UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

MIRIAM LÚCIA CAMPOS SERRA DOMINGUES

ABORDAGEM PARA O DESENVOLVIMENTO DE UM ETIQUETADOR DE ALTA ACURÁCIA PARA O PORTUGUÊS DO BRASIL

TD 10 / 2011

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém - Pará - Brasil
2011

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

MIRIAM LÚCIA CAMPOS SERRA DOMINGUES

ABORDAGEM PARA O DESENVOLVIMENTO DE UM ETIQUETADOR DE ALTA ACURÁCIA PARA O PORTUGUÊS DO BRASIL

TD 10 / 2011

Tese submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Pará como requisito parcial para obtenção do Grau de Doutor em Engenharia Elétrica.

Orientador: Prof. Dr. Eloi Luiz Favero

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém - Pará - Brasil
2011

Dados Internacionais de Catalogação na Publicação (CIP)

D671a Domingues, Miriam Lúcia Campos Serra

Abordagem para o desenvolvimento de um etiquetador de alta acurácia para o Português do Brasil / Miriam Lúcia Campos Serra Domingues; orientador, Eloi Luiz Favero. – 2011.

Tese (Doutorado) – Universidade Federal do Pará, Instituto de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, Belém, 2011.

1. Etiquetagem morfossintática. 2. Processamento de linguagem natural. 3. Linguística computacional. 4. Linguística de *corpus*. I. Título.

CDD: 22. ed. 410.285

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

MIRIAM LÚCIA CAMPOS SERRA DOMINGUES

ABORDAGEM PARA O DESENVOLVIMENTO DE UM ETIQUETADOR DE ALTA ACURÁCIA PARA O PORTUGUÊS DO BRASIL

TESE SUBMETIDA À AVALIAÇÃO DA BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA DA UNIVERSIDADE FEDERAL DO PARÁ E JULGADA ADEQUADA PARA OBTENÇÃO DO GRAU DE DOUTOR EM ENGENHARIA ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA.

APROVADA EM: 21 DE OUTUBRO DE 2011.

BANCA EXAMINADORA

Prof. Dr. Eloi Luiz Favero
(Orientador – UFPA)

Prof. Dr. Aldebaro Barreto da Rocha Klautau Júnior
(Membro – UFPA)

Prof. Dr. Roberto Célio Limão de Oliveira
(Membro – UFPA)

Profa. Dra. Regina Célia Fernandes Cruz
(Membro – UFPA)

Prof. Dr. Gustavo Augusto Lima de Campos
(Membro – UECE)

Visto:

Prof. Dr. Marcus Vinícius Alves Nunes
Coordenador do PPGEE/ITEC/UFPA

UFPA / ITEC / PPGEE

Dedico este trabalho à minha mãe Antonia, ao meu pai Miguel (*in memoriam*), ao meu esposo José Luiz e aos meus filhos Rafael e Rodrigo.

Agradecimentos

Agradeço a DEUS, por me conduzir e me fortalecer na realização desta jornada.

Ao Prof. Dr. Eloi Luiz Favero, pela orientação, auxílio e amizade, pelos ensinamentos repassados e pelos caminhos apontados para o sucesso deste trabalho.

Ao meu esposo, José Luiz, e aos meus filhos, Rafael e Rodrigo, pelo amor, carinho, incentivo, compreensão e por compartilharem comigo anseios e alegrias no decorrer das atividades do Curso.

À minha mãe, Antonia, pelo seu imenso amor e pela ajuda em todos os momentos.

Ao meu pai Miguel (*in memoriam*), meus irmãos, tios, primos, sobrinhos e cunhados que me incentivaram durante a realização do curso.

Ao Ivo Medeiros, ex-bolsista de iniciação científica, pela participação dedicada nas atividades desta pesquisa e pela amizade.

A todos os professores e amigos do PPGEE, pela ajuda e troca de conhecimentos durante as disciplinas e atividades do Curso.

Aos professores Aldebaro Klautau Júnior, Roberto Limão de Oliveira, Regina Cruz, Gustavo Lima de Campos e Renata Vieira, membros da banca de qualificação, pelas valiosas revisões, críticas e avaliações.

Aos amigos do Instituto de Ciências da Saúde e da Universidade Federal do Pará, pelo apoio e incentivo.

À amiga Vilma Bastos, bibliotecária, pela ajuda na elaboração das referências bibliográficas.

Aos amigos Thaís Tavares, Ana Carla Silva, Adriano Lino, Odila Ventura e Socorro Palheta, pelo apoio e pelas palavras de incentivo em todas as horas, e ao Prof. Nandamudi Vijaykumar, pelas sugestões na escrita de artigo.

Ao Instituto de Ciências da Saúde pela liberação e apoio para que eu pudesse realizar o Curso.

À Universidade Federal do Pará, pela oportunidade de capacitação profissional oferecida.

Ao Programa de Pós-Graduação em Engenharia Elétrica, pela realização do Curso e pelo apoio.

A todos que de alguma forma contribuíram para a realização desta tese.

“Progresso, da melhor espécie, é comparativamente lento. Grandes resultados não podem ser alcançados imediatamente; e devemos estar satisfeitos em avançar na vida como andamos, passo a passo.”

(Samuel Smiles)

Resumo

A etiquetagem morfossintática é uma tarefa básica requerida por muitas aplicações de processamento de linguagem natural, tais como análise gramatical e tradução automática, e por aplicações de processamento de fala, por exemplo, síntese de fala. Essa tarefa consiste em etiquetar palavras em uma sentença com as suas categorias gramaticais. Apesar dessas aplicações requererem etiquetadores que demandem maior precisão, os etiquetadores do estado da arte ainda alcançam acurácia de 96 a 97%. Nesta tese, são investigados recursos de *corpus* e de *software* para o desenvolvimento de um etiquetador com acurácia superior à do estado da arte para o português brasileiro. Centrada em uma solução híbrida que combina etiquetagem probabilística com etiquetagem baseada em regras, a proposta de tese se concentra em um estudo exploratório sobre o método de etiquetagem, o tamanho, a qualidade, o conjunto de etiquetas e o gênero dos *corpora* de treinamento e teste, além de avaliar a desambiguização de palavras novas ou desconhecidas presentes nos textos a serem etiquetados. Quatro *corpora* foram usados nos experimentos: CETENFolha, Bosque CF 7.4, Mac-Morpho e Selva Científica. O modelo de etiquetagem proposto partiu do uso do método de aprendizado baseado em transformação (TBL) ao qual foram adicionadas três estratégias, combinadas em uma arquitetura que integra as saídas (textos etiquetados) de duas ferramentas de uso livre, o TreeTagger e o μ-TBL, com os módulos adicionados ao modelo. No modelo de etiquetador treinado com o *corpus* Mac-Morpho, de gênero jornalístico, foram obtidas taxas de acurácia de 98,05% na etiquetagem de textos do Mac-Morpho e 98,27% em textos do Bosque CF 7.4, ambos de gênero jornalístico. Avaliou-se também o desempenho do modelo de etiquetador híbrido proposto na etiquetagem de textos do *corpus* Selva Científica, de gênero científico. Foram identificadas necessidades de ajustes no etiquetador e nos *corpora* e, como resultado, foram alcançadas taxas de acurácia de 98,07% no Selva Científica, 98,06% no conjunto de teste do Mac-Morpho e 98,30% em textos do Bosque CF 7.4. Esses resultados são significativos, pois as taxas de acurácia alcançadas são superiores às do estado da arte, validando o modelo proposto em busca de um etiquetador morfossintático mais confiável.

Palavras-chave. Etiquetagem morfossintática, processamento de linguagem natural, linguística computacional, linguística de *corpus*.

Abstract

Part-of-speech tagging is a basic task required by many applications of natural language processing, such as parsing and machine translation, and by applications of speech processing, for example, speech synthesis. This task consists of tagging words in a sentence with their grammatical categories. Although these applications require taggers with greater precision, the state of the art taggers still achieved accuracy of 96 to 97%. In this thesis, *corpus* and software resources are investigated for the development of a tagger with accuracy above of that of the state of the art for the Brazilian Portuguese language. Based on a hybrid solution that combines probabilistic tagging with rule-based tagging, the proposed thesis focuses on an exploratory study on the tagging method, size, quality, tag set, and the textual genre of the *corpora* available for training and testing, and evaluates the disambiguation of new or out-of-vocabulary words found in texts to be tagged. Four *corpora* were used in experiments: CETENFolha, Bosque CF 7.4, Mac-Morpho, and Selva Científica. The proposed tagging model was based on the use of the method of transformation-based learning (TBL) to which were added three strategies combined in a architecture that integrates the outputs (tagged texts) of two free tools, Treetagger and μ-TBL, with the modules that were added to the model. In the tagger model trained with Mac-Morpho *corpus* of journalistic genre, tagging accuracy rates of 98.05% on Mac-Morpho test set and 98.27% on Bosque CF 7.4 were achieved, both of journalistic genres. The performance of the proposed hybrid model tagger was also evaluated in the texts of Selva Científica *Corpus*, of the scientific genre. Needs of adjustments in the tagger and in *corpora* were identified and, as result, accuracy rates of 98.07% in Selva Científica, 98.06% in the text set of Mac-Morpho, and 98.30% in the texts of the Bosque CF 7.4 have been achieved. These results are significant because the accuracy rates achieved are higher than those of the state of the art, thus validating the proposed model to obtain a more reliable part-of-speech tagger.

Keywords. Part-of-speech tagging, natural language processing, computational linguistics, *corpus* linguistics.

Lista de Figuras

FIGURA 2.1 - Árvore sintática do analisador gramatical PALAVRAS 31
FIGURA 2.2 - Sequências possíveis de etiquetas para uma sentença de três palavras 38
FIGURA 2.3 - Exemplo de árvore de decisão 41
FIGURA 2.4 - Aprendizado Baseado em Transformação Dirigida por Erro 45
FIGURA 4.1 - Textos do *Corpus* CETENFolha em formato de ADs 62
FIGURA 4.2 - Textos do *Corpus* Mac-Morpho 63
FIGURA 4.3 - Treinamento e etiquetagem probabilística com o TreeTagger 67
FIGURA 4.4 - Treinamento e etiquetagem baseada em regras com o μ-TBL 68
FIGURA 4.5 - Treinamento do etiquetador proposto 69
FIGURA 4.6 - Arquitetura do etiquetador híbrido proposto 71
FIGURA 4.7 - Exemplo de um *corpus* com informações linguísticas em formato XML 72
FIGURA 4.8 - Formato do arquivo de treinamento do μ-TBL 76
FIGURA 4.9 - Lista ordenada de regras aprendidas com o μ-TBL 77
FIGURA 4.10 - Arquivo com os erros após a aplicação das regras aprendidas com o μ-TBL 77
FIGURA 5.1 - Regra para correção de erro de etiquetagem, programada em Java 80
FIGURA 5.2 - Exemplos dos conjuntos de etiquetas simples e modificadas 81
FIGURA 5.3 - Acurácia média da etiquetage probabilística com o QTag e TreeTagger 83
FIGURA 5.4 - Acurácia média com etiquetas simples e etiquetas modificadas 84
FIGURA 5.5 - Acurácia média da etiquetagem híbrida em *corpus* com etiquetas simples e a aplicação de 774 regras manuais 84
FIGURA 5.6 - Acurácia média da etiquetagem híbrida em *corpus* com etiquetas simples e a aplicação de 2.721 regras extraídas com o μ-TBL 85
FIGURA 5.7 - Acurácia média da etiquetagem híbrida em *corpus* com etiquetas modificadas e a aplicação de 281 regras manuais 86
FIGURA 5.8 - Acurácia média da etiquetagem híbrida em *corpus* com etiquetas modificadas e a aplicação de 1.598 regras extraídas com o μ-TBL 86
FIGURA 5.9 - Tela de configuração do etiquetador e analisador gramatical PALAVRAS. 88
FIGURA 5.10 - Acurácia média com o etiquetador do PALAVRAS 89
FIGURA 6.1 - Divisão do *corpus* em conjuntos de treinamento (TR) e teste (TE) 97
FIGURA 6.2 - Falsos erros visualizados em um arquivo gerado pelo μ-TBL 101
FIGURA 6.3 - Casos em que as palavras do termo “a respeito de” tem diferentes etiquetas 101

FIGURA 6.4 - Acurácia média de etiquetagem após a aplicação das estratégias de consulta ao léxico de nomes próprios e aplicação de regras adicionais sobre os conjuntos de testes TE_{VC} e TE do Mac-Morpho (MM) e conjunto de teste formado pelo Bosque CF 7.4 (CF 7.4)..106

FIGURA 6.5 - Acurácia média de etiquetagem após a aplicação das estratégias de consulta ao léxico de nomes próprios, aplicação de regras adicionais sobre os conjuntos de teste e uso de um léxico de grande tamanho sobre os conjuntos de testes TE_{VC} e TE do Mac-Morpho (MM) e conjunto de teste formado pelo Bosque CF 7.4 (CF 7.4) 107

FIGURA 6.6 - Número de novas palavras no léxico a cada acréscimo de 500 mil palavras.109

FIGURA 7.1 - Regra para corrigir a etiqueta para o pronome pessoal “se” 116

Lista de Tabelas

TABELA 1.1 - Número de erros de etiquetagem por percentual de acurácia em um texto de 1.130 palavras do português brasileiro 18

TABELA 2.1 - Classificação de *corpus* quanto ao tamanho 27

TABELA 2.2 - *Templates* não-lexicalizados 46

TABELA 2.3 - *Templates* lexicalizados 46

TABELA 2.4 - *Templates* para palavras desconhecidas 46

TABELA 4.1 - Conjunto de etiquetas do Bosque CF 7.4 63

TABELA 4.2 - Conjunto de etiquetas do *Corpus* Mac-Morpho 64

TABELA 4.3 - Tarefas de pré-processamento das palavras que serão etiquetadas nos arquivos de teste 74

TABELA 5.1 - Características do conjunto de etiquetas modificadas 81

TABELA 5.2 - Resultados da etiquetagem com o QTag e TreeTagger 87

TABELA 6.1 - Regras para correção de erros em uma etapa intermediária de etiquetagem ... 95

TABELA 6.2 - Regras para pós-correção de erros 96

TABELA 6.3 - Desempenho da etiquetagem híbrida no *Corpus* Mac-Morpho 99

TABELA 6.4 - Categorias mais frequentes em erros de etiquetagem no *Corpus* Mac-Morpho 100

TABELA 6.5 - Desempenho da etiquetagem híbrida sobre o conjunto de teste TE do Mac-Morpho após as estratégias propostas terem sido incluídas 103

TABELA 6.6 - Desempenho da etiquetagem híbrida no Bosque CETENFolha 7.4 após as estratégias propostas terem sido incluídas 105

TABELA 7.1 - Desempenho do etiquetador em textos do *Corpus* Selva Científica 115

TABELA 7.2 - Desempenho em textos do *Corpus* Selva Científica após os ajustes no etiquetador 118

TABELA 7.3 - Erros mais frequentes por categorias gramaticais 119

TABELA 7.4 - Erros mais frequentes por palavras 119

TABELA 7.5 - Desempenho em textos do *Corpus* Mac-Morpho e Bosque CF 7.4 após os ajustes no etiquetador 120

Lista de Abreviaturas

ADs	Árvores Deitadas
BD	Banco de Dados
CETENFolha	*Corpus de Extractos de Textos Electrónicos*/Folha de São Paulo
CF	CETENFolha
CG	*Constraint Grammar*
CRF	*Conditional Random Field*
EMs	Entidades Mencionadas
HMM	*Hidden Markov Models*
MBL	*Memory-Based Learning*
MLE	*Maximum Likelihood Estimation*
MSD	*Morpho-Syntactic Descriptions*
OOV	*Out-of-Vocabulary*
POS	*Part-of-Speech*
PLN	Processamento de Linguagem Natural
SVM	*Support Vector Machine*
TA	Tradução Automática
TBL	*Transformation-Based Learning*
TE_{VC}	Conjunto de teste da avaliação com validação cruzada
VISL	*Visual Interactive Syntax Learning*
WSD	*Word Sense Disambiguation*
WSJ	*Wall Street Journal*

Sumário

1 Introdução 15
1.1 Contextualização 15
1.2 Motivação e Relevância 17
1.3 Objetivos 18
1.4 Premissas 19
1.5 Metodologia 20
1.6 Trabalhos Publicados 21
1.7 Contribuições 22
1.8 Organização do Texto 22
2 Etiquetagem Morfossintática 25
2.1 Níveis de Análise da Linguagem 25
2.2 Linguística Computacional 26
2.2.1 Linguística de *Corpus* 27
2.2.2 Processamento de Linguagem Natural (PLN) 27
2.3 Etiquetagem Automática 28
2.3.1 Etiquetagem Morfossintática 29
2.4 Abordagens para Etiquetagem Morfossintática 32
2.4.1 Abordagem Baseada em Regras 32
2.4.2 Abordagem Probabilística 34
2.4.3 Abordagem Híbrida 43
2.5 Processo de Classificação 47
2.5.1 Métodos de Avaliação da Acurácia do Classificador 48
2.6 Adaptação de Domínio 49
3 Revisão Bibliográfica 51
3.1 Etiquetadores X Abordagens X Acurácia 51
3.1.1 Abordagem Baseada em Regras 51
3.1.2 Abordagem Probabilística 52
3.1.3 Abordagem Híbrida 53
3.2 Etiquetadores X Conjunto de Etiquetas X Acurácia 54
3.3 Outras Abordagens 55
3.3.1 Redes Neurais 55
3.3.2 *Latent Analogy* 55
3.3.3 Combinação de Etiquetadores 56
3.4 Considerações 56
4 Descrição de Abordagem de um Etiquetador de Alta Acurácia para o Português Brasileiro 57
4.1 Recursos Experimentais 57
4.1.1 Etiquetadores 58
4.1.2 Corpora 61
4.2 Metodologia 64
4.3 Arquitetura do Etiquetador 65
4.3.1 Etapa Probabilística 66
4.3.2 Etapa Baseada em Regras Aprendidas Automaticamente 67
4.3.3 Treinamento no Modelo Final de Etiquetador Proposto 68
4.3.4 Arquitetura do Etiquetador Proposto 70
4.4 Processo de Etiquetagem 71
4.4.1 Pré-Processamento 72
4.4.2 Etiquetagem Probabilística 74
4.4.3 Consulta ao Léxico de Nomes Próprios 75
4.4.4 Etiquetagem Baseada em Regras 75

4.4.5 Comparação dos Resultados da Etiquetagem Híbrida com os Textos Pré-Anotados 78
4.4.6 Métrica da Acurácia de Etiquetagem e a Obtenção das Estatísticas do Processo 78

5 Etiquetagem Baseada em Textos do Bosque CETENFolha 79

5.1 Estratégias Avaliadas 79
5.1.1 Aplicação de Regras para Correção de Erros da Etiquetagem Probabilística 79
5.1.2 Modificação no Conjunto Inicial de Etiquetas 80
5.2 Treinamento e Teste de Etiquetadores com o Bosque CF 7.4 81
5.2.1 Método de Avaliação 82
5.2.2 Parâmetros de Configuração das Ferramentas 82
5.2.3 Etiquetagem Probabilística 83
5.2.4 Etiquetagem Híbrida 84
5.2.5 Etiquetagem Baseada em Regras no Ambiente VISL 87
5.2.6 Avaliação dos Resultados da Etiquetagem Baseada no Bosque CF 7.4 89

6 Etiquetagem Baseda em Textos do Corpus Mac-Morpho 93

6.1 Estratégias Avaliadas 93
6.1.1 Consulta a um Léxico de Nomes Próprios 93
6.1.2 Adição de Regras ao Método TBL 94
6.1.3 Uso de um Léxico de Grande Tamanho para o TreeTagger 96
6.2 Treinamento do Etiquetador com o Corpus Mac-Morpho 97
6.2.1 Etiquetagem em Conjuntos de Testes do *Corpus* Mac-Morpho 97
6.2.2 Etiquetagem do Bosque CF 7.4 103
6.2.3 Avaliação dos Resultados da Etiquetagem Baseada no *Corpus* Mac-Morpho 107

7 Etiquetagem em Textos de Gêneros Diferentes 113

7.1 Metodologia dos Experimentos 113
7.2 Identificação e Remoção de Ruídos do *Corpus* Selva Científica 113
7.3 Etiquetagem do Corpus Selva Científica 114
7.3.1 Método de Avaliação 114
7.3.2 Experimentos 1 e 2 114
7.4 Avaliação dos Erros de Etiquetagem e Ajustes no Etiquetador 115
7.5 Etiquetagem do Selva Científica depois dos Ajustes 117
7.5.1 Experimentos 3 e 4 117
7.6 Etiquetagem do Mac-Morpho e Bosque CF 7.4 depois dos Ajustes 119
7.6.1 Experimentos 5 e 6 119
7.7 Avaliação dos Resultados da Etiquetagem em Textos de Gêneros Diferentes 120

8 Considerações Finais 121

Referências 123

Apêndice A – Treinamento e Etiquetagem com o QTag 133

Apêndice B – Treinamento e Etiquetagem com o TreeTagger 134

Apêndice C – Informações Estatísticas sobre o Treinamento e Teste com o μ-TBL 135

Apêndice D – Conjunto de Etiquetas Modificadas para o Bosque CF 7.4 136

Apêndice E – Arquivos de Classes Abertas para o TreeTagger 137

Apêndice F – Etiquetagem Sintática no Bosque CF 7.4 e no Ambiente VISL 138

Apêndice G – Comandos para Treinamento e Teste do TreeTagger 139

Apêndice H – Comandos para Extrair Regras com o μ-TBL 140

1 Introdução

"To many, the ability of computers to process language as skillfully as we do will signal the arrival of truly intelligent machines. The basis of this belief is the fact that effective use of language is intertwined with our general cognitive abilities. […] "
(JURAFSKY; MARTIN, 2000, p. 6)

Nos anos recentes, aplicações de processamento de linguagem natural (PLN), tais como análise gramatical ou *parsing*, tradução automática (TA) e simplificação de textos, bem como aplicações de processamento de fala, por exemplo, síntese de fala, têm sido cada vez mais requisitadas em virtude da grande demanda para processar, automaticamente, informação textual e falada (BELLEGARDA, 2010; GASPERIN, MAZIERO, ALUÍSIO, 2010; SILVA et al., 2010; POPOVIĆ, NEY, 2011). Nesta tese, é investigada uma tarefa básica e importante em aplicações de PLN: a etiquetagem morfossintática (*part-of-speech tagging – POS tagging*), que se situa nos níveis de análise da linguagem da morfologia e da sintaxe.

1.1 Contextualização

A investigação da linguagem e das línguas naturais com a utilização de recursos computacionais é tema da linguística computacional, área de pesquisa que, segundo Othero e Menuzzi (2005), pode ser subdividida em duas subáreas: 1) linguística de *corpus*, que se aplica ao estudo da língua pela observação de *corpora* eletrônicos, grandes bases de dados que contêm amostras de linguagem natural, e 2) PLN, que se ocupa do estudo da linguagem voltado para a construção de ferramentas computacionais específicas para o processamento automático de informação textual e falada.

