*, pp. 929–936. Association for Computational Linguistics.
- Hajič, J., M. Ciaramita, R. Johansson, D. Kawahara, M. A. Martí, L. Màrquez, A. Meyers, J. Nivre, S. Padó, J. Štěpánek, et al. (2009). The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the 13th Conference on Computational Natural Language Learning: Shared Task (CoNLL-2009)*, pp. 1–18. Association for Computational Linguistics.

- Hart, L., H. Alam, and A. Kumar (2012). Revisiting Arabic semantic role labeling using SVM kernel methods. In *Proceedings of the International Conference on Computational Linguistics (COLING-2012)*, pp. 207–214. The COLING 2012 Organizing Committee.
- Hartmann, N. S., L. Avanço, P. Balage, M. Duran, M. D. G. V. Nunes, T. Pardo, and S. Aluísio (2014). A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, pp. 3865–3871. European Language Resources Association (ELRA).
- Hartmann, N. S., M. Duran, and S. Aluísio (2014). Filling the gap: inserting an artificial constituent where a subject is omitted in portuguese. In *Proceedings of the Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish (ToRPorEsp-2014)*. Committee of the International Conference on Computational Processing of Portuguese (PROPOR-2014).
- Jannach, D., M. Zanker, A. Felfernig, and G. Friedrich (2010). Other text classification methods. In *Recommender systems: an introduction*, pp. 63–64. Cambridge University Press.
- Johansson, R. (2008). *Dependency-based Semantic Analysis of Natural-language*. Ph. D. thesis, University of Lund, Sweden.
- Johansson, R., K. F. Heppin, and D. Kokkinakis (2012). Semantic role labeling with the swedish framenet. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, pp. 3697–3700. European Language Resources Association (ELRA).
- Johansson, R. and P. Nugues (2008). Dependency-based semantic role labeling of prop-bank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, pp. 69–78. Association for Computational Linguistics.
- Jones, B., J. Andreas, D. Bauer, K. M. Hermann, and K. Knight (2012). Semantics-based machine translation with hyperedge replacement grammars. In *Proceedings of*

- the International Conference on Computational Linguistics (COLING 2012)*, pp. 1359–1376. The COLING 2012 Organizing Committee.
- Kaisser, M. and B. Webber (2007). Question answering based on semantic roles. In *Proceedings of the ACL Workshop on Deep Linguistic Processing*, pp. 41–48. Association for Computational Linguistics.
- Kipper, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph. D. thesis, Univerdity of Pennsylvania, USA.
- Knight, K. and D. Marcu (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence* 139(1), 91–107.
- Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit (MT Summit-2005)*, pp. 79–86. European Association for Machine Translation (EAMT).
- Kong, F. and G. Zhou (2012). Exploring local and global semantic information for event pronoun resolution. In *Proceedings of the International Conference on Computational Linguistics (COLING-2012)*, pp. 1475–1488. The COLING 2012 Organizing Committee.
- Kozhevnikov, M. and I. Titov (2013). Bootstrapping semantic role labelers from parallel data. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM-2013)*, pp. 317–327. Association for Computational Linguistics.
- Landis, J. R. and G. G. Koch (1977). The measurement of observer agreement for categorical data. *Biometrics* 33(1), pp. 159–174.
- Lang, J. and M. Lapata (2011). Unsupervised semantic role induction via split-merge clustering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, pp. 1117–1126. Association for Computational Linguistics.
- Liu, D. and D. Gildea (2010). Semantic role features for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010)*, pp. 716–724. Association for Computational Linguistics.

- Lluís, X., X. Carreras, and L. Màrquez (2013). Joint arc-factored parsing of syntactic and semantic dependencies. *Transactions of the Association for Computational Linguistics* 1, 219–230.
- Lo, C. and D. Wu (2011). Meant: an inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, Volume 1, pp. 220–229.
- López, R. and T. A. Pardo (2015). Experiments on sentence boundary detection in user-generated web content. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2015)*, pp. 227–237. Springer International Publishing.
- Maarouf, I. E. and J. Villaneau (2012). A french fairy tale corpus syntactically and semantically annotated. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, pp. 345–350. European Language Resources Association (ELRA).
- Marcus, M. P., M. A. Marcinkiewicz, and B. Santorini (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics* 19(2), 313–330.
- Mitchell, T. M. (1997). *Machine Learning* (1 ed.). McGraw-Hill, Inc.
- Monard, M. C. and J. A. Baranauskas (2003). Conceitos sobre aprendizado de máquina. In *Sistemas Inteligentes - Fundamentos e Aplicações*, pp. 89–114. Editora Manole.
- Palmer, M., H. T. Dang, and C. Fellbaum (2007). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering* 13(2), 137–163.
- Palmer, M., D. Gildea, and P. Kingsbury (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1), 71–106.
- Palmer, M., D. Gildea, and N. Xue (2010). *Semantic role Labeling*, Volume 3 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