A etiquetagem morfossintática é uma tarefa de pré-processamento importante em PLN que consiste em rotular palavras de uma sentença com etiquetas morfossintáticas que as identificam como categorias gramaticais tais como substantivos, verbos, adjetivos, etc. O etiquetador automático (*tagger*) é a ferramenta que usa um vasto conjunto de técnicas e métodos, em diferentes abordagens, para automatizar essa tarefa.

O problema central nessa tarefa é a ambiguidade que ocorre devido às diferentes categorias gramaticais que um item léxico pode receber em diferentes contextos. O problema é resolvido quando, em um contexto específico, o etiquetador atribui a categoria gramatical apropriada desse item. A palavra “volta”, por exemplo, foi encontrada nos *corpora* deste estudo com cinco categorias diferentes: substantivo, verbo, verbo auxiliar, preposição e advérbio. Assim, na sentença “Congresso pede volta de órgão do governo”, o etiquetador deve ser capaz de analisar o contexto da sentença e rotular “volta” como um substantivo.

Existem três abordagens principais utilizadas pelos etiquetadores para rotular sentenças: 1) baseada em regras, que aplica grandes conjuntos de regras codificadas manualmente ou de forma semi-automática para desambiguizar as palavras (GREENE, RUBIN, 1971; KARLSSON et al., 1995; BICK, 2000); 2) probabilística, que, dada uma palavra e um conjunto finito de etiquetas possíveis para essa palavra, as quais podem ser buscadas, por exemplo, em um *corpus* eletrônico, aplica métodos de aprendizado de máquina para determinar a sequência ótima de etiquetas *T*, dada uma sequência de palavras *W* (TOUTANOVA et al., 2003; KEPLER, FINGER, 2006); e 3) híbrida, que surgiu a partir do método de aprendizado baseado em transformação dirigida por erro (*transformation-based error-driven learning* – TBL) proposto por Brill (1995), a qual combina as duas primeiras abordagens e extrai, automaticamente, uma lista ordenada de regras de *corpora* anotados para etiquetar palavras de uma sentença (FINGER, 2000; KINOSHITA, SALVADOR, MENEZES, 2007; SANTOS, MILIDIÚ, RENTERÍA, 2008). Esta última abordagem é a adotada para o desenvolvimento do etiquetador desta tese.

Diversas aplicações de PLN dependem do alto desempenho dos etiquetadores em termos de acurácia. Esse desempenho costuma ser avaliado comparando-se suas saídas com conjuntos de testes etiquetados por especialistas humanos chamados de *Gold Standard*. Para fazer essa comparação é necessário que ambos os conjuntos tenham as mesmas sentenças. O percentual de acertos ou acurácia é igual ao percentual de etiquetas no conjunto de teste em que o etiquetador e o *Gold Standard* concordam (JURAFSKY; MARTIN, 2000, p. 308).

Para facilitar a pesquisa em PLN, recursos de *corpora* eletrônicos anotados de várias línguas estão disponíveis na *Web*, por exemplo, os *corpora* da língua portuguesa CETENFolha (LINGUATECA, 2007), Mac-Morpho (LÁCIO-WEB, 2007) e Selva (LINGUATECA, 2009) usados neste trabalho, os quais permitem que os etiquetadores realizem aprendizado de máquina e façam a etiquetagem de novas sentenças baseada em padrões presentes nos dados. Esses *corpora* ou parte deles também têm sido utilizados como *Gold Standard*. No início dos anos 90, essas atividades, dentre outras, contribuíram para o sucesso de aplicações de PLN baseadas em *corpus* (NUGUES, 2006, p. 26). Os *corpora*, construídos no escopo da linguística de *corpus*, além de terem atividades estritamente linguísticas (voltadas para o estudo da língua), passaram a ter atividades mais aplicadas, no sentido de testar aplicações de PLN (SANTOS, 2008, p. 50). Também estão disponíveis na *Web* etiquetadores de uso livre implementados nas abordagens citadas, anteriormente, para testar as aplicações de PLN, por exemplo, o TreeTagger (SCHMID, 1994a, 1995), o QTag (MASON; TUFIS, 1997), o μ-TBL (LAGGER, 1999) e o etiquetador do analisador gramatical PALAVRAS (BICK, 2000) usados nos experimentos desta tese.

1.2 Motivação e Relevância

A etiquetagem morfossintática é uma tarefa de importância crescente para muitas aplicações de PLN, tais como análise gramatical, TA, anotação de *corpus*, extração e recuperação de informação, simplificação de textos, geração automática de resumos e outras aplicações que precisem extrair a categoria gramatical das palavras, por exemplo, ferramentas para auxiliar a elicitação e modelagem de requisitos em engenharia de requisitos (SILVA; MARTINS, 2008) e aplicações de síntese de fala (BELLEGARDA, 2010). Por essa razão, requer uma acurácia elevada na resolução do problema de ambiguidade. A disponibilidade de etiquetadores morfossintáticos de alta acurácia representa ainda um desafio para o sucesso de aplicações em linguística computacional.

Embora seja uma tarefa bem compreendida e tenha sido tema de inúmeras pesquisas, no estado da arte, a maioria das ferramentas de etiquetagem tanto para o português brasileiro, quanto para outras línguas, alcança acurácia em torno de 96 a 97%. Por exemplo, o etiquetador proposto por Santos, Milidiú e Rentería (2008) alcançou acurácia de 96,75% em textos jornalísticos do *Corpus* Mac-Morpho do português brasileiro, e o de Toutanova et al. (2003) alcançou acurácia de 97,24% em textos jornalísticos do *Corpus Wall Street Journal* do *Penn Treebank* da língua inglesa. Para muitos autores, por exemplo, Umansky-Pesin, Reichart e Rappoport (2010), valores nessa faixa são considerados como alta acurácia para textos do gênero jornalístico. Contudo, uma acurácia de 97% significa que muitas sentenças contêm erros de etiquetagem, prejudicando os resultados de analisadores gramaticais, tradutores automáticos e outros.

Considere-se o exemplo de um texto do português brasileiro contendo 1.130 palavras distribuídas em 31 parágrafos, 51 sentenças e 109 linhas que, se obtivesse os percentuais de acurácia de 96%, 97%, 98% e 99%, apresentaria a quantidade de erros mostrados na Tabela 1.1. Com base nos valores mostrados nessa tabela, observa-se que o valor ideal para a acurácia deveria estar acima de 99% e o mais próximo possível de uma linha de teto que se baseia na análise humana. Segundo experimentos realizados por Voutilainen (1995a, p. 174), a discussão sobre as etiquetas por um grupo de especialistas pode levar a um consenso em 100% das etiquetas. Entretanto, os fatores críticos envolvidos na tarefa de etiquetagem, tais como o método de etiquetagem, o *corpus* de treinamento (tamanho, qualidade, conjunto de etiquetas e gênero textual) e a presença de novas palavras ou palavras desconhecidas nos textos a serem etiquetados, precisam ser melhor compreendidos para que se possa aperfeiçoá-los em um ambiente de PLN de maneira a elevar o desempenho dessa tarefa.

TABELA 1.1 - Número de erros de etiquetagem por percentual de acurácia em um texto de 1.130 palavras do português brasileiro

Acurácia	**96%**	**97%**	**98%**	**99%**
Nº erros	≈45	≈34	≈23	≈11
Nº erros/parágrafo em média	≈ 2/parágrafo	≈ 1/parágrafo	≈ 1 a cada 1,5 parágrafo	≈ 1 a cada 3 parágrafos
Nº erros/sentença em média	≈ 1/sentença	≈ 1 a cada 1,5 sentença	≈ 1 a cada 2 sentenças	≈ 1 a cada 5 sentenças
Nº erros/linha em média	≈ 1 a cada 2,5 linhas	≈ 1 a cada 3,2 linhas	≈ 1 a cada 5 linhas	≈ 1 a cada 10 linhas

Isto indica que o problema da etiquetagem é relevante e ainda não totalmente resolvido, motivo pelo qual tem sido tema de estudos atuais. Qualquer pequena melhora em termos de acurácia pode representar uma diferença significativa na aplicação final. Kepler (2010), por exemplo, pesquisou a etiquetagem morfossintática para duas das palavras mais problemáticas em textos do português: “que” e “a”.

Com a disponibilidade de recursos de *corpora* e etiquetadores e considerando-se a investigação dos fatores envolvidos no processo, esta tese se fundamenta na ideia de que é possível elevar os percentuais de acurácia do estado da arte na etiquetagem morfossintática de sentenças em textos de diferentes domínios.

1.3 Objetivos

O objetivo geral desta pesquisa é conceber um etiquetador morfossintático, de uso livre, capaz de alcançar acurácia de 98% em ambientes de pelo menos dois domínios diferentes a partir de *corpora* do português brasileiro de textos jornalísticos.

Os objetivos específicos envolvem:

- estudar as abordagens baseada em regras, probabilística e híbrida para etiquetagem;
- elevar o percentual de acurácia em relação aos etiquetadores do estado da arte;
- focar em *corpora* do português brasileiro do gênero jornalístico para treinamento e dos gêneros jornalístico e científico para testes[1];
- avaliar por intermédio de experimentos práticos *softwares* que se encontram disponíveis na *Web* para etiquetar palavras;
- formalizar a arquitetura do etiquetador;
- apresentar uma análise de problemas com soluções adotadas.

1.4 Premissas

As premissas adotadas nesta tese são as de que:

- os etiquetadores atuais, construídos com base em métodos de aprendizagem de máquina variados e/ou a partir da aplicação de regras codificadas manualmente, parecem convergir quanto ao desempenho em termos de acurácia;
- o uso de uma abordagem híbrida (probabilística + baseada em regras) eleva o percentual de acurácia do etiquetador;
- a análise dos erros resultantes desse processamento híbrido permite visualizar problemas no *corpus* e fazer correções no mesmo, proporcionando com isso, o aumento de acurácia na etiquetagem;
- a análise dos erros permite também melhorar e aumentar, gradativamente, o *corpus* de treinamento, disponibilizando-o para outras aplicações;
- um conjunto de etiquetas com mais informações lexicais, por exemplo, gênero, número, tempo e modos verbais, eleva o percentual de acurácia do etiquetador;
- o uso de estratégias eficientes para desambiguizar palavras desconhecidas nos textos a serem etiquetados eleva o percentual de acurácia do etiquetador;
- o método de avaliação da acurácia é importante para que o desempenho do etiquetador possa ser bem estimado em novos textos. Nesta tese, é usado o

[1] A escolha de textos jornalísticos para treinar o etiquetador, bem como de textos jornalísticos e científicos para testá-lo, foi feita por haver a disponibilidade de *corpora* anotados desses gêneros textuais, revisados por especialistas humanos, e por se pretender usar o etiquetador, futuramente, em um ambiente de ensino virtual que trabalha com textos acadêmicos.

método de validação cruzada (*cross-validation*) com 20 divisões para a estimativa de acurácia do etiquetador treinado com um *corpus* de pequeno tamanho. Para um *corpus* de maior tamanho, é usado o mesmo método, mas com 10 divisões em 90% do *corpus*; um subconjunto de 10% do *corpus* fica separado apenas para teste e não é visto no treinamento. Além disso, é feita a avaliação do etiquetador em textos do mesmo domínio e em textos de domínio diferente do qual foi treinado.

1.5 Metodologia

A abordagem de desenvolvimento de etiquetador proposta nesta tese modela um etiquetador baseado em *corpus* com um método de etiquetagem híbrido de duas etapas principais: na primeira etapa, usa um etiquetador probabilístico de uso livre, o TreeTagger; na segunda etapa, usa um etiquetador baseado em regras que combina módulos de regras codificadas manualmente com um módulo de regras extraídas automaticamente com o sistema μ-TBL. A metodologia para realização desta tese consistiu em:

- realizar revisão bibliográfica associada aos assuntos abordados na tese e sobre trabalhos correlatos;
- analisar os elementos do etiquetador por meio de experimentos práticos com etiquetadores de uso livre disponíveis na *Web* (QTag, TreeTagger e μ-TBL) e *corpora* do português brasileiro (CETENFolha, Bosque CF 7.4, Mac-Morpho e Selva Científica); e uso de diferentes técnicas de avaliação de acurácia (validação cruzada, *hold-out*), explorando como os seguintes fatores influenciam na acurácia de um etiquetador: 1) algoritmo de aprendizado (probabilísticos, baseados em regras, híbridos); 2) conjunto de regras empregado para a correção de erros (manuais, extraídas automaticamente); 3) *corpora* (tamanho, qualidade, conjunto de etiquetas e gênero textual); e 4) presença de palavras desconhecidas nos textos a serem etiquetados;
- elaborar *softwares* de apoio na linguagem Java para a realização dos experimentos com diferentes etiquetadores e *corpora*, por exemplo, extratores de palavras e etiquetas de *corpora*, segmentadores e concatenadores de palavras e sentenças, programas para formatar textos, integrar os módulos do

etiquetador, criar léxicos, aplicar regras, comparar arquivos de resultados, calcular frequências de palavras e calcular a acurácia dos etiquetadores;

- especificar os módulos constituintes do modelo de etiquetador proposto.

A metodologia de avaliação e validação do modelo de etiquetador proposto consistiu, resumidamente, em:

1) realizar treinamento e testes do etiquetador com textos de *corpora* do português brasileiro nas abordagens baseada em regras, probabilística e híbrida;
2) testar diferentes estratégias e avaliar o desempenho do etiquetador após a aplicação de cada uma das estratégias propostas até encontrar a menor taxa de erros possível.

1.6 Trabalhos Publicados

Os resultados obtidos da aplicação da metodologia foram apresentados nos trabalhos publicados:

1. DOMINGUES, M. L.; FAVERO, E. L. Domain Adaptation in Part-of-Speech Tagging. In: BANDYOPADHYAY, S.; NASKAR, S.K.; EKBAL, A. (Eds.). **Emerging Applications of Natural Language Processing**: Concepts and new research. IGI Global, 2011. (no prelo).

2. DOMINGUES, M. L.; MEDEIROS, I. P.; FAVERO, E. L. O desenvolvimento de um etiquetador morfossintático com alta acurácia para o português. In: TAGNIN, S. E. O.; VALE, O. A. (Eds.). **Avanços da Linguística de Corpus no Brasil.** São Paulo, SP: Humanitas, 2008. p. 267-286.

3. DOMINGUES, M. L.; MEDEIROS, I. P.; FAVERO, E. L. O Desenvolvimento de um Etiquetador Morfossintático com alta acurácia para o Português. In: ENCONTRO DE LINGUÍSTICA DE CORPUS, 6., 2007, São Paulo. **Anais**... São Paulo: FFLCH/USP. 2007. Disponível em:<http://www.nilc.icmc.usp.br/viencontro/anais.htm>.

4. DOMINGUES, M. L.; MEDEIROS, I. P.; FAVERO, E. L. Etiquetagem de Palavras para o Português do Brasil. In: WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 5., 2007, Rio de Janeiro. **Anais**... Rio de Janeiro: SBC - Sociedade Brasileira de Computação, 2007. p. 1721-1724. Disponível em: <http://www.nilc.icmc.usp.br/til2007/>.

1.7 Contribuições

A principal contribuição desta tese é conceber um etiquetador morfossintático baseado em TBL, capaz de alcançar acurácia superior a 98% em textos de dois domínios diferentes, jornalístico e científico, a partir de *corpora* do português brasileiro de textos jornalísticos, aplicando três estratégias (usar um léxico de nomes próprios, adicionar regras codificadas manualmente e treinar o etiquetador probabilístico com um léxico de grande tamanho) testadas experimentalmente e adicionadas ao modelo híbrido de etiquetador. Apresentar os problemas encontrados e analisados no trabalho também é de grande valia para proporcionar avanço significativo em sistemas da área, em busca de um etiquetador morfossintático mais confiável.

A abordagem TBL foi explorada em uma combinação de ferramentas que usam métodos bem sucedidos de etiquetagem desenvolvidos ao longo dos anos e aumenta seus desempenhos com as três estratégias propostas, tentando extrair benefícios dos recursos disponíveis para a língua portuguesa. Esta tese aprofunda a pesquisa sobre os fatores críticos da tarefa de etiquetagem e propõem soluções que elevam a acurácia do estado da arte em cerca de 1% a 2% para o português brasileiro.

1.8 Organização do Texto

Este trabalho está organizado em sete capítulos, resumidos as seguir:

Capítulo 1. INTRODUÇÃO. Apresenta-se a contextualização do tema desta tese, motivação e relevância, objetivos, premissas, metodologia, trabalhos publicados, contribuições e esta organização do texto.

Capítulo 2. ETIQUETAGEM MORFOSSINTÁTICA. Apresenta-se a fundamentação teórica sobre etiquetagem morfossintática, com os principais conceitos, as abordagens: baseada em regras, probabilística e híbrida, e os métodos e técnicas relacionados.

Capítulo 3. REVISÃO BIBLIOGRÁFICA. Faz-se uma revisão bibliográfica sobre as abordagens para etiquetadores morfossintáticos, com foco para a acurácia alcançada na etiquetagem.

Capítulo 4. DESCRIÇÃO DE ABORDAGEM DE UM ETIQUETADOR DE ALTA ACURÁCIA PARA O PORTUGUÊS BRASILEIRO. Descreve-se os recursos experimentais, a metodologia da pesquisa, a arquitetura do etiquetador proposto e as etapas do processo de etiquetagem.

Capítulo 5. ETIQUETAGEM BASEDA EM TEXTOS DO BOSQUE CETENFOLHA. Apresenta-se as duas estratégias usadas, inicialmente, nesta pesquisa (aplicação de regras para correção de erros da etiquetagem probabilística e modificação no conjunto inicial de etiquetas), avaliadas em experimentos nos quais o etiquetador é treinado com um *corpus* de pequeno tamanho, o Bosque CF 7.4, e testado em textos desse mesmo *corpus*, seus resultados e a sua discussão. A modificação no conjunto inicial de etiquetas não faz parte da solução final de etiquetador adotada nesta tese, mas poderá ser considerada em trabalhos futuros, com a aquisição de *corpora* de treinamento adequados.

Capítulo 6. ETIQUETAGEM BASEDA EM TEXTOS DO *CORPUS* MAC-MORPHO. Apresenta-se as três estratégias (o uso de um léxico de nomes próprios, a adição de regras codificadas manualmente e o uso de um léxico de grande tamanho para treinamento do TreeTagger) avaliadas em experimentos nos quais o etiquetador é treinado com um *corpus* de maior tamanho, o *Corpus* Mac-Morpho, e testado em textos dos *corpora* Mac-Morpho e Bosque CF 7.4, seus resultados e a sua discussão.

Capítulo 7. ETIQUETAGEM EM TEXTOS DE GÊNEROS DIFERENTES. Descreve-se a etiquetagem do *Corpus* Selva Científica do gênero científico a partir do etiquetador treinado com textos jornalísticos do *Corpus* Mac-Morpho, a avaliação dos resultados iniciais, os ajustes realizados no etiquetador para melhorar a acurácia, os resultados finais e sua discussão.

Capítulo 8. CONCLUSÕES E TRABALHOS FUTUROS. Faz-se a avaliação dos resultados obtidos e principais contribuições. Faz-se ainda sugestões para o sucesso de trabalhos futuros.

2 Etiquetagem Morfossintática

"A perhaps surprising fact about the six categories of linguistic knowledge is that most or all tasks in speech and language processing can be viewed as resolving ***ambiguity*** *at one of these levels. We say some input is ambiguous if there are multiple alternative linguistic structures than can be built for it. [...]" (JURAFSKY; MARTIN, 2000, p. 4)*

A etiquetagem morfossintática é uma tarefa de PLN que envolve o uso de variados métodos, técnicas e abordagens. Este capítulo apresenta a fundamentação teórica sobre os tópicos relevantes para a pesquisa e aplicação desta tarefa sob a ótica da linguística computacional.

2.1 Níveis de Análise da Linguagem

Sistemas de linguagem natural envolvem processos complexos como a análise sintática e inferências semânticas de alto nível, dentre outros. Isto requer um conhecimento considerável, não apenas sobre a estrutura da linguagem, mas também sobre o conhecimento geral do mundo e a habilidade de raciocínio dos humanos. Para lidar com esta complexidade, os linguistas definiram diferentes níveis de análise da linguagem falada (fala) e escrita (texto), a saber:

- **Fonética e fonologia**: estudam a relação entre as palavras e os sons que são combinados para formar a linguagem, conhecimento que é importante para sistemas de reconhecimento e geração computadorizada de fala (ALLEN, 1995; LUGER, 2004).
- **Morfologia**: aborda os componentes (morfemas) que constituem as palavras, as regras que governam a formação das palavras e a análise para determinar o papel de uma palavra em uma sentença (LUGER, 2004). Nesse nível, um analisador morfológico é capaz de identificar palavras ou expressões isoladas em uma sentença com o auxílio de delimitadores que variam conforme a ferramenta (por exemplo, sinais de tabulação e espaços em branco) (RICH; KNIGHT, 1993) e classificá-las em categorias gramaticais por exemplo.
- **Sintaxe**: estuda as regras para se combinar palavras, frases e sentenças e o uso destas regras para analisar e gerar sentenças corretas (LUGER, 2004). Determina o papel estrutural de cada palavra na sentença e quais frases são

subpartes de quais outras frases (ALLEN, 1995). Os analisadores sintáticos são as ferramentas responsáveis pela recuperação automática da estrutura sintática de sentenças em linguagem natural.

- **Semântica**: diz respeito ao significado das palavras e como esses significados se combinam em sentenças para expressar uma determinada ideia. O significado é estudado independente de contexto (ALLEN, 1995). Um exemplo é mapear palavras isoladas para os objetos apropriados em uma base de dados e criar estruturas corretas que correspondem ao modo como os significados das palavras isoladas combinam entre si (LUGER, 2004, p. 514).
- **Pragmática**: se aplica a como as sentenças são usadas em diferentes situações e como o seu uso afeta a interpretação da sentença (ALLEN, 1995). Por exemplo, tratar o motivo de “Sim” ser normalmente uma resposta inadequada à pergunta: “Você sabe que horas são?” (LUGER, 2004).
- **Discurso**: se refere a como as sentenças imediatamente anteriores afetam a interpretação da próxima sentença, tarefa que é importante na interpretação de pronomes e na interpretação de aspectos temporais da informação transmitida (ALLEN, 1995). Sistemas para resolução de anáfora intersentencial (anáforas pronominais, como ele, este, etc.) e resolução de co-referência textual são exemplos tratados nesse nível de análise linguística.

2.2 Linguística Computacional

A investigação da linguagem e das línguas naturais com a utilização de recursos computacionais é tema da linguística computacional, área de pesquisa que se originou da interseção entre a linguística e a uma subárea da ciência da computação, a inteligência artificial. A linguística computacional busca investigar, propor e desenvolver formalismos, modelos, técnicas, métodos e sistemas computacionais que têm como objeto primário a linguagem natural (NUNES, 2007).