- Pradhan, S., K. Hacıoglu, V. Krugler, W. Ward, J. H. Martin, and D. Jurafsky (2005). Support vector learning for semantic argument classification. *Machine Learning* 60(1-3), 11–39.
- Pradhan, S. S., W. Ward, and J. H. Martin (2008). Towards robust semantic role labeling. *Computational Linguistics* 34(2), 289–310.
- Ricci, F., L. Rokach, and B. Shapira (2011). Data mining methods for recommender systems. In *Recommender Systems Handbook*, pp. 48. Springer.
- Roland, D. (2002). *Verb Sense and Verb Subcategorization Probabilities*. Ph. D. thesis, University of Colorado, USA.
- Roth, D. and W.-t. Yih (2005). Integer linear programming inference for conditional random fields. In *Proceedings of the 22nd International Conference on Machine learning (ICML-2005)*, pp. 736–743. Association for Computing Machinery.
- Salza, E. (2013). Using NooJ for semantic annotation of Italian language corpora in the domain of motion: a cognitive-grounded approach. In *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing (FSMNLP-2013)*, pp. 72–80. Association for Computational Linguistics.
- Scarton, C. E. (2013). Verbnet.br: construção semiautomática de um léxico verbal online e independente de domínio para o português do brasil. Master’s thesis, University of São Paulo, Brazil.
- Sequeira, J., T. Gonçalves, and P. Quaresma (2012). Semantic role labeling for portuguese—a preliminary approach—. In *Proceedings of Computational Processing of the Portuguese Language (PROPOR-2012)*, pp. 193–203. Springer.
- Shamsfard, M. and M. S. Mousavi (2007). Thematic role extraction using shallow parsing. *International Journal of Computational, Intelligence*, 126–132.
- Shen, D. and M. Lapata (2007). Using semantic roles to improve question answering. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP-2007)*, pp. 12–21.

- Søgaard, A., A. Johannsen, B. Plank, D. Hovy, and H. Martinez (2014). What’s in a p-value in nlp? In *Proceedings of the 18th Conference on Computational Natural Language Learning (CONLL2014)*, pp. 1–10.
- Stevens, G., P. Monachesi, and A. van den Bosch (2007). A pilot study for automatic semantic role labeling in a dutch corpus. In *Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands (CLIN-2007)*, Volume 7, pp. 99–114.
- Surdeanu, M., S. Harabagiu, J. Williams, and P. Aarseth (2003). Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL-2003)*, pp. 8–15. Association for Computational Linguistics.
- Titov, I. and A. Klementiev (2012a). A bayesian approach to unsupervised semantic role induction. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2012)*, pp. 12–22. Association for Computational Linguistics.
- Titov, I. and A. Klementiev (2012b). Crosslingual induction of semantic roles. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL-2012)*, pp. 647–656. Association for Computational Linguistics.
- Titov, I. and A. Klementiev (2012c). Semi-supervised semantic role labeling: Approaching from an unsupervised perspective. In *Proceedings of the International Conference on Computational Linguistics (Coling-2012)*, pp. 2635–2652. The COLING 2012 Organizing Committee.
- Toutanova, K., A. Haghighi, and C. D. Manning (2008). A global joint model for semantic role labeling. *Computational Linguistics* 34(2), 161–191.
- Van der Plas, L., P. Merlo, and J. Henderson (2011). Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL-2011)*, pp. 299–304.
- Wu, D. and P. Fung (2009). Semantic roles for smt: A hybrid two-pass model. In *Proceedings of the Annual Conference of the North American Chapter of the Association for*

Computational Linguistics (NAACL-2009), pp. 13–16. Association for Computational Linguistics.

Xu, F., Q. Zhu, and G. Zhou (2012). A unified framework for discourse argument identification via shallow semantic parsing. In *Proceedings of the International Conference on Computational Linguistics (COLING 2012)*, pp. 1331–1340. The COLING 2012 Organizing Committee.

Yi, S.-T., E. Loper, and M. Palmer (2007). Can semantic roles generalize across genres? In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-2007)*, pp. 548–555. Association for Computational.

Yoshikawa, K., R. Iida, T. Hirao, and M. Okumura (2012). Sentence compression with semantic role constraints. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL-2012)*, pp. 349–353. Association for Computational Linguistics.