Segundo Othero e Menuzzi (2005), a linguística computacional pode ser subdividida em duas subáreas: a linguística de *corpus* e o processamento de linguagem natural (PLN).

2.2.1 Linguística de *Corpus*

A linguística de *corpus* se aplica ao estudo da língua pela observação de *corpora* eletrônicos, grandes bases de dados que contêm amostras de linguagem natural. No conceito de Berber Sardinha (2004), a linguística de *corpus*:

> Ocupa-se da coleta e da exploração de *corpora*, ou conjunto de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas extraídas por meio de computador.

Atualmente, estão disponíveis grandes *corpora* eletrônicos anotados, aqueles cujas palavras já se encontram etiquetadas, que permitem que etiquetadores automáticos realizem aprendizado de máquina (supervisionado ou não-supervisionado) e façam a etiquetagem de novas sentenças baseada em padrões presentes nos dados.

Segundo Berber Sardinha (2004), a classificação de um *corpus* quanto ao seu tamanho inclui as categoriais mostradas na Tabela 2.1. Em geral, quanto maior o *corpus*, mais representativo, pois engloba maior parcela da população estudada. Porém, uma amostra pequena pode ser representativa, na medida em que reflita a composição da população.

TABELA 2.1 - Classificação de *corpus* quanto ao tamanho

Tamanho em Palavras	**Classificação**
Menos de 80 mil	Pequeno
80 a 250 mil	Pequeno-médio
250 a 1 milhão	Médio
1 milhão a 10 milhões	Médio-grande
10 milhões ou mais	Grande

Fonte: Berber Sardinha, 2004, p. 26.

Os *corpora* etiquetados automaticamente costumam apresentar grande porcentagem de erros ou ruídos, uma vez que não existe um etiquetador capaz de alcançar 100% de acurácia. Para melhorar a qualidade, é feita uma revisão por especialistas para a remoção dos erros.

2.2.2 Processamento de Linguagem Natural (PLN)

PLN se ocupa do estudo da linguagem voltado para a construção de ferramentas computacionais específicas voltadas para tarefas básicas direcionadas ao processamento da informação baseada na linguagem humana tais como: delimitadores de sentenças (*sentence delimiters*), itemizadores (*tokenizers*), identificadores de radicais de palavras (*stemmers*), codificadores de partes do discurso oral ou escrito (*speech taggers*), identificadores de sintagmas nominais (*noun phrase recognizers*) e de nomes próprios e analisadores gramaticais (*parsers*), dentre outros (JACKSON, MOULINIER, 2002; NUNES, 2007).

A ideia é fragmentar um problema de PLN, por exemplo, análise gramatical, tradução automática e outras, e solucioná-lo por meio da combinação das soluções das partes. Isto ocorre em virtude do insucesso obtido nas pesquisas iniciais, em que a primeira tarefa de PLN estudada foi TA. Nesse caso, se almejou um tratamento textual de forma completa e sofisticada, em todos os seus níveis, similar à tradução humana, o que se revelou de grande complexidade e relegou a pesquisa de TA a um plano secundário por várias décadas (NUNES, 2007).

2.3 Etiquetagem Automática

A etiquetagem automática emprega recursos computacionais para a classificação das unidades do texto em diferentes níveis da análise linguística, por exemplo: morfossintáticas (etiquetadores morfossintáticos), sintáticas (analisadores gramaticais), semânticas (analisador semântico ou interpretador), discursivas (analisador discursivo).

Cada caso requer linguagens de anotação para representar as classes como: etiquetas ou símbolos para etiquetadores morfossintáticos (ADJ, ART, N, etc..); estruturas (árvores sintáticas) no caso dos analisadores gramaticais; relações (semânticas ou retóricas) para analisadores semânticos ou discursivos (NUNES, 2007).

Neste trabalho, é pesquisada a etiquetagem morfossintática, que se situa nos níveis da morfologia e da sintaxe e serve de alicerce, por exemplo, para a interpretação sintática, em que uma gramática poderia descrever um conjunto restrito de frases de uma linguagem com regras como: "*Uma sentença é formada por uma frase nominal e uma frase verbal*.", "*Uma frase nominal é formada por um artigo e um nome*" e "*Uma frase verbal é formada por um verbo e uma frase nominal*" (NUGUES, 2006, p. 9).

Conforme o etiquetador utilizado, algumas tarefas de pré-processamento devem ser feitas para que se possa realizar a etiquetagem automática, tais como:

- preparar as sentenças do arquivo de entrada de dados no formato adequado para o etiquetador em uso, tal como colocar uma palavra e sua etiqueta por linha para arquivos de treinamento ou uma palavra por linha para arquivos a serem etiquetados;

- separar palavras e etiquetas por sinais de tabulação, espaços em branco ou outros, ou separar palavras de sinais de pontuação;
- preparar arquivo de recursos léxicos contendo as palavras do *corpus* e a(s) etiqueta(s) que cada palavra pode receber (etiquetas possíveis);
- identificar, listar e disponibilizar informações estatísticas de palavras que não pertencem ao conjunto de treinamento, isto é, palavras desconhecidas (em inglês, *out-of-vocabulary words* – OOV).

2.3.1 Etiquetagem Morfossintática

A etiquetagem morfossintática é uma tarefa básica de PLN que rotula palavras, expressões multi-palavras (por exemplo, *apesar_de*) e sinais de pontuação de uma sentença como substantivos, verbos, adjetivos, etc. Por exemplo, dadas como entrada a um etiquetador apropriado as sequências de palavras “Congresso pede volta de órgão do governo.” e “Ele só volta ao Rio segunda-feira.” e um conjunto de etiquetas específico (*tagset*), esse etiquetador deve produzir como resultado as seguintes saídas anotadas:

- Congresso/NPROP pede/V volta/N de/PREP órgão/N de/PREP|+ o/ART governo/N ./.
- Ele/PROPESS só/PDEN volta/V a/PREP|+ o/ART Rio/NPROP segunda-feira/N ./.

As etiquetas do exemplo são provenientes do *corpus* Mac-Morpho (ALUÍSIO et al., 2003) e são descritas como: NPROP=nome próprio, V=verbo, N=nome ou substantivo, PREP=preposição, PREP|+=preposição com contração, ART=artigo, PROPESS=pronome pessoal, PDEN=palavra denotativa e o sinal de pontuação .=.

Quando as palavras possuem mais de uma categoria gramatical possível, como a palavra “volta” nos exemplos acima, que pode se referir ao nome *volta* ou ao verbo *voltar*, diz-se que a palavra é ambígua. A tarefa de etiquetagem não é trivial, uma vez que é necessário resolver o problema da ambiguidade de acordo com o contexto da palavra na sentença e, além disso, sustentar alta acurácia para textos novos. Para essa finalidade, os etiquetadores usam um vasto conjunto de técnicas e métodos em diferentes abordagens para etiquetar, de forma automática, as palavras com a maior acurácia possível.

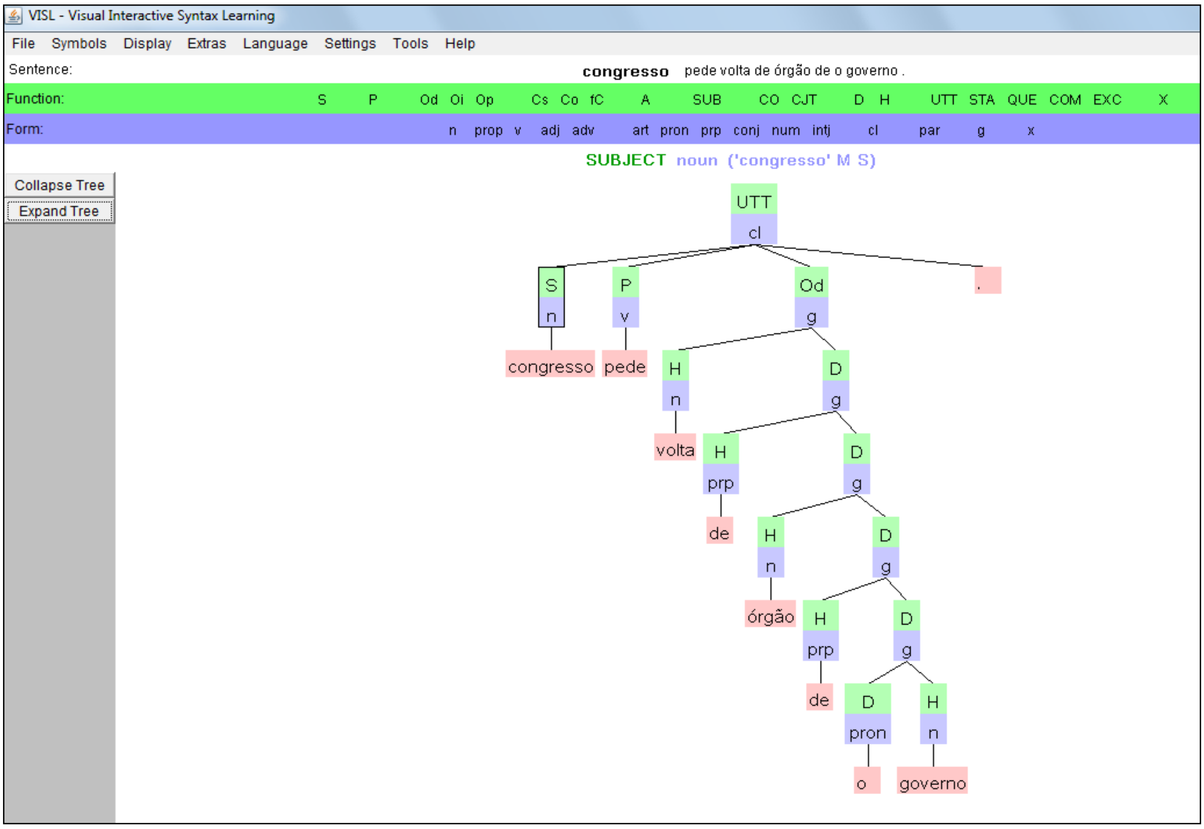
As etiquetas podem incluir também outros atributos léxicos como gênero, número, pessoa, tempo e modo verbais, dentre outros. Na frase “Congresso pede volta de órgão do governo.”, além das categorias gramaticais, podem fazer parte das etiquetas atributos como gênero feminino ou masculino (F, M), número singular (S), terceira pessoa verbal (3a), tempo presente (PR) e modo verbal indicativo (IND) com as seguintes saídas anotadas:

- Congresso/NPROPMS pede/ V3aSPRIND volta/NFS de/PREP órgão/NMS de/PREP|+ o/ARTMS governo/NMS ./.

A etiquetagem morfossintática costuma ser usada como tarefa de pré-processamento por outras aplicações, por exemplo, por um analisador gramatical ou *parser* que, inicialmente, usa um etiquetador para atribuir as etiquetas morfossintáticas às palavras de uma sentença e, posteriormente, usa a saída anotada pelo etiquetador para subsidiar a análise gramatical e, então, rotular os elementos dessa sentença com suas categorias sintáticas. Na Figura 2.1, é ilustrado um exemplo, em que a frase “Congresso pede volta de órgão do governo.” é dada como entrada ao analisador gramatical PALAVRAS (Bick, 2000) no ambiente *Visual Interactive Syntax Learning* – VISL (VISL, 2011) e, ao ser processada pelo analisador, este fornece como saída uma árvore sintática, na qual se observa na parte inferior das etiquetas as categorias gramaticais (n=nome, v=verbo, etc.) atribuídas pelo etiquetador e na parte superior das etiquetas as categorias sintáticas (S=sujeito, P=predicador, Od=objeto direto, etc.) atribuídas pelo *parser*[2]. Ao passar o *mouse* sobre as etiquetas sintáticas, o analisador exibe a descrição completa de cada constituinte da sentença, por exemplo, sobre as etiquetas da palavra “congresso”, em que se vê na figura a informação “SUBJECT noun (‘congresso’ M S)” que corresponde a: etiqueta sintática=sujeito; etiqueta gramatical=nome; palavra=congresso; e etiquetas de inflexão de gênero=masculino e número=singular.

[2] Os *links* para as descrições dos vários conjuntos de etiquetas usados pelo PALAVRAS são listados em: http://beta.visl.sdu.dk/visl/pt/info/

FIGURA 2.1 - Árvore sintática do analisador gramatical PALAVRAS



Fonte: VISL, 2011.

2.3.1.1 Linha básica para avaliação da acurácia

Charniak et al. (1993) estabeleceram uma linha básica para avaliação da acurácia de um etiquetador ao demonstrarem que um algoritmo estocástico que simplesmente atribui a etiqueta mais frequente para uma determinada palavra em um conjunto de treinamento atinge uma acurácia de 90 a 91%. A linha de teto se baseia na análise humana e segundo estudos de Voutilainen (1995a), se um grupo de especialistas for autorizado a discutir sobre as etiquetas, podem chegar a um consenso em 100% das etiquetas.

2.3.1.2 Classes abertas e fechadas

Alguns etiquetadores morfossintáticos utilizam no processamento informações sobre as classes de palavras, as quais podem pertencer a duas supercategorias: classes abertas e classes fechadas.

- Classes abertas: englobam um conjunto potencialmente infinito de palavras, passíveis de receber novas unidades sempre que se justifique. São classes abertas: substantivos, adjetivos, verbos e advérbios terminados em –*mente*.
- Classes fechadas: são aquelas cujos elementos são finitos e contáveis e às quais a evolução da língua só muito raramente acrescenta novos membros. São classes fechadas: determinantes (artigos, possessivos, demonstrativos, indefinidos, interrogativos); pronomes (pessoais, possessivos, demonstrativos, indefinidos, interrogativos, relativos); numerais (cardinais, ordinais, coletivos); advérbios (de tempo, de modo, de quantidade, de afirmação, de negação, de dúvida, de exclusão, de inclusão, de designação) e locuções adverbiais; preposições; conjunções (coordenativas, subordinativas) e interjeições.

2.4 Abordagens para Etiquetagem Morfossintática

No decorrer dos últimos trinta anos, as pesquisas sobre etiquetagem morfossintática têm focado em três abordagens principais: baseada em regras (VOUTILAINEN, 1995; BICK, 2000), probabilística (KEMPE, 1993; KEPLER, 2010) e híbrida (BRILL, 1995; SANTOS, MILIDIÚ, RENTERÍA, 2008).

2.4.1 Abordagem Baseada em Regras

Etiquetadores baseados em regras geralmente abrangem uma grande base de regras de desambiguização codificadas manualmente, por exemplo, a regra:

> “... uma palavra ambígua é *nome* em vez de *verbo* se ela é precedida de um *artigo*.” (JURAFSKY; MARTIN, 2000, p. 300).

Os primeiros algoritmos possuíam uma arquitetura de dois estágios: no primeiro, era utilizado um dicionário para atribuir a cada palavra uma lista de categorias gramaticais possíveis; no segundo, grandes listas de regras de desambiguização codificadas manualmente eram utilizadas de forma a atribuir uma única categoria para cada palavra (KLEIN, SIMMONS, 1963; GREENE, RUBIN, 1971). Anos depois, surgiram etiquetadores baseados em uma arquitetura de dois estágios com a diferença de que o léxico e as regras de desambiguização eram mais sofisticados do que os anteriores.

Para exemplificar essa abordagem, é apresentado o etiquetador ENGTWOL (VOUTILAINEN, 1995a), baseado na arquitetura de Gramática Constritiva (*Constraint Grammar* – CG) de Karlsson et al. (1995) com um léxico de cerca de 56.000 entradas léxicas para radicais de palavras (*stems*) em inglês (HEIKKILÄ, 1995), em que palavras com múltiplas categorias gramaticais são contadas como entradas separadas. Cada entrada é anotada com um conjunto de atributos morfológicos e sintáticos (morfologia de dois níveis). Por exemplo, a palavra inglesa *she* possui a etiqueta *PRON* (pronome) e os atributos adicionais *PERSONAL FEMININE NOMINATIVE SG3* (pessoal, feminino, nominativo, 3ª pessoa do singular). No primeiro estágio, cada palavra de uma sentença é processada pelo transdutor léxico de dois níveis e as entradas para todas as etiquetas possíveis são retornadas. Depois, um conjunto de cerca de 1.100 restrições linguísticas (*constraints*) é aplicado à sentença de entrada para excluir as etiquetas incorretas. As restrições são usadas de forma negativa para eliminar etiquetas inconsistentes com o contexto. Assim, uma versão simplificada de restrição para o advérbio inglês *that* da sentença *it isn't that odd* (não é assim <u>tão</u> estranho) é escrita da seguinte forma (JURAFSKY; MARTIN, 2000, p. 302):

Entrada: "that"
se
(+1 A/ADV/QUANT); /*a próxima palavra é adjetivo, advérbio ou quantificador*/
(+2 SENT-LIM); /*e é seguido de um limitador de sentença,*/
(NOT -1 SVOC/A); /*e a palavra anterior não é um verbo como 'consider',*/
/*que permite adjetivos como complemento do objeto*/
então eliminar etiquetas não-ADV
senão eliminar etiqueta ADV

As duas primeiras cláusulas da regra checam se *that* precede diretamente um adjetivo, advérbio ou quantificador no final da sentença. Em todos os outros casos, o advérbio é eliminado. A última cláusula elimina casos precedidos por verbos como *consider* ou *believe* que podem ser um substantivo e um adjetivo. Isto evita a etiquetagem de *that* como advérbio em frases do tipo: *I consider that odd* (Considero <u>isso</u> estranho).

A etiquetagem baseada em regras é sensível a mudanças no conjunto de palavras como neologismos e palavras que caem em desuso, ficando assim suscetível a falhas.

2.4.2 Abordagem Probabilística

Os algoritmos estocásticos empregados na etiquetagem são centrados na ideia de rotular palavras com suas etiquetas mais prováveis (JURAFSKY; MARTIN, 2000). Na etiquetagem probabilística, cada palavra possui um conjunto finito de etiquetas possíveis. Essas etiquetas podem ser buscadas em *corpora* eletrônicos, por exemplo. “Quando uma palavra possui mais de uma etiqueta possível, métodos estocásticos permitem determinar a sequência ótima de etiquetas $T = t_1, t_2, ..., t_n$, dada a sequência de palavras $W = w_1, w_2, ..., w_n$.” (NUGUES, 2006).

Conforme foi mencionado anteriormente, um algoritmo estocástico que simplesmente atribui a etiqueta mais frequente para uma determinada palavra em um conjunto de treinamento atinge uma acurácia de 90 a 91% (CHARNIAK et al., 1993). O desempenho desse etiquetador se tornou uma linha básica aos estudos subsequentes, levando os etiquetadores a adotarem uma combinação de informação contextual (informação sobre as sequências de etiquetas) e informação léxica (atribuir uma etiqueta com base na palavra) (MANNING; SCHUTZE, 1999).

2.4.2.1 Modelos de *N*-gramas

Modelos de *N*-gramas são modelos probabilísticos empregados para atribuir probabilidades a cadeias de palavras, quer para calcular a probabilidade de uma sentença inteira, quer para fazer uma predição probabilística do que será a próxima palavra em uma sequência. Para cada palavra, são analisadas as $n-1$ palavras anteriores. O modelo mais simples possível dessas sequências é aquele em que qualquer palavra de uma língua pode seguir qualquer outra palavra (JURAFSKY; MARTIN, 2000, p. 196).

Para calcular a probabilidade de uma cadeia completa de palavras (que pode ser representada como $w_1, w_2, ..., w_n$ ou w_1^n), se for considerada cada palavra ocorrendo em sua posição correta como um evento independente, essa probabilidade pode ser representada da seguinte forma, após ser decomposta pela aplicação da regra da cadeia (JURAFSKY; MARTIN, 2000, p. 197):

$$P(w_1^n) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1^2)...P(w_n \mid w_1^{n-1})$$

$$= \prod_{k=1}^{n} P\left(w_k \mid w_1^{k-1}\right) \tag{2.1}$$

Para calcular a probabilidade de uma palavra, dada uma longa sequência de palavras que a precedem, utiliza-se uma simplificação útil com a seguinte aproximação: calcula-se a probabilidade da palavra dada apenas a palavra anterior (bigrama). O modelo de bigrama aproxima a probabilidade de uma palavra, dadas todas as palavras anteriores $P\left(w_n \mid w_1^{n-1}\right)$ pela probabilidade condicional da palavra anterior $P\left(w_n \mid w_{n-1}\right)$. Assim, para uma gramática de bigramas, calcula-se a probabilidade de uma cadeia completa com a seguinte equação (JURAFSKY; MARTIN, 2000, p. 198):

$$P\left(w_1^n\right) \approx \prod_{k=1}^{n} P\left(w_k \mid w_{k-1}\right) \tag{2.2}$$

A hipótese de que a probabilidade de uma palavra depende apenas da palavra anterior é chamada de hipótese de Markov. Os **bigramas**, em que se analisa apenas a palavra anterior, são conhecidos como modelos de Markov de primeira ordem. Os **trigramas**, em que são analisadas as duas palavras anteriores, são chamados de modelos de Markov de segunda ordem.

2.4.2.2 Modelo de Markov Oculto (HMM)

Dada uma sentença ou uma sequência de palavras, os etiquetadores baseados em Modelo de Markov Oculto ou *Hidden Markov Model* (HMM) escolhem a sequência de etiquetas que maximiza a seguinte fórmula (JURAFSKY; MARTIN, 2000, p. 303):

$$P(\text{palavra} \mid \text{etiqueta})\,P(\text{etiqueta} \mid n\ \text{etiquetas prévias}) \tag{2.3}$$

A explicação para se chegar à fórmula 2.3 é descrita a seguir. Esta abordagem assume que está sendo calculada para cada sentença, a sequência de etiquetas mais provável $T = t_1, t_2, ..., t_n$ para a sequência de palavras na sentença ($W = w_1, w_2, ..., w_n$), dado um conjunto de etiquetas possíveis τ. A sequência de etiquetas ótima, dada uma sequência de palavras, corresponde à maximização da probabilidade condicional (JURAFSKY, MARTIN, 2000, p. 305; NUGUES, 2006, p. 163):

$$\hat{T} = P(t_1, t_2, ...t_n \mid w_1, w_2, ..., w_n) \qquad (2.4)$$

Seja $P(W) = P(w_1, w_2, ..., w_n)$ e $P(T) = P(t_1, t_2, ..., t_n)$.

$$\hat{T} = \underset{T \in \tau}{\arg\max}\, P(T \mid W) \qquad (2.5)$$

Pelo Teorema de Bayes, $P(T|W)$ pode ser representado como:

$$P(T|W) = \frac{P(T)P(W \mid T)}{P(W)} \qquad (2.6)$$

A sequência de etiquetas que maximiza a equação 2.6 é dada por:

$$\hat{T} = \underset{T \in \tau}{\arg\max}\, \frac{P(T)P(W \mid T)}{P(W)} \qquad (2.7)$$

Uma vez que o objetivo é encontrar a sequência de etiquetas ótima para uma sentença, dada uma sequência particular de palavras, $P(W)$ é constante e pode ser ignorada. Assim, a fórmula pode ser reescrita como:

$$\hat{T} = \underset{T \in \tau}{\arg\max}\, P(T)P(W \mid T) \qquad (2.8)$$

2.4.2.2.1 Trigrama ou Modelo de Markov de Segunda Ordem

Uma vez que é impossível de se obter estatísticas sobre sequências de um tamanho qualquer, é necessário realizar aproximações em $P(T)$ e $P(T|W)$ para tornar a estimativa tratável. Com o uso da hipótese de N-gramas para modelar a probabilidade de sequências de palavras, apresenta-se a seguir a definição do modelo de trigramas, por ser, frequentemente, o mais usado. O modelo de trigramas é usado pelo etiquetador probabilístico dos experimentos desta tese, o TreeTagger. Em geral, um produto de trigramas aproxima a sequência completa de etiquetas:

$$P(T) = P(t_1, t_2, ..., t_n) \approx P(t_1)P(t_2 \mid t_1)\prod_{i=3}^{n} P(t_i \mid t_{i-2}, t_{i-1}) \tag{2.9}$$

Se for usado um delimitador de início de sentença <s>, os dois primeiros termos do produto $P(t_1)$ $P(t_2|t_1)$ são reescritos como $P(\text{<s>})P(t_1|\text{<s>})$ $P(t_2|\text{<s>}, t_1)$, em que $P(\text{<s>})=1$.