Zavaglia, C. (2003). Base de conhecimento léxico-ontológico para o português do brasil: uma proposta de modelo. In *Anais do 1º Workshop em Tecnologia da Informação e Linguagem Humana (TIL-2003)*, evento associado ao XVI Brazilian Symposium on Computer Graphics and Image Processing - (SIBGRAPI).

Zavaglia, C. (2006). O papel do léxico na elaboração de ontologias computacionais: de seu resgate à sua disponibilização. *Léxico e morfologia: perspectivas e análises.*, 229–269.

Quantidade de Instâncias Seleccionadas por Verbo do Corpus Buscapé

Na Tabela [41](#) deste apêndice são listados todos os verbos do corpus Buscapé seleccionados para anotação, bem como a quantidade de instâncias de cada um. A descrição do procedimento de anotação é apresentada em *Seleção do Buscapé*, na seção [4.1.1](#).

Verbo	Instâncias	Verbo	Instâncias	Verbo	Instâncias	Verbo	Instâncias
reclamar	11	ver	11	desligar	11	trocar	11
adquirir	11	precisar	11	ligar	11	conhecer	11
faltar	11	quebrar	11	saber	11	ler	11
tornar	11	colocar	11	oferecer	11	durar	11
lavar	11	procurar	11	pensar	11	adorar	11
ir	11	chegar	11	possuir	11	falar	11
gostar	11	demorar	11	ficar	11	vir	11
utilizar	11	receber	11	rodar	11	comprar	11
instalar	11	perder	11	resolver	11	encontrar	11
levar	11	fazer	11	jogar	11	pesquisar	11
assistir	11	estar	11	dar	11	querer	11
entrar	11	sair	11	amar	11	deixar	11
acabar	11	desejar	11	atender	11	valer	11
permitir	11	travar	11	usar	11	acreditar	11
tirar	11	comparar	11	arrepender	11	esperar	11
dizer	11	existir	11	haver	11	facilitar	11
ter	11	achar	11	parecer	11	limpar	11
carregar	11	poder	11	passar	11	caber	10
ser	11	cumprir	11	funcionar	11	conseguir	10
apresentar	11	pagar	11	decidir	11	dever	6

Tabela 41: Distribuição das instâncias selecionadas por verbo do corpus Buscapé.

Tabela com as Etiquetas do Guia de Anotação do PropBank.Br

Na Tabela 42 deste apêndice são listados todas as etiquetas de papéis semânticos utilizadas na anotação do PLN-Br e também do Buscapé. Exceto alguns papéis, todos os demais são também utilizados no PropBank inglês e, portanto, possuem o seu significado em inglês apresentado na tabela. Os argumentos Arg2, Arg3, Arg4 e Arg5 possuem uso genérico e, portanto, não possuem um significado atribuído na Tabela 42.

Numerados	PropBank	PropBank.Br
Arg0	<i>Prototypical Agent</i>	Agente Prototípico
Arg1	<i>Prototypical Patient</i>	Paciente Prototípico
Arg2		
Arg3		
Arg4		
Arg5		
Modificadores	PropBank	PropBank.Br
ArgM-Cau	<i>Cause</i>	Causa
ArgM-Loc	<i>Locative</i>	Local
ArgM-Neg	<i>Negation</i>	Negação
ArgM-Mnr	<i>Manner</i>	Maneira
ArgM-Prp	<i>Purpose</i>	Propósito
ArgM-Ext	<i>Extent</i>	Quantidade
ArgM-Com	<i>Commitative</i>	Companhia
ArgM-Gol	<i>Goal</i>	Objetivo
ArgM-Dir	<i>Directional</i>	Direção
Auxiliares	PropBank	PropBank.Br
ArgM-Adv	<i>Adverbial</i>	Advérbios que modificam toda a oração
ArgM-Prd	<i>Secondary Predication</i>	Predicação Secundária
ArgM-Rec	<i>Reciprocals</i>	Reciprocidade
ArgM-Dis	<i>Discourse</i>	Marcador Discursivo
ArgM-Mod	<i>Modals</i>	Verbo Auxiliar de Modo
ArgM-Asp		Verbo Auxiliar de Aspecto
ArgM-Pas		Verbo Auxiliar de Voz Passiva
ArgM-Tml		Verbo Auxiliar de Tempo
ArgM-Exp		Expletivos
ArgM-Nse		Pronome Reflexivo não Argumental

Tabela 42: Conjunto de papéis semânticos utilizado nesse trabalho de mestrado.