Comumente, pode ser usada a estimativa de máxima verossimilhança (*maximum likelihood estimation* – MLE) das frequências relativas para estimar essas probabilidades:

$$P_{MLE}(t_i \mid t_{i-2}, t_{i-1}) = \frac{C(t_{i-2}, t_{i-1}, t_i)}{C(t_{i-2}, t_{i-1})} \tag{2.10}$$

em que $C(t_{i-2},t_{i-1},t_i)$ se refere ao número de ocorrências ou freqüência do trigrama t_{i-2},t_{i-1},t_i no *corpus* de treinamento e $C(t_{i-2},t_{i-1})$ se refere ao número de ocorrências do bigrama t_{i-2},t_{i-1}.

A sequência completa de palavras, uma vez que se sabe a sequência de etiquetas, é aproximada com a hipótese simplificada de que a probabilidade condicional de uma palavra depende apenas de sua etiqueta:

$$P(W \mid T) = P(w_1, w_2, ..., w_n \mid t_1, t_2, ..., t_n) \approx \prod_{i=1}^{n} P(w_i \mid t_i) \tag{2.11}$$

Da mesma forma que foi feito anteriormente, $P(w_i|t_i)$ é estimada de *corpora* anotados com o uso da MLE:

$$P_{MLE}(w_i \mid t_i) = \frac{C(w_i, t_i)}{C(t_i)} \tag{2.12}$$

em que $C(w_i,t_i)$ se refere ao número de ocorrências do par palavra e etiqueta w_i,t_i no *corpus* de treinamento e $C(t_i)$ se refere ao número de ocorrências da etiqueta t_i.

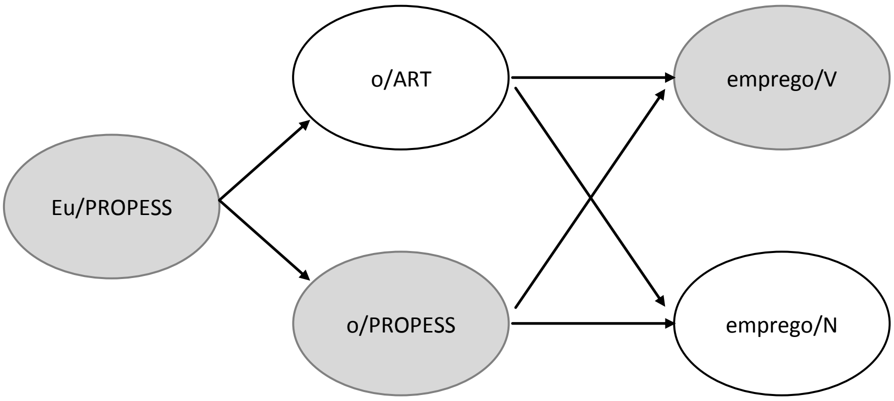
Assim, é escolhida a sequência de etiquetas que maximiza a fórmula 2.13, que corresponde à fórmula 2.3 aproximada para trigramas:

$$P(t_1)P(t_2 \mid t_1)\prod_{i=3}^{n} P(t_i \mid t_{i-2}, t_{i-1})\left[\prod_{i=1}^{n} P(w_i \mid t_i)\right] \tag{2.13}$$

As probabilidades sobre trigramas $P(t_i \mid t_{i-2}, t_{i-1})$ requerem uma estimativa para qualquer sequência de três etiquetas, a qual é obtida de *corpora* anotados. Se N_p é o número de diferentes etiquetas, há N_p x N_p x N_p valores para serem estimados. Muitas vezes, os dados anotados não são suficientes e algumas sequências estão ausentes, o que configura um problema de dados esparsos que pode ser evitado com o uso de técnicas de *smoothing* (NUGUES, 2006).

Como exemplo da etiquetagem probabilística, dada a sentença "Eu o emprego", na qual a palavra "Eu" é um pronome não ambíguo, a palavra "o" pode ser um artigo ou um pronome e a palavra "emprego" pode ser um nome ou um verbo, essa etiquetagem consiste em encontrar a sequência ótima dentre os quatro caminhos possíveis, conforme é ilustrado na Figura 2.2 (NUGUES, 2006).

FIGURA 2.2 - Sequências possíveis de etiquetas para uma sentença de três palavras



Fonte: Baseada em Nugues (2006, p. 165).

As etiquetas são do *corpus* Mac-Morpho (ALUÍSIO ET AL., 2003) e são descritas como: PROPESS=pronome pessoal, ART=artigo, V=verbo, N=nome ou substantivo comum.

No exemplo, cada transição é associada com um produto de probabilidade: $P(w_i \mid t_i)P(t_i \mid t_{i-2}, t_{i-1})$. A estimativa das sequências de etiquetas ao longo dos quatro caminhos é calculada pela multiplicação das probabilidades. A etiquetagem ótima corresponde ao máximo dos quatro valores:

1. *P*(PROPESS|<s>) × *P*(ART|<s>,PROPESS) × *P*(V|PROPESS,ART) × *P*(Eu|PROPESS) × *P*(o|ART) × *P*(emprego|V)
2. *P*(PROPESS|<s>) × *P*(ART|<s>,PROPESS) × *P*(N|PROPESS,ART) × *P*(Eu|PROPESS) × *P*(o|ART) × *P*(emprego|N)
3. *P*(PROPESS|<s>) × *P*(PROPESS|<s>,PROPESS) × *P*(V|PROPESS,PROPESS) × *P*(Eu|PROPESS) × *P*(o|PROPESS) × *P*(emprego|V)
4. *P*(PROPESS|<s>) × *P*(PROPESS|<s>,PROPESS) × *P*(N|PROPESS,PROPESS) × *P*(Eu|PROPESS) × *P*(o|PROPESS) × *P*(emprego|N)

Apesar de simples, esse método é custoso para sequências longas, uma vez que o cálculo com uma sentença de N palavras e um conjunto de etiquetas de T etiquetas terá uma complexidade com limite superior de N^T, isto é, exponencial. Os etiquetadores baseados em HMM costumam usar o algoritmo de Viterbi (VITERBI, 1967) para escolher a sequência de etiquetas mais provável para cada sentença de entrada.

2.4.2.3 Técnicas de *smoothing*

Modelos de *N*-gramas apresentam o problema de dados esparsos que ocorre quando palavras e sequências de etiquetas raras ou não vistas nos dados de treinamento recebem probabilidade zero no modelo treinado. Quando o modelo for aplicado na etiquetagem de uma dessas ocorrências, com a multiplicação das probabilidades, a probabilidade de toda a sentença será zero prejudicando a classificação das palavras. Para resolver esse problema, são aplicadas técnicas de *smoothing* (vide JURAFSKY; MARTIN, 2000, p. 206), por exemplo, *Add-One Smoothing*, *Witten-Bell Discounting*, *Good-Turing Discounting*, *Deleted Interpolation* e *Backoff*, dentre outras, que atribuem probabilidades diferentes de zero a *N*-gramas de probabilidade zero.

2.4.2.4 Outros métodos usados para etiquetagem

2.4.2.4.1 Árvores de Decisão

Um método conhecido de aplicação de árvore de decisão para etiquetar sentenças é o do etiquetador TreeTagger de Schmid (1994a, 1995), o qual é usado nos experimentos desta tese. Assim como os etiquetadores baseados em *N*-gramas convencionais (CHURCH, 1988; KEMPE, 1993), o TreeTagger modela a probabilidade de uma sequência etiquetada de palavras (no caso de um trigrama) recursivamente por:

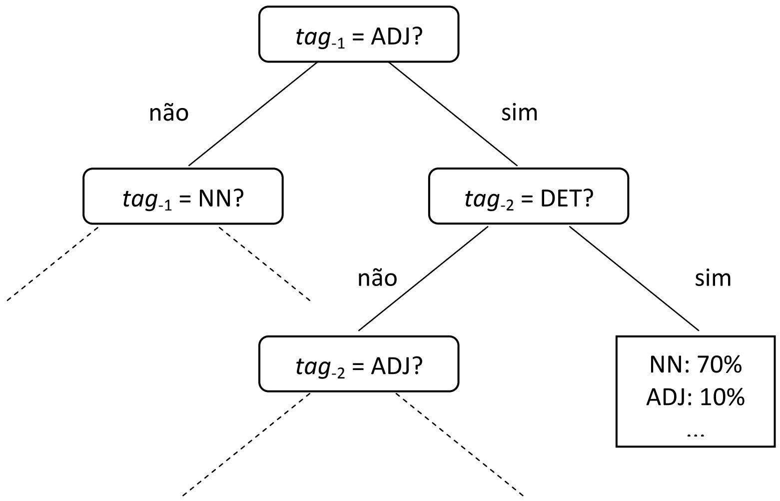
$$P(w_1 w_2 ... w_n, t_1 t_2 ... t_n) := P(t_n \mid t_{n-2} t_{n-1}) P(w_n \mid t_n) P(w_1 w_2 ... w_{n-1}, t_1 t_2 ... t_{n-1}) \qquad (2.14)$$

Os métodos diferem na forma pela qual a probabilidade de transição $p(t_n|t_{n\text{-}2}\ t_{n\text{-}1})$ é estimada. Os etiquetadores baseados em *N*-gramas costumam estimar a probabilidade usando a equação 2.12 baseada no princípio da MLE. Segundo Schmid (1994a), esse método de estimativa é problemático, visto que muitas frequências são pequenas, de forma que as probabilidades correspondentes não podem ser estimadas com robustez. Há dificuldades, sobretudo para os casos de frequências zero, em que é difícil decidir se o trigrama correspondente está sintaticamente incorreto (caso em que a probabilidade é zero) ou é apenas raro (caso em que a probabilidade tem um pequeno valor positivo).

Outro fator importante é que um etiquetador robusto deve ser capaz de lidar com erros gramaticais nos dados de entrada. Esses erros podem levar à probabilidade zero para toda a sentença independente da sequência de etiquetas, o que deve ser evitado. Por esse motivo, a fórmula acima é quase sempre mudada, por exemplo, com um método de *smoothing* que substitui as probabilidades zero por um pequeno valor, renormalizando-as, de forma que sua soma seja igual a 1.

A vantagem do TreeTagger em relação aos outros etiquetadores baseados em *N*-gramas é que este substitui o método usado para estimar as probabilidades de transição (MLE) por uma árvore de decisão binária. Esta substituição melhora o desempenho nos casos em que o conjunto de treinamento é pequeno. A Figura 2.3 mostra um exemplo (não realístico, apenas para ilustração) dessa árvore. A probabilidade de um dado trigrama é determinada seguindo-se o caminho correspondente pela árvore até que uma folha seja alcançada. Por exemplo, para encontrar a probabilidade de um substantivo, o qual é precedido por um determinativo e um adjetivo[3] $p(NN|DET,ADJ)$, faz-se o teste no nó raiz. Se a etiqueta da primeira palavra do trigrama for *ADJ*, segue-se o caminho *sim*. O próximo teste (tag_{-2} = *DET*) é verdadeiro e conduz ao nó folha. Então, é só procurar pela probabilidade da etiqueta *NN* na tabela que está anexada a esse nó.

FIGURA 2.3 - Exemplo de árvore de decisão



Fonte: Schmid, 1994a, p. 46.

Além da árvore de decisão, as técnicas de realizar *smoothing* com classes de equivalência e de tratar palavras de início de sentença, que também fazem parte do algoritmo do TreeTagger, aumentam a acurácia desse etiquetador, uma vez que melhoram a estimativa da probabilidade léxica em *corpora* de pequenos tamanhos (SCHMID, 1995). No primeiro caso, parte-se do princípio que as palavras com o mesmo conjunto de etiquetas possíveis têm distribuições de probabilidades similares, o que facilita a estimativa de probabilidades para palavras raras, a qual passa a se basear em probabilidades de classes de equivalência. No segundo caso, o algoritmo busca por duas instâncias de uma mesma palavra: com inicial maiúscula no começo de uma sentença e com inicial minúscula no léxico de treinamento. Se ambas são encontradas, os vetores de probabilidades são ponderados pela frequência relativa de suas formas correspondentes e as frequências ponderadas são somadas. Dessa forma, por exemplo, a palavra “Vozes” na sentença “Vozes esbravejantes nos diziam ...”, terá maior probabilidade de ser etiquetada como nome (N) do que como nome próprio (NPROP), como é o caso em “Editora Vozes”.

[3] Exemplo referente à língua inglesa.

2.4.2.4.2 Memory-Based Learning

Memory-Based Learning (MBL) é um modelo descendente da abordagem de classificação *k-Nearest Neighbour*. Conforme descrito por Daelemans et al. (1996), o algoritmo inicia guardando um conjunto de exemplos de casos na memória. Cada caso consiste de uma palavra (ou representação léxica para a palavra) com o contexto anterior e posterior e a categoria correspondente para essa palavra nesse contexto. Uma nova sentença é etiquetada pela seleção, para cada palavra da sentença, do(s) caso(s) mais similares em memória e obtém a categoria da palavra desses "vizinhos mais próximos". A métrica de similaridade usada considera o número de casamentos de atributos entre os casos e calcula o peso da importância relativa de cada atributo por um fator de ganho de informação. Esse número mede a utilidade do atributo na predição da classificação correta.

2.4.2.4.3 Maximum Entropy

É um modelo que integra diversas fontes de informação heterogêneas para classificação. Os dados a serem classificados são descritos por um grande número de atributos complexos que utilizam conhecimento prévio sobre que tipo de informação é importante para a classificação tais como: prefixos e sufixos de palavras; se as palavras contêm números, hífens ou outros sinais, etc. Cada atributo corresponde a uma restrição sobre o modelo. O modelo de máxima entropia é aquele que prevalece sobre todos os modelos que satisfazem as restrições (MANNING; SCHUTZE, 1999). Esse modelo é melhor compreendido no trabalho de Ratnaparkhi (1996), que define o conjunto de possíveis contextos de palavras e etiquetas como um *history* h_i disponível nos dados de treinamento quando realiza a atribuição de uma etiqueta t_i para uma palavra w_i, por exemplo, $h_i = \{ w_i, w_{i+1}, w_{i+2}, w_{i-1}, w_{i-2}, t_{i-1}, t_{i-2}\}$. *Templates* de atributos são usados para gerar o espaço de atributos pelo exame de cada par (h_i, t_i). Um exemplo de um *template* de atributos é o seguinte: "se w_i é rara, X é sufixo de uma palavra w_i, $|X| \leq 4$ e a etiqueta $t_i = T$".

2.4.2.4.4 Inferência bidirecional

O método de inferência bidirecional é explorado por Toutanova et al. (2003) em um etiquetador concebido para a língua inglesa, que também explora conjuntos de atributos para a etiquetagem de sentenças. Esse método explora os contextos das etiquetas precedentes e posteriores à etiqueta corrente em uma representação de rede de dependência cíclica, com a adição de técnicas como o uso de atributos léxicos que incluem condicionamento conjunto em múltiplas palavras consecutivas, uso efetivo de *priors* em modelos log-linear condicionais e modelagens refinadas de atributos de palavras desconhecidas.

2.4.2.4.5 Conditional Random Field

O modelo de *conditional random field* (CRF) também lida com variados atributos para etiquetar as palavras. CRFs são modelos gráficos não direcionais usados para construir modelos probabilísticos para segmentar e etiquetar dados sequenciais (LAFFERTY, MCCALLUM, PEREIRA, 2001; SUTTON, ROHANIMANESH , MCCALLUM, 2004). São usados para calcular a probabilidade condicional dos valores em nós de saída quando foram dados determinados valores nos nós de entrada e conseguem representar relacionamentos complexos entre etiquetas, tais como dependências de longo alcance entre estas (SUTTON, ROHANIMANESH , MCCALLUM, 2004).

2.4.2.4.6 Diversos

Há ainda muitos outros modelos para a etiquetagem de sentenças dentre os quais, baseados em redes neurais (SCHMID, 1994b), baseados em *support vector machine* (SVM) (GIMÉNEZ; MÁRQUEZ, 2004) e em algoritmo genético (WILSON; HEYWOOD, 2005).

2.4.3 Abordagem Híbrida

A partir dessas abordagens, surgiu uma terceira abordagem chamada “híbrida” (BRILL, 1995), que envolve a combinação das abordagens baseada em regras e probabilística e se caracteriza pela etiquetagem de textos com regras aprendidas automaticamente de *corpora* anotados. Um etiquetador com essa abordagem pode ser estruturado de várias maneiras, por exemplo, com um estado inicial baseado em regras e um final probabilístico ou, ao contrário, com um estado inicial estocástico e um final baseado em regras.

2.4.3.1 Aprendizado baseado em transformação dirigida por erro

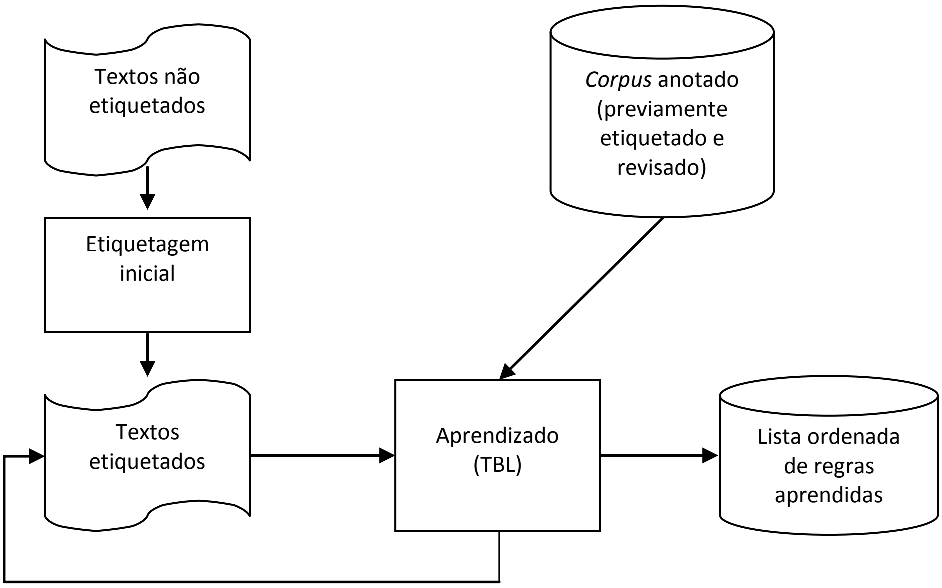
A abordagem híbrida é centrada no algoritmo de Aprendizado Baseado em Transformação Dirigida por Erro (TBL) desenvolvido por Brill (1994, 1995, 1997) e tem sido usada em diversas aplicações, por exemplo, em análise sintática e em etiquetagem morfossintática.

O processamento do TBL é realizado da seguinte maneira (BRILL, 1995):

1. **Etiquetagem inicial**: o texto não etiquetado passa por um processo de etiquetagem inicial, o qual pode ser realizado utilizando-se várias soluções, dentre as quais:
   - Uso de um etiquetador estocástico.
   - Atribuição às palavras de suas etiquetas mais frequentes, como indicado no *corpus* de treinamento.
   - Atribuição às palavras de uma determinada etiqueta (“substantivo”, por exemplo).
2. **Aprendizado**:
   - Após a etiquetagem inicial, o texto é comparado com o seu correspondente que se encontra corretamente etiquetado em um *corpus* de referência previamente anotado e revisado.
   - Uma lista ordenada de transformações (lista de regras) é aprendida e aplicada ao texto após a etiquetagem inicial para tentar aproximá-lo do *corpus* de referência.
   - Uma busca é aplicada sobre o texto que está sendo etiquetado para a derivação de uma lista de transformações (lista ordenada de regras aprendidas). A cada iteração de aprendizado, é encontrada a transformação que resulta na melhor pontuação de acordo com uma função objetivo pré-definida (por exemplo, o número total de erros). Essa transformação é adicionada à lista ordenada de transformações e o texto de treinamento é atualizado com essa transformação.
   - O aprendizado continua até que nenhuma outra transformação melhore o texto etiquetado.

A Figura 2.4 ilustra os passos do TBL.

FIGURA 2.4 - Aprendizado Baseado em Transformação Dirigida por Erro



Fonte: Baseada em Brill, 1995, p. 4.

Uma transformação possui dois componentes básicos: uma regra de substituição e o contexto ao qual se aplica esta regra.

Um exemplo de regra de substituição é:

Mudar a etiqueta de *verbo* (V) para *nome* (N).

Um exemplo de contexto dessa regra é:

A palavra anterior é um *artigo* (ART).

Cada transformação se baseia na instanciação de um metamodelo de regra denominado *template*. Brill (1995) descreve para aplicações de etiquetadores, duas versões de *templates* sensíveis ao contexto: não-lexicalizados e lexicalizados.

Os *templates* não-lexicalizados (Tabela 2.2) não fazem referência a palavras específicas. O foco são as estruturas contextuais (etiquetas).

TABELA 2.2 - *Templates* não-lexicalizados

	Mudar a etiqueta *a* para *b* quando:
1	A palavra anterior (posterior) foi etiquetada como *z*.
2	A palavra duas posições antes (depois) foi etiquetada como *z*.
3	Uma das duas palavras anteriores (posteriores) foi etiquetada como *z*.
4	Uma das três palavras anteriores (posteriores) foi etiquetada como *z*.
5	A palavra anterior foi etiquetada como *z* e a posterior como *w*.
6	A palavra anterior (posterior) foi etiquetada como *z* e a palavra duas posições antes (depois) foi etiquetada como *w*.

em que *a*, *b*, *z* e *w* são variáveis que representam as etiquetas do *corpus* utilizado.

Os *templates* lexicalizados (Tabela 2.3) buscam relações entre as palavras e as etiquetas, adicionando informação léxica ao modelo, isto é, fazendo uso de palavras e etiquetas do *corpus*.

TABELA 2.3 - *Templates* lexicalizados

	Mudar a etiqueta *a* para *b* quando:
1	A palavra anterior (posterior) é *w*.
2	A palavra duas posições antes (depois) é *w*.
3	Uma das duas palavras anteriores (posteriores) é *w*.
4	A palavra corrente é *w* e a palavra anterior (posterior) é *x*.
5	A palavra corrente é *w* e a palavra anterior (posterior) foi etiquetada como *z*.
6	A palavra corrente é *w*.
7	A palavra anterior (posterior) é *w* e a etiqueta anterior (posterior) é *t*.
8	A palavra corrente é *w*, a palavra anterior (posterior) é w_2 e a etiqueta anterior (posterior) é *t*.

em que *w* e *x* são variáveis que representam todas as palavras do *corpus* de treinamento e *z* é uma variável que representa todas as etiquetas desse *corpus*.

Para etiquetar palavras desconhecidas, o TBL atribui na etiquetagem inicial como etiquetas mais frequentes às palavras desconhecidas: nome próprio se a palavra possui inicial maiúscula e nome ou substantivo se a inicial é minúscula. Para encontrar as transformações, são utilizados os *templates* listados na Tabela 2.4.

TABELA 2.4 - *Templates* para palavras desconhecidas

	Mudar a etiqueta de uma palavra desconhecida de *X* para *Y* se:
1	Apagando prefixo (sufixo) *x*, |*x*| <=4, resulta em uma palavra (*x* é qualquer cadeia de comprimento 1 a 4).
2	Os primeiros (últimos) (1, 2, 3, 4) caracteres da palavra são *x*.
3	Adicionando a cadeia de caracteres *x* como prefixo (sufixo) resulta em uma palavra (|*x*| <=4).
4	A palavra *W* sempre aparece imediatamente à esquerda (direita) da palavra.
5	O caractere *Z* aparece na palavra.

Segundo Brill (1995), uma vantagem do TBL é que a lista de transformações é usada como um processador e não como um classificador, sendo adequada para o pós-processamento de qualquer sistema de etiquetagem. Dessa forma, as regras aprendidas podem melhorar sistemas de etiquetadores estocásticos maduros utilizados como entrada na etiquetagem inicial do TBL.

2.5 Processo de Classificação

A finalidade de conceber um sistema de classificação consiste em criar um modelo com base em dados conhecidos, que serve para analisar porque determinada classificação foi feita ou para classificar novos dados. Os classificadores utilizam aprendizado supervisionado para extração do modelo. Os dados de entrada para esse tipo de sistema consistem em um conjunto de valores de determinados atributos e uma classificação para esses valores. A análise de dados já classificados revela as características que levaram à classificação anterior. O modelo de classificação resultante pode, assim, ser usado para prever as classes de novos objetos apresentados ao sistema.

A classificação é realizada em dois passos:

1. **Construção de um modelo** que descreve conceitos ou classes de um conjunto de dados, pré-determinados pela análise de cada linha de uma tabela de banco de dados (BD) ou população de exemplos descrita por seus atributos.
   - Assume-se que cada linha do BD pertence a uma classe pré-definida, conforme foi determinado por um dos atributos chamado de *atributo classe*. Este conjunto de dados (exemplos ou objetos) é chamado de *conjunto de treinamento*, o qual pode ser selecionado randomicamente da população de exemplos.
   - Tipicamente, o modelo aprendido é representado na forma de regras de classificação, árvores de decisão ou fórmulas matemáticas ou lógicas.
2. **O modelo é usado para a classificação**:
   - É realizada a fase de teste, na qual é avaliada a precisão deste modelo.
   - Os resultados da fase de teste são comparados com o modelo (*Gold Standard*), para verificar a taxa de erros/acertos do classificador.

- A acurácia preditiva do modelo (ou classificador) é estimada. Se a acurácia é considerada aceitável, o modelo pode ser utilizado para prever situações (casos) não conhecidas.
- A acurácia se refere ao percentual de exemplos de um conjunto de testes que estão corretamente classificados pelo modelo.
- O modelo pode incorporar algumas anomalias particulares do conjunto de treinamento, que não estão presentes na maioria dos exemplos (de uma mesma classe) na população de exemplos, levando à classificação incorreta (apresentada em uma matriz de desordem ou matriz de confusão).

2.5.1 Métodos de Avaliação da Acurácia do Classificador

2.5.1.1 Cálculo da acurácia ou precisão

O desempenho de um classificador é medido em termos da taxa de erro. O classificador atribui uma classe para cada instância. Se a classificação for correta, é contada como sucesso; caso contrário, é um erro. A taxa de erros é a proporção de erros ocorridos sobre o conjunto total de instâncias que está sendo avaliado e mede o desempenho global do classificador (WITTEN; FRANK, 2005).

2.5.1.2 Métodos para divisão dos conjuntos de treinamento e teste

A taxa de erro calculada sobre um conjunto de treinamento não indica que o desempenho será o mesmo em um conjunto de teste (dados novos). É importante observar que para predizer o desempenho de um classificador em dados novos, deve ser verificada a sua taxa de erro em um conjunto de dados que não fez parte do treinamento do classificador (conjunto de teste). Além disso, ambos os dados de treinamento e teste devem ser exemplos representativos do problema delineado. O conjunto de teste não deve ser usado na criação do classificador (WITTEN; FRANK, 2005).

Para obter uma taxa de erro mais precisa, métodos para divisão dos conjuntos de treinamento e testes utilizados são empregados para avaliar o desempenho em dados futuros. Os métodos mais utilizados são:

- ***Holdout***: é um método de divisão de conjuntos mais simples, favorável para grandes volumes de dados, que separa, por exemplo, dois terços dos dados para treinamento e um terço para teste. Em etiquetagem morfossintática, são muito comuns as avaliações em que se separa 80% x 20% ou 90% x 10% dos dados para treinamento e teste, respectivamente. Em alguns casos se reserva uma parte dos dados para formar um conjunto de avaliação, por exemplo, 80% para treinamento, 10% para avaliação e 10% para teste.
- **Validação Cruzada**: é a repetição do método *holdout* para a estimativa da taxa de erro. Este método costuma ser utilizado quando a quantidade de dados é limitada. É também conhecido por *k-fold cross-validation,* em que o *k* é o número de partições mutuamente exclusivas. A taxa de erro é a média das taxas de erro calculadas para as diversas partições. Todos os exemplos são utilizados para treinar e testar. Por exemplo, se k = 10, o conjunto de dados é dividido em 10 partes randomicamente. Em cada uma das 10 iterações, cada parte será tomada para teste e as outras nove partes são usadas para treinamento. Cada uma das 10 iterações produz uma taxa de erro. Ao final, é calculada a média dessas 10 taxas de erro para encontrar a uma estimativa da taxa de erro global. Nos experimentos desta tese, este método é usado com k=20 para a avaliação de um *corpus* pequeno (cerca de 80.000 palavras) e k=10 para um corpus de tamanho maior (446.676 palavras).

2.6 Adaptação de Domínio

Os etiquetadores do estado da arte apresentam boa acurácia para o domínio ou gênero em que foram treinados, mas quando usados para etiquetar textos de domínios ou gêneros diferentes, apresentam baixa acurácia. Esse problema é típico em aplicações de aprendizado supervisionado, em que se tem uma quantidade suficiente de dados anotados para o domínio de treinamento (fonte), mas não se tem dados suficientes do novo domínio (alvo) (JIANG, 2008). Nesse contexto, os problemas mais comuns são as diferentes distribuições dos dados e a presença de palavras desconhecidas. Nesses casos, faz-se necessário realizar uma adaptação de domínio, de maneira que o etiquetador passe a apresentar bom desempenho tanto no domínio fonte, quanto no novo domínio alvo. Algumas pesquisas importantes em adaptação de domínio são as de Blitzer (2007), Jiang (2008), Huang e Yates (2010) e Umansky-Pesin et al. (2010). Em Domingues e Favero (2011), é apresentado um levantamento dos métodos usados no estado da arte para adaptação de domínio, no entanto ainda não há um método que seja o mais adequado para resolver o problema. No capítulo 7 desta tese, são apresentados experimentos no sentido de adaptar o etiquetador proposto para os domínios jornalístico e científico.

3 Revisão Bibliográfica

"Os estudos aperfeiçoam a natureza e são aperfeiçoados pela experiência"
(Francis Bacon)

Pesquisas em etiquetagem morfossintática para as mais diversas línguas têm sido desenvolvidas com o uso de abordagens e tecnologias variadas, em busca de resultados cada vez mais acuráveis. Neste capítulo, são listados alguns trabalhos citados na literatura e seu desempenho em termos de acurácia.

3.1 Etiquetadores X Abordagens X Acurácia

3.1.1 Abordagem Baseada em Regras

Ao longo dos anos, percentuais de acurácia de 99% já foram relatados em trabalhos sobre etiquetadores baseados em regras que utilizam *Constraint Grammar*, a saber:

- O etiquetador do analisador gramatical EngCG (*English Constraint Grammar*) (VOUTILAINEN, 1995b), reavaliado por Samuelsson e Voutilainen (1997), apresenta acurácia global de 99,26% para a língua inglesa. Comparada aos percentuais de acurácia de 95-97% alcançados pelos etiquetadores probabilísticos da época, essa acurácia sugere que a desambiguização de etiquetas é um problema sintático.
- O etiquetador do analisador gramatical PALAVRAS (BICK, 2000), para o português, pode alcançar acurácia em torno de 99% dependendo do tipo de texto e da granularidade do conjunto de etiquetas. Bick (2000) relata acurácia de 99% em experimentos, com a aplicação de milhares de regras gramaticais, em pequenos conjuntos de testes de 2.500, 3.140 e 4.800 palavras.

Apesar dos relatos de alta acurácia, etiquetadores dessa abordagem requerem um grande esforço dos especialistas na construção de regras. Mesmo que na atualidade a construção e teste de regras possam ser feitos com menor esforço, é sabido que etiquetadores probabilísticos com acurácia de 95-97% podem ser produzidos com um esforço bem menor (SAMUELSSON; VOUTILAINEN, 1997).

3.1.2 Abordagem Probabilística

Uma vez que os etiquetadores probabilísticos requerem menor esforço na sua construção, etiquetadores envolvendo o uso dos mais diversos métodos probabilísticos e que extraem informações de *corpora* têm motivado inúmeras pesquisas. No decorrer dos anos, tem-se:

- O etiquetador baseado em trigrama de Kempe (1993) e o etiquetador TreeTagger baseado em HMM e Árvore de Decisão de Schmid (1994a), ambos treinados em dados do *Wall Street Journal* (WSJ) do *corpus Penn-Treebank* da língua inglesa, com dois milhões de palavras usadas para treinamento e 100 mil palavras usadas para teste, obtiveram acurácia de 96,06% e 96,36%, respectivamente.
- Com esse mesmo *corpus*, o etiquetador MXPOST (RATNAPARKHI, 1996) obteve acurácia de 96,6% com o modelo de *Maximum Entropy*, porém com um número de palavras diferente: 962.687 palavras para treinamento, 192.826 para avaliação e 133.805 para teste. O MXPOST combina vantagens de outros etiquetadores, como a representação de atributos, a qual também é explorada pelo TBL (BRILL, 1995) e alguns etiquetadores baseados em árvore de decisão, e gera uma distribuição de probabilidade de etiquetas para cada palavra, similar às técnicas de árvore de decisão e HMM. Nessa ocasião, diversos etiquetadores alcançaram acurácia em torno de 96,5% e houve suposição de que isso fosse motivado por problemas de consistência no *corpus*.
- Ainda com o mesmo *corpus*, o etiquetador IGTree proposto por Daelemans et al. (1996) utilizou o modelo *Memory-based*, usou 2 milhões de palavras para treinamento e 200 mil para teste obtendo acurácia global de 96,4%; o etiquetador TnT de Brants (2000), baseado em trigramas, obteve acurácia de 96,7%.
- Para a língua alemã, Schmid (1995) realizou melhorias no TreeTagger e obteve acurácia de 97,5%. Neste caso, foi utilizado um *corpus* manualmente etiquetado com 20.000 palavras para treinamento e 5.000 palavras para teste, e um arquivo com 350.000 entradas léxicas (requerido pelo TreeTagger).
- Gimenez e Márquez (2004) usaram um etiquetador baseado em SVM que alcançou acurácia de 97,16% para o inglês.

- Alguns trabalhos mais recentes para a língua inglesa são os etiquetadores Stanford POS Tagger (TOUTANOVA et al., 2003) e Postagger (TSURUOKA; TSUJII, 2005), com experimentos usando 1.173.766 palavras do WSJ do *corpus Penn-Treebank III* e estratégia de avaliação que empregou 78% do *corpus* para treinamento, 11% para desenvolvimento e 11% para teste, combinada com novas abordagens probabilísticas de classificação sequencial bidirecional. O Stanford POS Tagger alcançou um dos melhores resultados para o inglês com acurácia de 97.24%, enquanto que o Postagger alcançou acurácia de 97,15%.
- Para o árabe, foi proposta uma aplicação do modelo *Memory-based*, que utilizou 166.068 palavras do *corpus Arabic Treebank* usando um método *10-fold-cross-validation*. A acurácia obtida foi de 91,5% (VAN DEN BOSCH; MARSI; SOUDI, 2007).
- Para a língua portuguesa, o trabalho de Aires (2000a) realizou experimentos com o etiquetador MXPOST. A acurácia média foi de 97% com um *corpus* de treinamento de cerca de 100.000 palavras e uma estratégia de validação cruzada de 10 subconjuntos para testes. Outro trabalho, disponibilizado no *site* do projeto Lácio-Web (2007), fez experimentos com os etiquetadores probabilísticos TreeTagger e MXPOST e com o etiquetador híbrido Brill TBL, em uma estratégia de 80% para treinamento e 20% para testes do conjunto de 1.221.468 palavras do *corpus* Mac-Morpho. Os resultados apresentaram acurácia que vai de 82,36% a 96,98%, variando conforme o etiquetador utilizado e o caderno jornalístico dos parágrafos do *corpus* etiquetado.
- Ekbal, Haque e Bandyopadhyay (2007) propuseram um etiquetador para o bengali baseado em um modelo de CRF que alcançou acurácia de 90,3%.

3.1.3 Abordagem Híbrida

A avaliação do etiquetador híbrido de Brill (1995), que usa o método TBL, foi realizada a partir do WSJ do *corpus Penn-Treebank* com 1.100.000 palavras, das quais 950.000 foram usadas para treinamento e 150.000 para teste. A versão com regras lexicalizadas alcançou acurácia global de 96,6% e acurácia em palavras desconhecidas de 82,2%. A versão sem regras lexicalizadas alcançou 96,3% de acurácia global e 82% de acurácia em palavras desconhecidas. A versão apenas com palavras conhecidas alcançou 97% de acurácia. O método TBL tornou-se fonte de pesquisa para muitos trabalhos recentes. Essa abordagem foi adotada, por exemplo:

- Para o hebreu (ITAI; SEGAL, 2003) com acurácia de 96,2%.
- Para a língua húngara, em uma abordagem que combina um modelo baseado em trigramas com *Maximum Entropy*, utiliza o *Szeged Corpus* e obteve acurácia de 98,17% (HALÁCSY et al., 2006).
- Há trabalhos que atingem um percentual menor de acurácia por limitações de recursos linguísticos ou outras. Para o khmer, língua oficial do Camboja, que possui características linguísticas bem diferentes do inglês, por exemplo, Nou e Kameyama (2007) propõem uma adaptação da abordagem híbrida baseada em TBL (BRILL, 1995) com um *corpus* de 41.061 palavras (78% para treinamento e 22% para teste) e relatam 91,96% de acurácia.
- Outro tipo de etiquetador proposto é o que utiliza algoritmo genético baseado em TBL (WILSON; HEYWOOD, 2005). Em experimentos com dados do WSJ, a melhor acurácia alcançada foi de 89,8%, considerada baixa em comparação com a obtida pelo TBL (BRILL, 1995) sobre o mesmo *corpus* que é de 97%.
- Um etiquetador híbrido para o português brasileiro é o inserido na ferramenta CoGrOO (KINOSHITA, SALVADOR, MENEZES, 2006; KINOSHITA et al., 2007) com acurácia de 95% em média.
- Há também o etiquetador híbrido para o português brasileiro de Santos, Milidiú e Rentería (2008), que foi treinado com 1 milhão de palavras do *corpus* Mac-Morpho e testado em 200.000 palavras desse mesmo *corpus*, usando um conjunto de 22 etiquetas, e alcançou acurácia de 96,75%. Em experimento com o *corpus* histórico do Português Tycho Brahe, treinado com 775.602 palavras e testado em 259.991 palavras, com um conjunto de 383 etiquetas, foi alcançada acurácia de 96,64%.

3.2 Etiquetadores X Conjunto de Etiquetas X Acurácia

Outro assunto relevante na pesquisa sobre etiquetagem morfossintática diz respeito ao conjunto de etiquetas.

- Tufis (2000) utiliza o QTag, etiquetador baseado em HMM (MASON e TUFIS, 1997), para a etiquetagem de textos da língua romana. Nesse estudo, é explorado o uso de um conjunto de 614 etiquetas chamadas descritores morfossintáticos (*morpho-syntactic descriptions* – MSD), que incluem atributos como gênero e número, dentre muitos outros. Para lidar com esse grande número de etiquetas, foi projetado um conjunto reduzido de 92 etiquetas morfossintáticas e mais 10 etiquetas de pontuação, utilizado em uma etapa intermediária no processo de etiquetagem. O *corpus* de treinamento foi anotado manualmente com os MSDs e depois teve essas etiquetas mapeadas automaticamente para o conjunto reduzido de etiquetas. Em uma primeira etapa, um novo texto é etiquetado com um número reduzido de etiquetas; em uma segunda etapa, essas etiquetas são substituídas pelas etiquetas mais informativas. Esse processo é chamado de **etiquetagem escalonada** (*tiered tagging*). Para o romano, esse processo gerou bem poucos erros no mapeamento de etiquetas, que foram resolvidos com 14 regras contextuais. Em Tufis e Mason (1998), experimentos com essa abordagem, com um número menor de etiquetas, foram realizados em conjuntos de testes de *corpus* diferentes, a acurácia variou de 96,22% a 98,39%.

3.3 Outras Abordagens

3.3.1 Redes Neurais

Schmid (1994b) também propôs a etiquetagem com o uso de redes neurais em experimentos semelhantes aos apresentados em Schmid (1994a) e obteve acurácia de 96,22%.

3.3.2 *Latent Analogy*

O método proposto por Bellegarda (2010) alcançou acurácia de 96,5% no corpus *Penn Treebank* e 96,4% em um *corpus* de síntese de fala.

3.3.3 Combinação de Etiquetadores

Além dos enfoques relatados acima, há trabalhos que propõem combinação de etiquetadores:

- No trabalho de Brill e Wu (1998), experimentos constataram que etiquetadores erram em casos diferentes e que essa complementaridade pode ser aproveitada para elevar a acurácia. Nesses experimentos, várias combinações foram testadas e o melhor resultado foi de 97,2%.
- Aires (2000b) utilizou abordagem semelhante para o português brasileiro, mas a melhor acurácia obtida foi de 90,91%. Nesse caso, a baixa acurácia foi atribuída mais a problemas em relação ao *corpus* do que propriamente com os métodos de etiquetagem e de combinação de etiquetadores.
- Em Tufis (2000), além da etiquetagem escalonada, foi proposto outro método para combinação de etiquetadores. Nesse caso, em vez de se utilizar vários etiquetadores e os mesmos dados de treinamento, foi utilizado um etiquetador e dados de treinamentos de vários *corpora*. A acurácia média da etiquetagem dos vários subconjuntos de testes usados ficou em 98,5%. Ao considerar somente a informação morfossintática das etiquetas (desprezando o restante dos atributos léxicos), a acurácia ultrapassou 99%.

3.4 Considerações

Os resultados dos trabalhos citados acima nem sempre permitem uma comparação, para que se possa afirmar com exatidão que um etiquetador é melhor do que outro, uma vez que, além dos algoritmos e abordagens para classificação, outras variáveis estão envolvidas, por exemplo: o *corpus* e suas características (tamanho, qualidade, particularidades de cada língua, gêneros textuais, conjunto de etiquetas) e os métodos de avaliação (divisão dos conjuntos de treinamento e testes, números de palavras em cada conjunto). No entanto, observa-se que, na maioria dos casos, abordagens híbridas exigem um menor esforço na geração de regras e funcionam melhor que a aplicação de uma só abordagem.

4 Descrição de Abordagem de um Etiquetador de Alta Acurácia para o Português Brasileiro

"Se você quer os acertos, esteja preparado para os erros."
(Carl Yastrzemski)

Neste capítulo, são apresentados os recursos experimentais usados na pesquisa, a metodologia para o desenvolvimento do etiquetador, a arquitetura do modelo de etiquetagem de alta acurácia proposto e a descrição dos passos do processo de etiquetagem aplicado.

Esta proposta de tese surgiu das necessidades de um projeto de uma plataforma de ensino virtual multiparadigmática (HARB et al., 2003), que trabalha com textos de gêneros acadêmicos, cujo objetivo é fazer a correção de respostas de questões discursivas avaliadas automaticamente por algoritmos de aprendizado de máquina usando inteligência artificial. Os resultados subsidiarão o desenvolvimento de um novo conjunto de ferramentas computacionais para PLN, que deverá incluir etiquetadores morfossintáticos, analisadores sintáticos e corretores gramaticais e poderá ficar disponível não só para esse projeto, mas também para atender às necessidades de outros projetos.

4.1 Recursos Experimentais

Recursos de etiquetadores e *corpora* foram avaliados, experimentalmente, para subsidiar a construção de um modelo de etiquetador morfossintático de alta acurácia. Esses recursos são de uso livre e se encontram disponíveis para *download* na *Web*.

Com a premissa de que os etiquetadores atuais, construídos com base em métodos de aprendizagem de máquina variados e/ou a partir da aplicação de regras codificadas manualmente, parecem convergir quanto ao desempenho em termos de acurácia, buscou-se aumentar essa acurácia pelo aproveitamento de recursos de etiquetadores existentes combinados em um modelo híbrido e pela minimização de problemas inerentes ao processo de etiquetagem.

Os **etiquetadores** avaliados nesta pesquisa foram:

- Etiquetadores probabilísticos:

  - QTag[4] e TreeTagger[5]
- Analisador gramatical e etiquetador *online* baseado em regras:
  - PALAVRAS, no ambiente *Visual Interactive Syntax Learning* (VISL)[6]
- Sistema de extração de regras automático:
  - μ-TBL[7]

Para realizar treinamentos e testes com os etiquetadores, foram usados os ***corpora*** do português brasileiro:

- *Corpus* CETENFolha[8]
- Bosque CF 7.4 do *Corpus* CETENFolha[9]
- *Corpus* Mac-Morpho[10]
- *Corpus* Selva Científica[11]

4.1.1 Etiquetadores

4.1.1.1 Etiquetadores probabilísticos

Dois etiquetadores probabilísticos, QTag e TreeTagger, foram escolhidos para os experimentos deste estudo por utilizarem métodos de etiquetagem diferentes, por apresentarem facilidade de uso e por serem multilíngues, o que permitiu a adaptação para o português. Ambos trabalham com fontes de informação extraídas de um *corpus* de treinamento pré-etiquetado. Após vários experimentos, o etiquetador TreeTagger foi escolhido para subsidiar o modelo proposto por ter apresentado melhor desempenho no processo de etiquetagem do português brasileiro, confirmando o que diz a literatura sobre o fato de que esse etiquetador tem bom desempenho quando os *corpora* de treinamento são de tamanho pequeno.

[4] http://www.softpedia.com/get/System/File-Management/OM-QTag.shtml
[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
[6] http://visl.sdu.dk/visl/pt/parsing/automatic/trees.php
[7] http://www.ling.gu.se/~lager/mutbl.html
[8] http://www.linguateca.pt/Floresta/
[9] http://www.linguateca.pt/Floresta/
[10] http://www.nilc.icmc.usp.br/lacioweb/corpora.htm
[11] http://www.linguateca.pt/Floresta/

4.1.1.1.1 QTag

O QTag (MASON; TUFIS, 1997) se baseia na técnica de HMM de segunda ordem (Capítulo 2 – Seção 2.4.2.3). Esse etiquetador combina duas fontes de informação: um dicionário de palavras com suas etiquetas possíveis e uma matriz de sequências de etiquetas, ambas com suas probabilidades anotadas. Caso as palavras não sejam encontradas no dicionário, há um módulo para a etiquetagem de palavras desconhecidas que analisa os sufixos das palavras (as três últimas letras das palavras) (TUFIS; MASON, 1998). O *software* possui dois elementos principais: o etiquetador em si, que realiza um processamento genérico, e as instruções de etiquetagem específicas de cada idioma que são reunidas em um arquivo chamado *resource file* (arquivo de recursos léxicos) ao ser realizado o treinamento do etiquetador. Para a criação desse arquivo de recursos, o QTag extrai informações de um *corpus* de treinamento pré-etiquetado, do qual obtém um conjunto de etiquetas e um conjunto de probabilidades de transição entre etiquetas, codificados de modo que o etiquetador possa lê-los (MASON, 2006; BERBER SARDINHA, 2004).

4.1.1.1.2 TreeTagger

O TreeTagger (SCHMID, 1994a) se baseia na combinação dos métodos HMM de segunda ordem e Árvore de Decisão (Capítulo 2 – Seção 2.4.2.4). Para etiquetar as palavras, o TreeTagger pesquisa um léxico que contém as probabilidades *a priori* das etiquetas ocorrerem para cada palavra. O léxico possui três partes: o léxico *fullform*, o léxico de sufixos e a entrada *default*. A busca de uma palavra inicia no léxico *fullform*. Se a palavra é encontrada, é retornado o vetor de probabilidade de etiquetagem correspondente. Caso contrário, as letras maiúsculas da palavra são transformadas para minúsculas e a busca é refeita. Se houver falha, novamente, a próxima busca é realizada no léxico de sufixos, o qual é organizado em forma de árvore. Cada nó da árvore, com exceção do nó raiz, é rotulado com um caractere. Nas folhas, são anexados vetores de probabilidades de etiquetas. Durante uma busca, a árvore de sufixos é pesquisada a partir do nó raiz. Em cada etapa, é seguido o caminho que estiver etiquetado com o próximo caractere a partir do final do sufixo da palavra. O léxico de sufixos é construído automaticamente a partir do *corpus* de treinamento. No trabalho de Schmid (1994a), por exemplo, a árvore de sufixos foi construída com sufixos de comprimento 5 de todas as palavras que foram anotadas como classe aberta e as frequências de suas etiquetas foram contadas para todos os sufixos e armazenadas nos nós correspondentes. Se ainda houver falha na busca, é retornada a entrada *default* do léxico.

Em Schmid (1995), o TreeTagger recebeu melhorias para se adequar à língua alemã, dentre elas a inclusão de: técnica de *smoothing* com o uso de classes equivalentes; léxico de prefixos; módulo de recuperação de informação dos *corpora* eletrônicos que utiliza aprendizado não-supervisionado; módulo identificador de palavras que iniciam sentenças e a simplificação da fórmula que faz a etiquetagem de palavras.

4.1.1.2 Analisador gramatical e etiquetador *online* baseado em regras PALAVRAS

O PALAVRAS (BICK, 2000) é um analisador gramatical baseado em regras com CGs (Capítulo 2 – Seção 2.4.1) que contém um etiquetador e se encontra disponibilizado no ambiente *online* VISL de análise automática do português. Esse analisador permite etiquetar sentenças em várias opções, por exemplo, apenas com a atribuição de etiquetas morfológicas ou fazendo a etiquetagem e a análise gramatical completa das sentenças. Os *corpora* escolhidos para esta pesquisa foram etiquetados, automaticamente, pelo PALAVRAS e revisados, manualmente, por especialistas que removeram boa parte dos ruídos resultantes do processo automático. Os resultados obtidos da etiquetagem com esse analisador, sem revisão de ruídos por especialistas, serão usados como um parâmetro de comparação para se avaliar os resultados de experimentos desta pesquisa.

4.1.1.3 Sistema de extração de regras automático μ-TBL

Como método para a extração automática de regras, foi escolhido o sistema μ-TBL (LAGER, 1999, 2007) por conter recursos que facilitam a análise e compreensão das regras extraídas, por exemplo, a geração de um arquivo no formato HTML que permite verificar os erros do processo de etiquetagem, suas estatísticas e identificar falsas regras.

O μ-TBL utiliza os recursos da linguagem de programação Prolog para realizar uma forma generalizada de aprendizado baseado em transformação (TBL), método introduzido por Eric Brill (BRILL, 1995) (Capítulo 2 – Seção 2.4.3). O sistema suporta um formalismo composicional de regras/*template* e três algoritmos (*brill*, *simple* e *lazy*) que podem ser adaptados para diferentes tarefas de aprendizagem. Um exemplo de regra de substituição extraída pelo μ-TBL é mostrado a seguir:

```
tag:adv>adj <- wd:apenas@[0] & tag:pron-det@[-1,-2].
```

A regra diz: "*substitua a etiqueta advérbio (adv) pela etiqueta adjetivo (adj), se a palavra corrente for "apenas" e se a etiqueta da palavra que a antecede ou a etiqueta da palavra anterior a que a antecede for um pronome determinativo (pron-det)*". Essas regras são instâncias de *templates*, os quais contêm variáveis no lugar de valores dos atributos. Por exemplo, o *template* abaixo diz: "*substitua a etiqueta A pela etiqueta B, se a palavra corrente for C e se a etiqueta da palavra que a antecede ou a etiqueta da palavra anterior a que a antecede for D*":

```
tag:A>B <- wd:C@[0] & tag:D@[-1,-2].
```

O µ-TBL disponibiliza, dentre outros, um conjunto de 26 *templates* de Brill, ilustrados no Capítulo 2, Seção 2.4.3, Tabelas 2.2 e 2.3.

A ferramenta possui dois métodos para a avaliação da qualidade das regras: o *score* e a acurácia. O *score* de uma regra é dado pelo número de instâncias positivas menos o número de instâncias negativas, enquanto que a acurácia de uma regra é dada pelo número de instâncias positivas dividido pela soma do número total de instâncias da regra. A instanciação de uma regra é positiva se o consequente e o antecedente da regra são verdadeiros e é negativa se o consequente for falso e o antecedente for verdadeiro. O sistema permite ao usuário configurar parâmetros da lista de regras que será aprendida, por exemplo: o algoritmo, o conjunto de *templates* e os limites mínimos de *score* e acurácia desejados (LAGER, 2007).

4.1.2 Corpora

Os *corpora* eletrônicos usados como conjuntos de treinamento e teste dos etiquetadores foram escolhidos por conterem textos jornalísticos e científicos do português brasileiro e por estarem afinados com os objetivos desta pesquisa.

4.1.2.1 *Corpus* CETENFolha

O *Corpus* CETENFolha (*Corpus de Extractos de Textos Electrónicos* NILC/Folha de São Paulo) possui cerca de 28 milhões de palavras em português brasileiro e pode ser consultado no projeto Floresta Sintá(c)tica (LINGUATECA, 2007). O CETENFolha contém textos do jornal Folha de São Paulo que foram etiquetados automaticamente pelo analisador gramatical PALAVRAS (BICK, 2000) e se apresentam no formato de árvores deitadas (ADs) conforme se exemplifica na Figura 4.1. As características do formato, descrição do conjunto de etiquetas e as opções de análise gramatical do CETENFolha se encontram em Afonso (2006).

FIGURA 4.1 - Textos do *Corpus* CETENFolha em formato de ADs

```
<ext id=1 cad="Opinião" sec="opi" sem="94a">
<t>
<s>
SOURCE: CETENFolha n=2 cad="TV Folha" sec="clt-soc"
sem="94a"
CF2-5 Manchete estréia novo jornalístico
A1
STA:fcl
=SUBJ:prop('Manchete' F S)    Manchete
=P:v-fin('estrear' PR 3S IND) estréia
=ACC:np
==>N:adj('novo' M S)    novo
==H:n('jornalístico' M S)     jornalístico
</s>
</t>
```

Fonte: LINGUATECA, 2007.

4.1.2.2 Bosque CF 7.4 do *Corpus* CETENFolha

O Bosque CETENFolha (CF) 7.4 é a primeira parte do CETENFolha revista por analistas humanos. É considerado de tamanho pequeno, já que possui um total de 80.078 palavras, das quais se tem 17.873 palavras diferentes (entradas léxicas). Em experimentos iniciais deste estudo realizados com um subconjunto de textos contendo quatro milhões de palavras do CETENFolha (etiquetados automaticamente, mas não revistos por analistas humanos), as taxas de acertos ficaram bem abaixo de 90%, fato que levou à escolha do Bosque para experimentos desta pesquisa, por se acreditar que a baixa taxa de acertos obtida previamente pode ter sido causada pelo número elevado de ruídos no *Corpus* CETENFolha ainda não revisto. O conjunto de etiquetas do Bosque CF 7.4 é descrito por Afonso (2006) e contém 18 etiquetas mostradas na Tabela 4.1.

TABELA 4.1 - Conjunto de etiquetas do Bosque CF 7.4

Conjunto de Etiquetas		
adj (adjetivo)	n (nome ou substantivo)	prp (preposição)
adv (advérbio)	num (numeral)	v-fin (verbo finito)
art (artigo)	pron-det (pronome determinativo)	v-ger (verbo no gerúndio)
conj-c (conjunção coordenativa)	pron-indp (pronome independente)	v-inf (verbo no infinitivo)
conj-s (conjunção subordinativa)	pron-pers (pronome pessoal)	v-pcp (verbo no particípio)
in (interjeição)	prop (nome próprio)	pt (atribuída a todos os sinais de pontuação)

4.1.2.3 *Corpus* Mac-Morpho

O Mac-Morpho é um *corpus* fechado e anotado, que foi primeiramente etiquetado pelo analisador gramatical PALAVRAS (BICK, 2000) e, posteriormente, teve suas etiquetas mapeadas para o conjunto de etiquetas do Projeto Lácio-Web (LÁCIO-WEB, 2007). Foi revisado manualmente quanto à anotação morfossintática. É constituído de textos jornalísticos da Folha de São Paulo do ano de 1984, dos cadernos Esporte (ES), Dinheiro (DI), Ciência (FC), Agronomia (AG), Informática (IF), Ilustrada (cultura e artes) (IL), Mais! (suplemento de cultura e artes dominical) (MA), Mundo (internacional) (MU), Brasil (BR) e Cotidiano (CO). A versão disponível contém 1.221.538 palavras, das quais se tem 65.347 palavras diferentes; está em um formato adequado para o treinamento de etiquetadores, no qual cada linha apresenta uma palavra e sua etiqueta, conforme se vê na Figura 4.2.

FIGURA 4.2 - Textos do *Corpus* Mac-Morpho

```
Estima_V|+
se_PROPESS
que_KS
existam_V
cerca_PREP
de_PREP
mil_NUM
pandas_N
em_PREP
liberdade_N
._.
```

Fonte: LÁCIO-WEB, 2007.

O conjunto de etiquetas do *Corpus* Mac-Morpho é descrito por Aluísio et al. (2003) e contém 78 etiquetas: 22 regulares, 18 sinais de pontuação e 38 regulares combinadas com complementares. A Tabela 4.2 apresenta esse conjunto de etiquetas.

TABELA 4.2 - Conjunto de etiquetas do *Corpus* Mac-Morpho

Etiquetas Regulares		**Etiquetas Complementares**
ADJ (adjetivo)	PCP (particípio)	|EST (estrangeirismo)
ADV (advérbio)	PDEN (palavra denotativa)	|AP (apostos)
ADV-KS (advérbio conectivo subordinativo)	PREP (preposição)	|+ (contrações/ênclises)
ADV-KS-REL (advérbio relativo subordinativo)	PROADJ (pronome adjetivo)	|! (mesóclises)
ART (artigo)	PRO-KS (pronome conectivo subordinativo)	|[(início)
KC (conjunção coordenativa)	PROPESS (pronome pessoal)	|… (meio)
KS (conjunção subordinativa)	PRO-KS-REL (pronome relativo conectivo subordinativo)	|] (e fim de composições descontínuas)
IN (interjeição)	PROSUB (pronome substantivo)	|TEL (número de telephone)
N (nome)	V (verbo)	|DAT (data)
NPROP (nome próprio)	VAUX (verbo auxiliar)	|HOR (hora)
NUM (numeral)	CUR (símbolo de moeda corrente).	|DAD (dados).

4.1.2.4 *Corpus* Selva Científica

O Selva Científica é um subcorpus do *Corpus* Selva (LINGUATECA, 2009) que contém 141.361 palavras em textos de relatórios do Banco Central do Brasil e do Banco Central Europeu e capítulos de teses e artigos da Wikipedia coletados em setembro de 2008 de tópicos relacionados à ciência. O Selva foi anotado, automaticamente, pelo analisador gramatical PALAVRAS (BICK, 2000) e é um *corpus* parcialmente revisto por especialistas. Neste estudo, para facilitar a métrica da acurácia do etiquetador, as etiquetas de algumas palavras foram ajustadas para ficarem iguais às do Bosque CETENFolha (LINGUATECA, 2007), uma vez que os conjuntos de etiquetas dos dois *corpora* são similares. Adicionalmente, o *Corpus* Selva possui uma etiqueta "*n-adj*" que é atribuída às palavras que podem ser tanto um nome quanto um adjetivo.

4.2 Metodologia

Com base nas premissas apresentadas na Introdução (Capítulo 1 – Seção 1.4), a pesquisa foi conduzida por intermédio da realização de uma série de experimentos que envolveram as atividades de:

- realizar experimentos com etiquetadores em textos do português brasileiro nas três abordagens: baseada em regras, probabilística e híbrida;

- realizar etiquetagem em textos do português brasileiro para avaliar ferramentas e *corpora* eletrônicos existentes;
- construir ferramentas de apoio para viabilizar a realização da etiquetagem;
- empregar métodos de avaliação de acurácia;
- realizar experimentos e analisar resultados segundo os fatores:
  - método de etiquetagem híbrido: método probabilístico combinado com método baseado em regras, com duas formas de extração dos conjuntos de regras: manual e automática;
  - *corpus* de treinamento: tamanho, qualidade, conjunto de etiquetas, e gênero textual;
  - palavras desconhecidas nos textos a serem etiquetados.
- propor e avaliar estratégias para melhorar a acurácia do etiquetador;
- analisar erros das etapas probabilísticas e baseada em regras extraídas automaticamente, após a aplicação de cada estratégia, para codificar novas regras manuais e fazer ajustes no etiquetador;
- realizar experimentos e avaliar se houve melhora no desempenho do etiquetador, até que haja o mínimo de erros possível;
- definir a arquitetura do novo etiquetador.

Existem duas aplicações que devem ser consideradas com relação ao processo de etiquetagem automática:

- a aplicação para treinamento e avaliação do sistema de etiquetagem, a qual é apresentada neste capítulo;
- a aplicação em dados reais, a qual já possui as informações obtidas do treinamento e um conjunto de regras embutidos no etiquetador, que apenas etiqueta novos textos, caso em que não há um modelo para medir a acurácia de forma automática.

4.3 Arquitetura do Etiquetador

Nesta tese, desenvolveu-se um etiquetador baseado no modelo TBL, o qual, originalmente, possui duas grandes etapas: probabilística e baseada em regras aprendidas automaticamente. Para realizar a etapa inicial ou etapa probabilística do modelo híbrido de etiquetador proposto, é usado o etiquetador TreeTagger, que além de um arquivo com o *corpus* de treinamento e um arquivo de classes abertas, usa no treinamento um arquivo léxico contendo as palavras e suas possíveis etiquetas presentes no *corpus* de treinamento. Na etapa baseada em regras, é usado o sistema µ-TBL para extrair regras, automaticamente, e aplicá-las para a correção dos erros da etapa inicial. A essa estrutura básica, são adicionadas: 1) entradas léxicas de outros *corpora* que não o de treinamento no arquivo léxico do TreeTagger, 2) a consulta a um léxico de nomes próprios para correção de etiquetas dessa categoria de palavras e 3) a aplicação de regras codificadas, manualmente, para correção de erros (intermediárias e de pós-correção). A seguir, serão apresentadas as estruturas básicas e adicionais do etiquetador proposto.

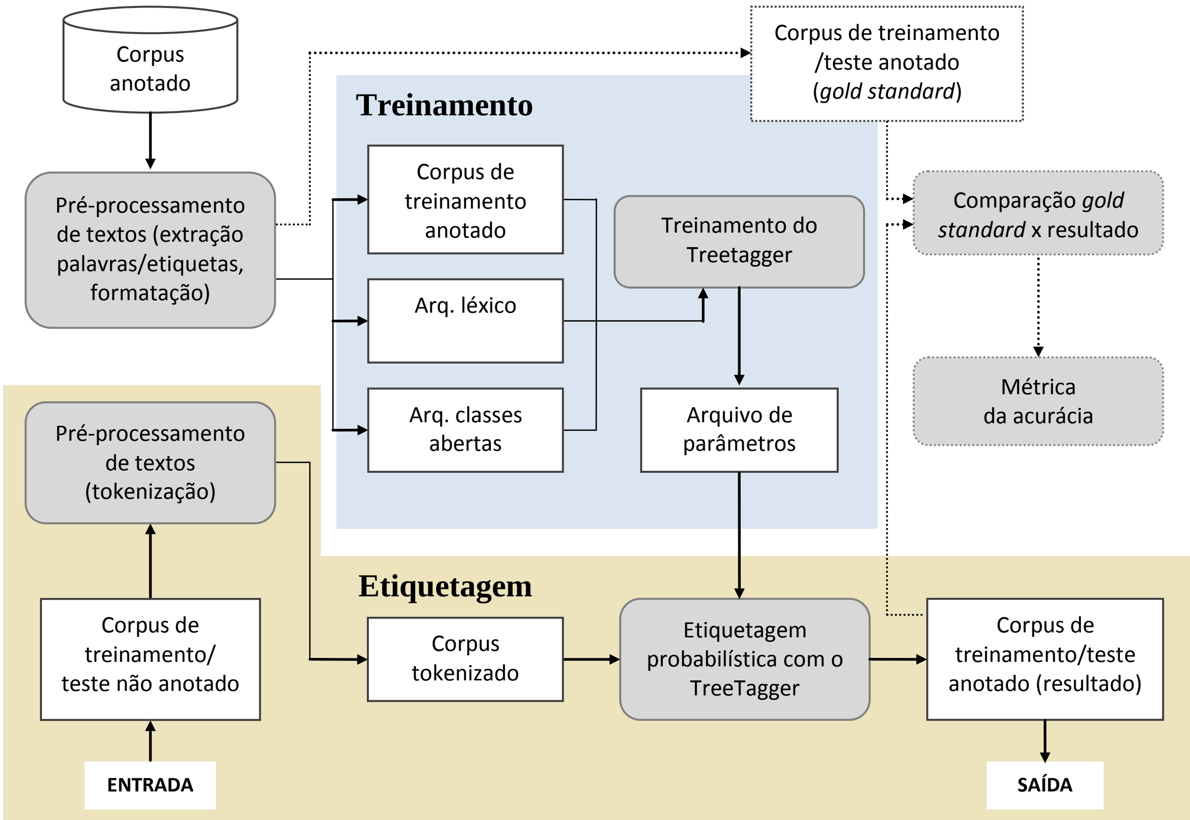
4.3.1 Etapa Probabilística

Na etapa probabilística, o etiquetador apresenta os módulos listados abaixo, conforme se ilustra na Figura 4.3:

- Treinamento:
  1) Pré-processamento de textos.
  2) Treinamento do TreeTagger.
- Etiquetagem:
  1) Pré-processamento de textos.
  2) Etiquetagem probabilística com o TreeTagger.
  3) Comparação de textos *gold standard* x resultado.
  4) Métrica da acurácia.

Os subitens 3 e 4 do item de Etiquetagem não fazem parte do processo de etiquetagem em si, são módulos usados para o processo de avaliação do etiquetador. Por esse motivo, serão representados com linhas pontilhadas nas figuras desta seção.

FIGURA 4.3 - Treinamento e etiquetagem probabilística com o TreeTagger

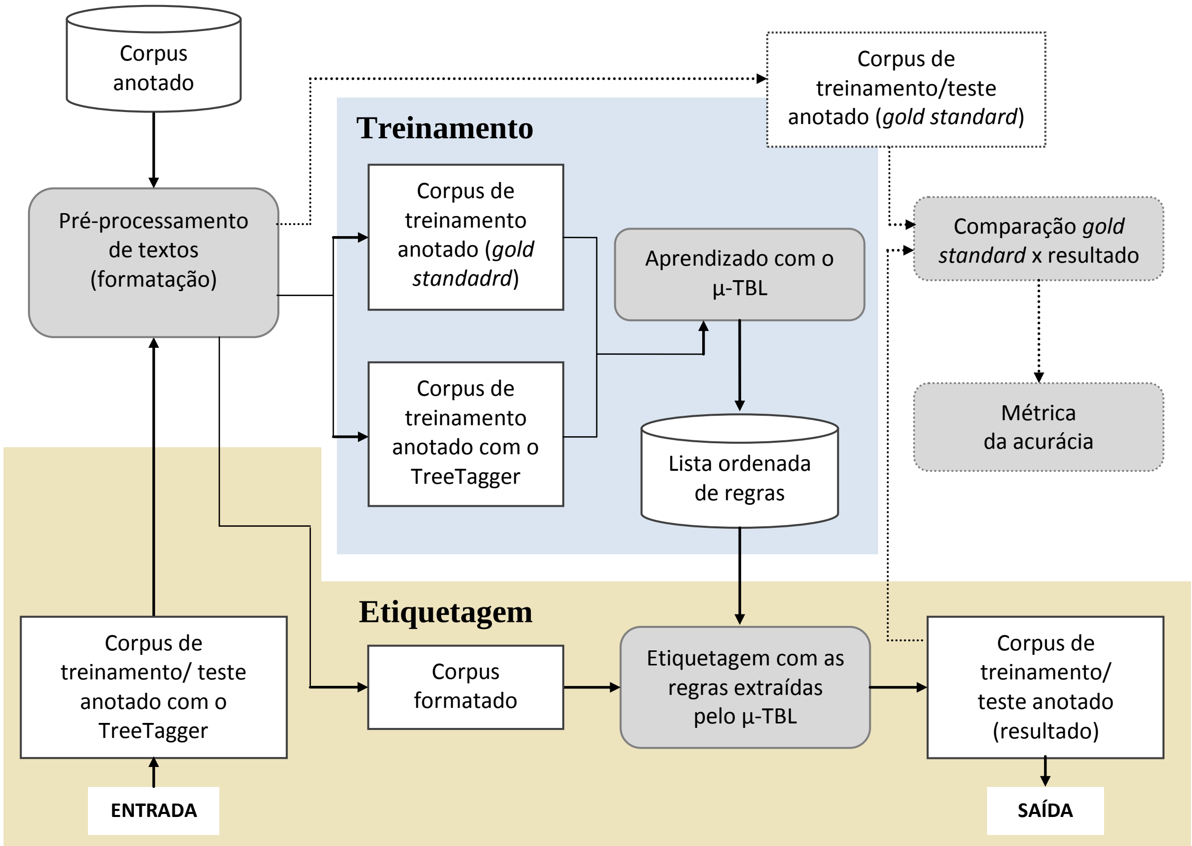


4.3.2 Etapa Baseada em Regras Aprendidas Automaticamente

Na etapa baseada em regras aprendidas automaticamente, o etiquetador apresenta os módulos listados abaixo, conforme se ilustra na Figura 4.4:

- Treinamento:
  1) Pré-processamento de textos.
  2) Aprendizado com o µ-TBL.
- Etiquetagem:
  1) Pré-processamento de textos.
  2) Etiquetagem com a lista ordenada de regras extraídas pelo µ-TBL.
  3) Comparação de textos *gold standard* x resultado.
  4) Métrica da acurácia.

FIGURA 4.4 - Treinamento e etiquetagem baseada em regras com o µ-TBL

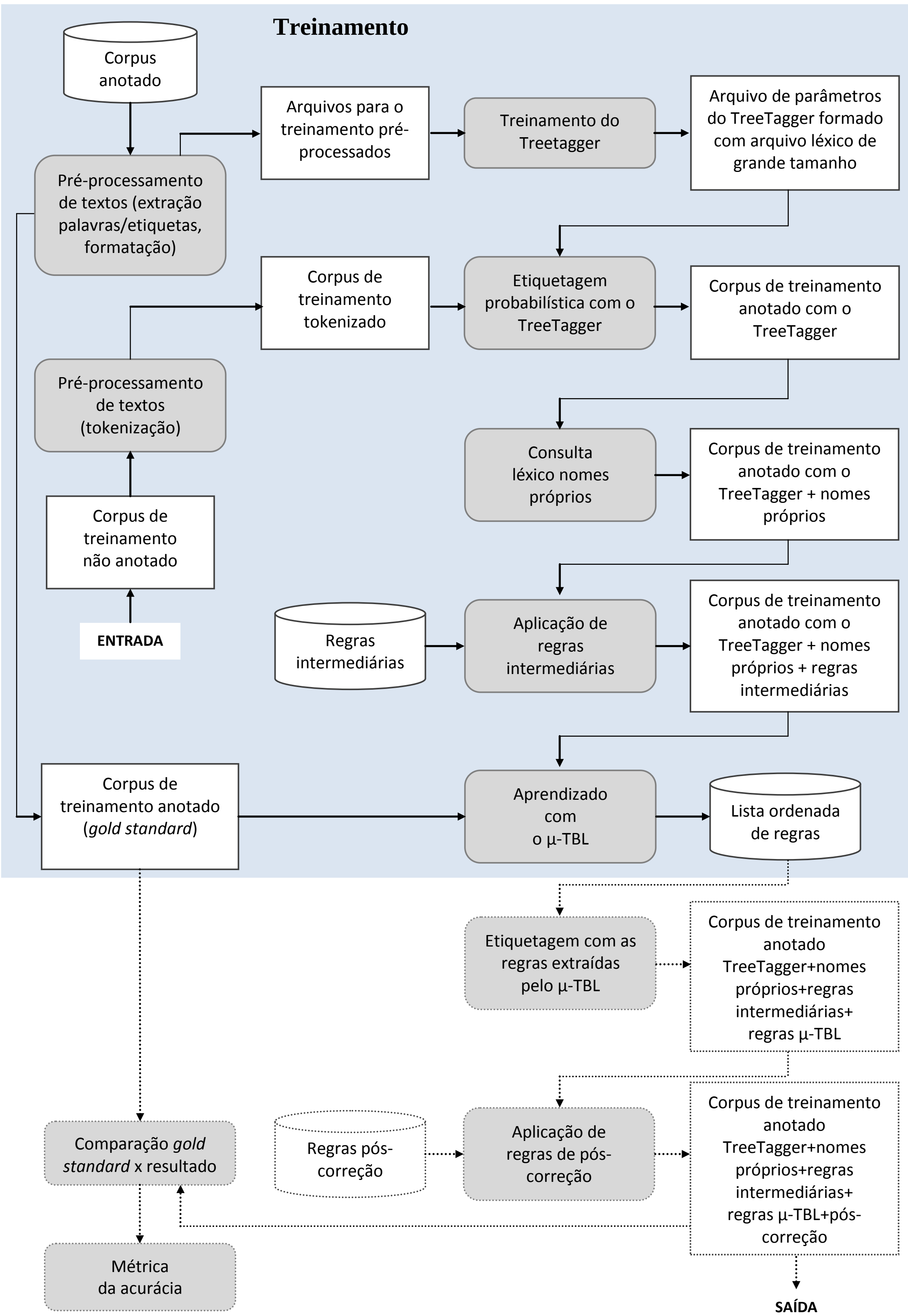


4.3.3 Treinamento no Modelo Final de Etiquetador Proposto

No modelo final, o etiquetador passou a ser treinado, na etapa probabilística, com um arquivo léxico de grande tamanho. Foram adicionados os módulos referentes à consulta ao léxico de nomes próprios e à aplicação da base de regras intermediárias antes do aprendizado automático de regras com o µ-TBL. Essas alterações na estrutura para treinamento do etiquetador são ilustradas na Figura 4.5. Os módulos com linhas pontilhadas não fazem parte do treinamento.

- Treinamento:
  1) Pré-processamento de textos.
  2) Treinamento do TreeTagger.
  3) Consulta ao léxico de nomes próprios.
  4) Aplicação de regras intermediárias
  5) Treinamento do µ-TBL.

FIGURA 4.5 - Treinamento do etiquetador proposto

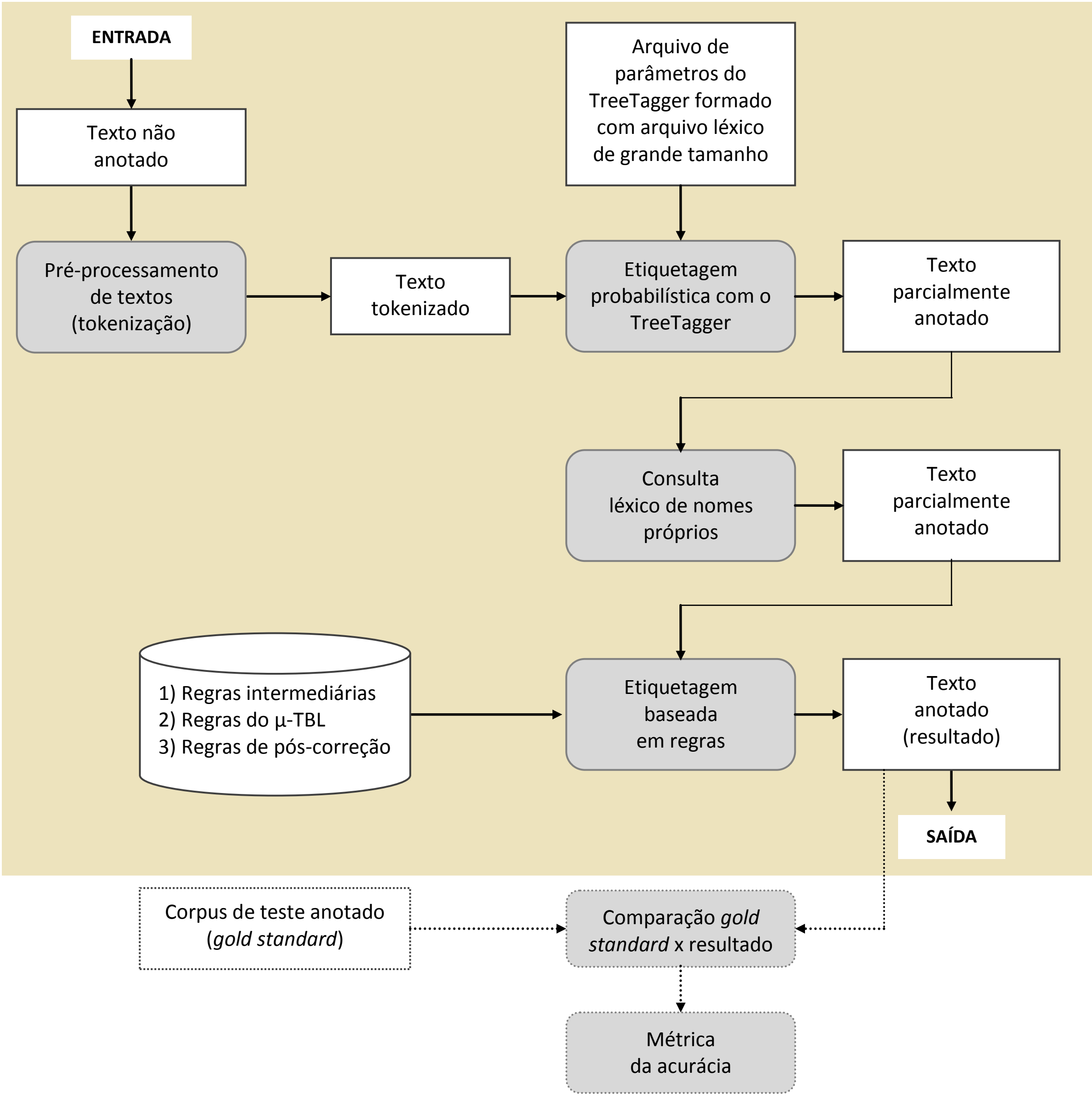


4.3.4 Arquitetura do Etiquetador Proposto

A arquitetura do modelo de etiquetagem proposto apresenta os módulos listados abaixo e é ilustrada na Figura 4.6.

- Etiquetagem:
  1) Pré-processamento de textos.
  2) Etiquetagem probabilística com o TreeTagger.
  3) Consulta ao léxico de nomes próprios.
  4) Etiquetagem com um conjunto intermediário de regras codificadas manualmente.
  5) Etiquetagem com regras extraídas automaticamente com o µ-TBL.
  6) Etiquetagem com um conjunto de regras de pós correção codificadas manualmente.
  7) Comparação dos textos pré-anotados com os resultantes da etiquetagem híbrida.
  8) Métrica da acurácia de etiquetagem e a obtenção das estatísticas do processo.

FIGURA 4.6 - Arquitetura do etiquetador híbrido proposto



4.4 Processo de Etiquetagem

Para realizar o processo de etiquetagem completo de textos do português brasileiro (incluindo todas as etapas, como pré-processamento de textos e treinamento do etiquetador), de acordo com a arquitetura do etiquetador proposto, são realizadas as etapas descritas a seguir.

4.4.1 Pré-Processamento

Nessa etapa, é feita a adequação dos textos ao processo de etiquetagem, que envolve a realização das tarefas de: 1) extração de palavras e etiquetas do *corpus* para formar arquivos de treinamento e arquivos de teste e 2) modificações nos arquivos de teste para permitir a comparação e métrica da acurácia as quais se aplicam conforme o formato do *corpus* e os requisitos da ferramenta de etiquetagem.

4.4.1.1 Extração de palavras e etiquetas dos *corpora* para formar arquivos de treinamento e arquivos de teste

Os *corpora* armazenam informações linguísticas nos mais variados formatos, por exemplo, no formato de ADs como é o caso do CETENFolha (Figura 4.1), em formato já adequado para a entrada do etiquetador como é o caso do Mac-Morpho (Figura 4.2), em XML (Figura 4.7), SQL, ADCG e outros. O passo inicial é extrair a informação desejada e colocá-la em arquivos no formato requerido pelo etiquetador.

FIGURA 4.7 - Exemplo de um *corpus* com informações linguísticas em formato XML

```
<source>CF2-1 «Confissões» chega a Portugal
  </source>
    <tree cat="fcl" fun="STA">
      <punct ort="«"/>
      <tree cat="np" fun="SUBJ">
        <t cat="prop" lemma="Confissões" args="F S"
fun="H">Confissões</t>
      </tree>
      <punct ort="»"/>
      <tree cat="vp" fun="P">
        <t cat="v-fin" lemma="chegar" args="PR 3S IND"
fun="MV">chega</t>
      </tree>
      <tree cat="pp" fun="SA">
        <t cat="prp" lemma="a" fun="H">a</t>
        <tree cat="np" fun="P&lt;">
          <t cat="prop" lemma="Portugal" args="M S"
fun="H">Portugal</t>
        </tree>
      </tree>
    </tree>
  </extract>
  <extract>
```

Fonte: LINGUATECA, 2009.

Os etiquetadores multilíngues necessitam de arquivos de treinamento que servirão de entrada para a formação de arquivos de recursos léxicos de uma determinada língua e também de arquivos de teste para avaliar o sistema de etiquetagem. Neste trabalho, foram programadas ferramentas de apoio na linguagem Java para realizar essas tarefas de acordo com o *corpus* e as necessidades de cada etiquetador utilizado.

Para o QTag, os arquivos de treinamento e teste no formato *txt* devem apresentar, em cada linha, uma palavra e sua etiqueta separadas por espaço em branco, tabulação ou quebra de linha.

Para o TreeTagger, são necessários três arquivos para gerar o arquivo de parâmetros ou arquivo de recursos do português brasileiro:

- **arquivo de treinamento**, que deve apresentar, em cada linha, uma palavra e sua etiqueta separadas por tabulação (semelhante ao do QTag);
- **arquivo léxico**, que deve apresentar, em cada linha, a palavra e as etiquetas possíveis para essa palavra, extraídas do *corpus*, separadas por tabulação e, ao lado de cada etiqueta, separado por espaço, o seu *lemma*; se este não estiver disponível, o *lemma* é substituído por hífen;
- **arquivo de classes abertas**, que contém possíveis etiquetas para palavras desconhecidas, contendo uma classe ou categoria (etiqueta) por linha.

Os arquivos de teste possuem o mesmo formato dos arquivos de treinamento.

4.4.1.2 Modificações nos arquivos de teste para permitir a comparação e métrica da acurácia

A etapa de etiquetagem considera não apenas a etiquetagem das palavras em si, mas também deve propiciar uma forma de medir a acurácia dos resultados dessa etiquetagem. Para medir a acurácia, é necessário fazer uma comparação dos dados etiquetados corretamente (*arquivo modelo*) obtidos do *corpus* com os dados que contêm erros (*arquivo de resultado ou saída*) resultantes da etiquetagem com as ferramentas escolhidas. Dessa forma, os dois arquivos devem ter o mesmo número de linhas, pois os resultados são produzidos na forma de pares *palavra_etiqueta* por linha. Essa comparação só será realizada no pós-processamento. No entanto, no caso do Bosque CF 7.4, desde o início é necessário realizar alguns tratamentos nas palavras que serão etiquetadas para que fiquem compatíveis com o *corpus* de treinamento. Assim, foi implementado um módulo de pré-processamento para fazer a concatenação/separação de palavras conforme resumido na Tabela 4.3. Para identificar os casos mencionados nessa tabela, foi construído um léxico (dicionário) extraído do *corpus* contendo esses casos.

TABELA 4.3 - Tarefas de pré-processamento das palavras que serão etiquetadas nos arquivos de teste

Problema	**Característica**	**Tarefa de Pré-processamento**
Concatenação de palavras	O CETENFolha reúne expressões multipalavras definidas como Entidades Mencionadas (EMs) em uma só palavra. Essas expressões são nomes próprios e palavras que formam uma única entidade, p. ex., *secretário de Estado*, que é etiquetada como um nome comum: *secretário_de_Estado n*; *São Paulo*, que é etiquetada como um nome próprio: *São_Paulo prop* (AFONSO, 2006). Da mesma forma que as EMs, algumas preposições, pronomes e outros são reunidos em uma só palavra, p. ex.: ao contrário de (ao_contrário_de), ainda que (ainda_que), de o que (do_que), mesmo que (mesmo_que), etc. Correção de separações que não devem ser realizadas: av . (av.), tel . (tel.), 1 . (1.), etc.	Identificar as ocorrências desses casos e concatenar as palavras.
Separação de palavras	Preposições, pronomes, conjunções, etc.: no *Corpus*, palavras como as exemplificadas a seguir aparecem separadas: da (de a), do (de o), dele (de ele), disto (de isto), daquilo (de aquilo), no (em o), nele (em ele), nisto (em isto), naquilo (em aquilo), etc. Sinais de pontuação, p. ex., "descasadas" (" descasadas "); sinais de crase ou contração preposição+artigo: à (a a), ao (a o), àquele (a aquele), etc;	Identificar as ocorrências desses casos e separar as palavras.

4.4.2 Etiquetagem Probabilística

A etiquetagem probabilística é realizada em duas etapas: 1) construção do arquivo de parâmetros ou recursos léxicos e 2) etiquetagem de sentenças.

4.4.2.1 Construção do arquivo de parâmetros ou recursos léxicos (treinamento do etiquetador)

Nessa etapa, é realizado o treinamento do etiquetador, que resulta na construção de um arquivo de parâmetros ou recursos léxicos, o qual contém as informações da língua em estudo que são: todas as palavras do *corpus*, as suas possíveis etiquetas, as matrizes de sequências de etiquetas e suas probabilidades anotadas. Os comandos para criação desse arquivo com o QTag e TreeTagger são apresentados no Apêndice A e B, respectivamente.

4.4.2.2 Etiquetagem de sentenças

A etiquetagem das sentenças dos conjuntos de teste é feita com o arquivo de parâmetros. Os comandos para essa tarefa com o QTag e TreeTagger são apresentados nos Apêndices A e B, respectivamente.

4.4.3 Consulta ao Léxico de Nomes Próprios

Nessa etapa, o algoritmo de etiquetagem consulta um léxico de cerca de 60.000 nomes próprios extraído do corpus CETENFolha para reetiquetar nomes próprios que foram etiquetados incorretamente na etapa de etiquetagem probabilística.

4.4.4 Etiquetagem Baseada em Regras

4.4.4.1 Com um conjunto intermediário de regras codificadas manualmente

Nessa etapa, são aplicadas 116 regras codificadas manualmente e escritas em Java para corrigir erros da etapa anterior do processo.

4.4.4.2 Com regras extraídas automaticamente com o μ-TBL

O sistema μ-TBL usado para a extração de regras de forma automática requer um arquivo de treinamento e arquivos de teste para avaliação das regras aprendidas. O arquivo de treinamento é preparado em um formato a ser aceito pelo μ-TBL, o qual executa comandos na linguagem Prolog. Esse arquivo, a cada sequência de três linhas contém os seguintes dados: na primeira linha, o número da linha e a palavra a ser etiquetada em uma sentença; na segunda linha, o número da linha e a etiqueta atribuída pelo etiquetador à palavra, que pode estar certa ou errada; e, na terceira linha, a etiqueta atribuída, a etiqueta correta e o número da linha. Na Figura 4.8, por exemplo, é possível verificar um erro na etiqueta da palavra *publicada*, que é a sétima palavra da sentença que está sendo etiquetada. O arquivo de teste é construído de forma semelhante ao arquivo de treinamento. Para o treinamento, é preciso informar alguns parâmetros, tais como a escolha de um dos três algoritmos (*brill*, *simple* ou *lazy*), o conjunto de *templates* (*brill_templates* disponível para a tarefa de etiquetagem) e os valores mínimos de *score* e acurácia da regra desejados.

FIGURA 4.8 - Formato do arquivo de treinamento do µ-TBL

```
:- dynamic wd/2, tag/2, tag/3.
:- set data_size=76098.

wd(1,'PT').
tag(1,'prop').
tag('prop','prop',1).
wd(2,em).
tag(2,'prp').
tag('prp','prp',2).
wd(3,o).
tag(3,'art').
tag('art','art',3).
wd(4,governo).
tag(4,'n').
tag('n','n',4).
wd(5,'Brasília').
tag(5,'prop').
tag('prop','prop',5).
wd(6,'Pesquisa_Datafolha').
tag(6,'n').
tag('n','n',6).
wd(7,publicada).
tag(7,'adj').
tag('adj','v-pcp',7).
...
```

Após o treinamento com os dados descritos acima, o µ-TBL extrai uma lista ordenada de regras para correção de erros de etiquetagem. A Figura 4.9 mostra uma pequena seção dessas regras. O µ-TBL também produz um arquivo no formato *html* que informa o número de ocorrências etiquetadas de forma errada em uma certa categoria gramatical e qual deveria ser a etiqueta correta, listando todas as sentenças nas quais ocorreu esse tipo de erro (Figura 4.10). A aplicação das regras extraídas sobre o arquivo de teste gera informações estatísticas sobre o treinamento e teste que podem ser examinadas no Apêndice C.

FIGURA 4.9 - Lista ordenada de regras aprendidas com o μ-TBL

```
% 2721 rules
tag:'v-fin'>prop <- wd:'São_Paulo'@[0] o
tag:'pron-det'>'pron-indp' <- wd:o_que@[0] o
tag:prp>adv <- tag:prp@[1] o
tag:n>prop <- wd:'Estados_Unidos'@[0] o
tag:'pron-det'>'pron-pers' <- wd:o@[0] & tag:'v-fin'@[1] o
tag:'v-fin'>'conj-s' <- wd:do_que@[0] o
tag:'v-fin'>prp <- wd:além_de@[0] o
...
```

FIGURA 4.10 - Arquivo com os erros após a aplicação das regras aprendidas com o μ-TBL

5 occurrences tagged as art that should be prp:

113: . « Confissões » chega **a** Portugal Desde o último dia
418: governo de destinar o FSE **a** investimentos sociais . O assessor
3115: morte que Rúbio teria feito **a** Claudirene » , disse Borlina
3494: maior impacto de o Nafta **a** curto prazo será sobre as
3612: e elitistas . Benetton volta **a** chocar com cartaz A Benetton
...

O μ-TBL aplica a lista ordenada de regras e informa a acurácia obtida, porém não substitui as etiquetas erradas pelas corretas no arquivo de teste. Para fazer essa substituição, utiliza-se um programa chamado *rule_compiler.pl* disponível no mesmo *site* do μ-TBL e que é executado a partir de um sistema em Prolog, por exemplo, o SWI-Prolog. Nesse caso, é fornecido ao sistema o arquivo com as palavras e as etiquetas que lhes foram atribuídas na etiquetagem probabilística (arquivo de teste) e um novo arquivo contendo o resultado é produzido após a aplicação das regras. Nos experimentos desta tese, usou-se uma forma opcional de realizar essa tarefa. As regras extraídas foram convertidas automaticamente para a linguagem Java e depois foram incorporadas à base de regras juntamente com as regras codificadas manualmente.

4.4.4.3 Com um conjunto de regras de pós-correção codificadas manualmente

Nessa etapa, 203 regras codificadas manualmente e escritas em Java são aplicadas para corrigir erros da etapa anterior do processo.

4.4.5 Comparação dos Resultados da Etiquetagem Híbrida com os Textos Pré-Anotados

Os textos submetidos à etiquetagem podem ser comparados com o seu equivalente pré-anotado, se este estiver disponível. Os arquivos são comparados linha a linha por um programa em Java, as linhas diferentes são contabilizadas como erros e as iguais como acertos.

Na etiquetagem probabilística com o Bosque CF 7.4, o arquivo resultado pode apresentar problemas que geram números de linhas diferentes e que prejudicam a comparação entre o *modelo* e o *resultado* para a métrica da acurácia obtida. Nesse caso, é preciso identificar pares *palavras_etiquetas* que ora aparecem com as palavras ligadas por um *underline* e possuem uma determinada etiqueta, ora não aparecem ligadas e possuem outras etiquetas. Alguns exemplos são os seguintes: *do_que conj-s* ou *de prp o pron-det que pron-indp* ; *o_que pron-indp* ou *o pron-det que pron-indp* ; *por_que adv* ou *por prp que pron-indp* ; *como_se conj-s ou como adv se pron-pers*, etc. Para identificar qual a forma correta dessas palavras (se ligadas ou não por *underline* e que etiquetas devem receber), foi implementado um módulo de programa em Java que, baseado em regras obtidas pela observação de padrões dessas palavras, realiza a concatenação ou separação das mesmas.

4.4.6 Métrica da Acurácia de Etiquetagem e a Obtenção das Estatísticas do Processo

Ao final da comparação do arquivo resultado com o arquivo modelo, é realizada a métrica da acurácia pela contagem do número de erros ou do número de acertos dividido pelo número total de palavras etiquetadas multiplicado por 100, obtendo-se assim as taxas de erro ou de acerto do etiquetador. Por exemplo, se foram etiquetadas 3980 palavras e houve 43 erros, a taxa de erro foi de 1,08% e a taxa de acerto de 98,92%. São também realizadas as métricas de acurácia global, em palavras conhecidas e em palavras desconhecidas; calculados o número de palavras desconhecidas no conjunto de teste, o número de palavras desconhecidas no conjunto de erros e o número de erros no conjunto de teste.

5 Etiquetagem Baseada em Textos do Bosque CETENFolha

"As idéias e as estratégias são importantes, mas o verdadeiro desafio é a sua execução." (Percy Barnevick)

Neste capítulo, são apresentadas as duas estratégias avaliadas, experimentalmente, em textos do Bosque CF 7.4: a aplicação de regras para corrigir erros que permaneceram após a etiquetagem probabilística e a modificação no conjunto inicial de etiquetas do *corpus*. Estas estratégias foram propostas na investigação inicial desta tese e contribuíram para que se chegasse ao modelo de etiquetador proposto. São também apresentados os experimentos mais relevantes com o Bosque CF 7.4 seguidos da análise de seus resultados.

5.1 Estratégias Avaliadas

5.1.1 Aplicação de Regras para Correção de Erros da Etiquetagem Probabilística

A primeira estratégia avaliada em experimentos com o Bosque CF 7.4 foi a aplicação de um conjunto de regras codificadas manualmente para a correção de erros da etiquetagem probabilística. Essas regras foram obtidas com dois métodos de extração:

1) Primeiro, foi formada uma base de dados dos contextos extraídos do *corpus* em uma janela de sete posições envolvendo as palavras e as etiquetas mais frequentes nos erros observados nos resultados da etiquetagem probabilística. Essa base de dados foi submetida a estudos de mineração de dados com algoritmos de classificação e associação da ferramenta WEKA (WITTEN; FRANK, 2005). Desse estudo, foram extraídas regras de forma automática, que por serem muito gerais, precisaram ser refinadas manualmente para corrigir os erros de etiquetagem, uma vez que quando havia exceções a essas regras, os erros continuavam acontecendo. Para a etiquetagem baseada no *corpus* com o conjunto de etiquetas inicial (etiquetas simples) do Bosque CF 7.4 (Tabela 4.1 – Capítulo 4), foram construídas 774 regras para correção de erros. Para a etiquetagem baseada no *corpus* com o conjunto de etiquetas modificadas, que será discutido na Seção 5.1.2, foram construídas 281 regras para correção de

erros. Na Figura 5.1, é mostrado um exemplo dessas regras codificado em Java.

2) Apesar dos bons resultados obtidos com essa solução, a análise de cada regra para verificar casos que não se enquadravam nos padrões obtidos tornava o trabalho lento e custoso. Por isso, nos experimentos seguintes adotou-se o uso do sistema μ-TBL para a extração automática de regras. Com esse sistema foram obtidas 2.721 regras para o conjunto de etiquetas simples e 1.598 regras para o conjunto de etiquetas modificadas.

FIGURA 5.1 - Regra para correção de erro de etiquetagem, programada em Java

```
/* Regra: Corrigir etiqueta da palavra “que”
 * “Se é encontrada vírgula seguida de que etiquetado como advérbio seguido de um verbo finito
 * seguido de um verbo no infinitivo, então trocar a etiqueta do que para pronome independente.”
 */
 if ((sl[i].equals(",_pt")) &&
      (sl[i+1].equals("que_adv"))&&
      (sl[i+2].endsWith("_v-fin")) &&
      (sl[i+3].endsWith("_v-inf")))
 {
       sl[i+1] = “que_pron-indp";
 }
```

5.1.2 Modificação no Conjunto Inicial de Etiquetas

A segunda estratégia avaliada em experimentos com o Bosque CF 7.4 foi a modificação do conjunto inicial de etiquetas do Bosque CF 7.4. Esta modificação foi proposta por se observar, durante a codificação de regras manuais, que em muitos casos não era possível formular uma regra contextual sem que houvesse informações sobre gênero, pessoa, tempo e modo verbais e outros atributos léxicos. Por exemplo, nas sentenças “ ‘*Confissões’ chega a Portugal*.” e “*Benneton volta a chocar com cartaz*.”, para resolver a ambiguidade da palavra “a” e escolher entre as etiquetas artigo ou preposição, a presença desses atributos léxicos nas etiquetas das palavras vizinhas facilita ao algoritmo a classificação correta, no caso o “a” deve ser etiquetado como preposição.

O novo conjunto de etiquetas modificadas, baseado no formato usado em Bick (2000), consistiu em combinar 10 das 18 categorias gramaticais das etiquetas simples com informações de gênero, número, função sintática, pessoa, tempo e modo verbais disponíveis no *corpus*. Essa combinação surgiu da análise de resultados após alguns experimentos práticos. O conjunto de etiquetas modificadas ficou com 152 etiquetas (Apêndice D) que passaram a incluir as informações mostradas na Tabela 5.1.

TABELA 5.1 - Características do conjunto de etiquetas modificadas

Conjunto de Etiquetas	Informações Associadas
adj (adjetivo), art (artigo), n (nome ou substantivo), pron-det (pronome determinativo), pron-indp (pronome independente), prop (nome próprio), v-pcp (verbo no particípio)	gênero e número
pron-pers (pronome pessoal)	gênero, número e função sintática
v-fin (verbo finito), v-inf (verbo no infinitivo)	pessoa, tempo e modo verbais
adv (advérbio), conj-c (conjunção coordenativa), conj-s (conjunção subordinativa), in (interjeição), num (numeral), prp (preposição), v-ger (verbo no gerúndio), pt (pontuação)	nenhuma

A Figura 5.2 mostra um exemplo de palavras de uma sentença do Bosque CF 7.4 com etiquetas do conjunto inicial (etiquetas simples) e do conjunto modificado (etiquetas modificadas).

FIGURA 5.2 - Exemplos dos conjuntos de etiquetas simples e modificadas

Palavra	Etiqueta simples	Etiqueta modificada	Descrição
Manchete	prop	propFS	nome próprio feminino singular
estréia	v-fin	v-finPR3SIND	verbo finito presente do indicativo 3ª pessoa do singular
novo	adj	adjMS	adjetivo masculino singular
jornalístico	n	nMS	nome masculino singular

Fonte: Bosque CETENFolha (LINGUATECA, 2007).

5.2 Treinamento e Teste de Etiquetadores com o Bosque CF 7.4

Os experimentos desta Seção tiveram como objetivo avaliar o desempenho dos etiquetadores quanto à acurácia nas três abordagens para etiquetagem – probabilística, híbrida e baseada em regras – em textos de um *corpus* pequeno, mas que apresenta menor quantidade de ruído, visto que já passou por várias revisões de especialistas humanos.

Na etiquetagem probabilística, foram usados os etiquetadores QTag e TreeTagger; na etiquetagem híbrida, a etiquetagem probabilística com o QTag e TreeTagger foi combinada com a etiquetagem baseada em regras codificadas manualmente e com a etiquetagem baseada em regras extraídas automaticamente pelo sistema µ-TBL; e na etiquetagem baseada em regras, foi avaliado o desempenho do etiquetador do analisador gramatical PALAVRAS no ambiente VISL. Os etiquetadores foram treinados com textos do Bosque CF 7.4 e testados em conjuntos disjuntos desse mesmo *corpus*.

Oito experimentos são apresentados: os dois primeiros com etiquetagem probabilística, os quatro seguintes com etiquetagem híbrida e os dois últimos, com etiquetagem baseada em regras.

5.2.1 Método de Avaliação

Devido ao pequeno tamanho do Bosque CF 7.4 (80.078 palavras), utilizou-se o método de avaliação *20-fold-cross-validation*, com o *corpus* subdividido em 20 subconjuntos (4.004 palavras por subconjunto, em média). Em cada uma das 20 iterações do algoritmo de etiquetagem, 19 subconjuntos (95% do *corpus*) foram tomados para treinamento e 01 subconjunto (5% do *corpus*) para teste, de forma que os subconjuntos de treinamento e teste eram disjuntos, significando que os dados de teste não eram vistos no treinamento, conforme é requerido pela tarefa de classificação. A acurácia do processo de etiquetagem por subconjunto de textos foi dada pela soma do número de acertos dividida pelo número total de palavras etiquetadas em cada repetição do experimento. Ao final, foi calculada a acurácia média das 20 repetições.

5.2.2 Parâmetros de Configuração das Ferramentas

Quanto aos parâmetros informados para treinamento dos etiquetadores:

- O QTag não requer parâmetros a serem configurados pelo usuário.
- Para o TreeTagger, foi mantida a configuração padrão com o parâmetro *contexto* informado como *trigramas*. O arquivo léxico foi extraído do *corpus* de treinamento e possui 14.280 entradas léxicas. Os arquivos de classes abertas usados para os conjuntos de etiquetas simples e modificadas são apresentados no Apêndice E.

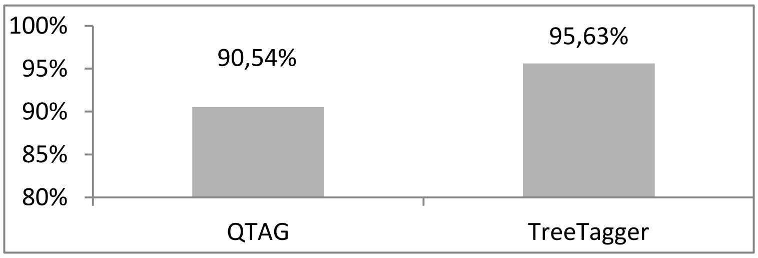
- O µ-TBL foi configurado com os parâmetros: algoritmo: *brill*, conjunto de *templates*: *brill_templates*, *score* mínimo: 1 e acurácia mínima: 0,5. Esses valores foram escolhidos baseados em resultados experimentais.

5.2.3 Etiquetagem Probabilística

5.2.3.1 Experimento 1

Este experimento compara o desempenho dos etiquetadores probabilísticos QTag e TreeTagger em textos do Bosque CF 7.4. O conjunto de etiquetas do *corpus* de treinamento foi o conjunto de 18 etiquetas simples. Os resultados são mostrados na Figura 5.3.

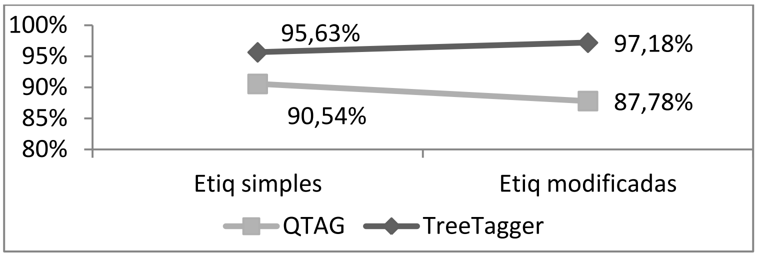
FIGURA 5.3 - Acurácia média da etiquetage probabilística com o QTag e TreeTagger



5.2.3.2 Experimento 2

Neste experimento, os etiquetadores QTag e TreeTagger foram treinados com os dois conjuntos diferentes de etiquetas do Bosque CF 7.4: simples e modificado. Os resultados são mostrados na Figura 5.4. A etiquetagem com o TreeTagger e o uso das 152 etiquetas modificadas aumentou em 1,55% a acurácia média em comparação com a acurácia média obtida com a utilização das 18 etiquetas simples. Com o QTag, a utilização de etiquetas modificadas provocou um decréscimo na acurácia média de 2,76%.

FIGURA 5.4 - Acurácia média com etiquetas simples e etiquetas modificadas

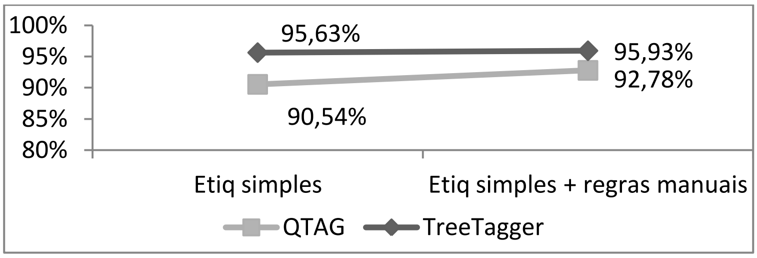


5.2.4 Etiquetagem Híbrida

5.2.4.1 Experimento 3

Este experimento apresenta o desempenho da etiquetagem híbrida, na qual é realizada a etiquetagem probabilística com o QTag e TreeTagger e, sobre os resultados dessa primeira etapa, é realizada a etiquetagem baseada em regras codificadas manualmente. Para treinamento e testes, foi usada a versão do *corpus* com o conjunto de 18 etiquetas simples. Na etapa baseada em regras, foram aplicadas as 774 regras manuais. A aplicação dessas regras aumentou a acurácia média em 2,24% e 0,30% com os etiquetadores QTag e TreeTagger, respectivamente (Figura 5.5). Ressalta-se que o menor aumento de acurácia para o TreeTagger ocorre pelo fato de a maioria das regras terem sido construídas baseadas nos erros do QTag.

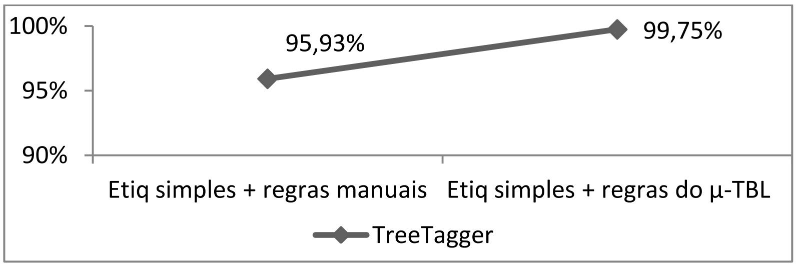
FIGURA 5.5 - Acurácia média da etiquetagem híbrida em *corpus* com etiquetas simples e a aplicação de 774 regras manuais



5.2.4.2 Experimento 4

Este experimento apresenta o desempenho da etiquetagem híbrida, na qual é realizada a etiquetagem probabilística com o TreeTagger e, sobre os resultados dessa primeira etapa, é realizada a etiquetagem baseada em regras extraídas automaticamente com o µ-TBL. Para treinamento e testes, foi usada a versão do *corpus* com o conjunto de 18 etiquetas simples. Na etapa baseada em regras, foram aplicadas as 2.721 regras extraídas com o µ-TBL para esse conjunto de etiquetas. A aplicação dessas regras elevou a acurácia média em 3,82% em relação à acurácia média obtida com as 774 regras codificadas manualmente (95,93%), conforme se observa na Figura 5.6.

FIGURA 5.6 - Acurácia média da etiquetagem híbrida em *corpus* com etiquetas simples e a aplicação de 2.721 regras extraídas com o µ-TBL

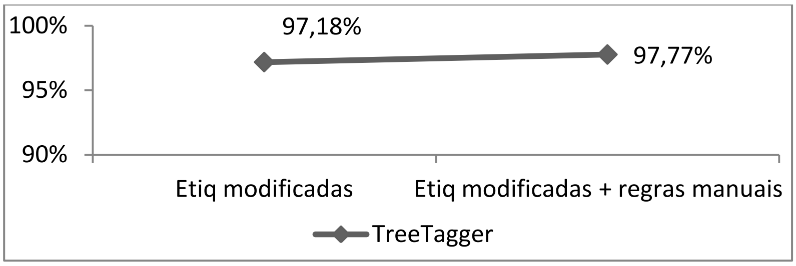


A partir do Experimento 4, em virtude do baixo desempenho apresentado pelo QTag, o que causava demora excessiva para a extração automática de regras que, muitas vezes, era interrompida por problemas no *hardware* ou outros, sem que se conseguisse concluí-la, os experimentos com essa ferramenta foram abandonados.

5.2.4.3 Experimento 5

Este experimento apresenta o desempenho da etiquetagem híbrida, na qual é realizada a etiquetagem probabilística com o TreeTagger e, sobre os resultados dessa primeira etapa, é realizada a etiquetagem baseada em regras codificadas manualmente. Para treinamento e testes, foi usada a versão do *corpus* com o conjunto de 152 etiquetas modificadas. Na etapa baseada em regras, foram aplicadas as 281 regras manuais. A aplicação dessas regras aumentou a acurácia média em 0,59%, conforme se observa na Figura 5.7.

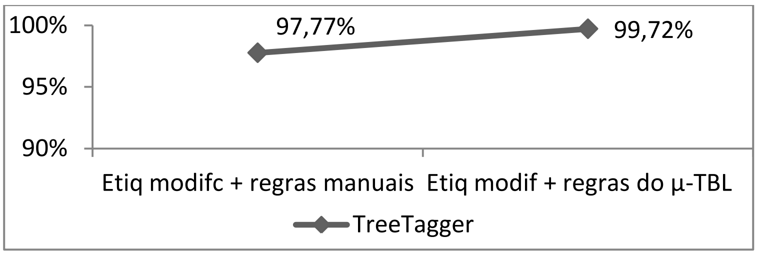
FIGURA 5.7 - Acurácia média da etiquetagem híbrida em *corpus* com etiquetas modificadas e a aplicação de 281 regras manuais



5.2.4.4 Experimento 6

Este experimento apresenta o desempenho da etiquetagem híbrida, na qual é realizada a etiquetagem probabilística com o TreeTagger e, sobre os resultados dessa primeira etapa, é realizada a etiquetagem baseada em regras extraídas automaticamente com o µ-TBL. Para treinamento e testes, foi usada a versão do *corpus* com o conjunto de 152 etiquetas modificadas. Na etapa baseada em regras, foram aplicadas as 1.598 regras extraídas com o µ-TBL. A aplicação dessas regras elevou a acurácia média em 1,95% em relação à acurácia média obtida com as 774 regras codificadas manualmente (97,77%), conforme se observa na Figura 5.8.

FIGURA 5.8 - Acurácia média da etiquetagem híbrida em *corpus* com etiquetas modificadas e a aplicação de 1.598 regras extraídas com o µ-TBL



Os resultados dos experimentos são resumidos na Tabela 5.2. É interessante observar nessa Tabela que, no resultado do Experimento 5, a diferença entre o percentual de acurácia do melhor subconjunto de teste (99,60%) e o percentual de acurácia média (97,77%) é bem maior (1,83%) do que nos resultados dos Experimentos 4 (99,90% no melhor subconjunto de teste, 99,75% de acurácia média e diferença de 0,15%) e 6 (99,95% no melhor subconjunto de teste, 99,72% de acurácia e diferença de 0,23%). Em virtude de as regras terem sido construídas à medida que os erros iam sendo analisados em cada um dos 20 subconjuntos de teste, essa diferença é maior no resultado do Experimento 5 devido não ter sido possível construir regras refinadas manualmente para corrigir erros de todos os 20 subconjuntos de teste.

TABELA 5.2 - Resultados da etiquetagem com o QTag e TreeTagger

Exp	**Configuração do Experimento**	**QTAG**			**TreeTagger**		
		Nº erros/ subconjunto em média	**Acurácia média**	**Acurácia no melhor subconjunto de teste**	**Nº erros/ subconjunto em média**	**Acurácia média**	**Acurácia no melhor subconjunto de teste**
1	Etiq probabilística: conj. de etiquetas simples	379	90,54%	91,57%	175	95,63%	96,64%
2	Etiq probabilística: conj. de etiquetas modificadas	490	87,78%	89,61%	113	97,18%	97,89%
3	Etiq híbrida: conj. de etiquetas simples + 774 regras manuais	290	92,78%	97,66%	163	95,93%	97,21%
4	Etiq híbrida: conj. de etiquetas simples + 2.721 regras do µ-TBL	-	-	-	90	97,77%	99,60%
5	Etiq híbrida: conj. de etiquetas modificadas + 281 regras manuais	-	-	-	10	99,75%	99,90%
6	Etiq híbrida: conj. de etiquetas modificadas + 1.598 regras do µ-TBL	-	-	-	11	99,72%	99,95%

5.2.5 Etiquetagem Baseada em Regras no Ambiente VISL

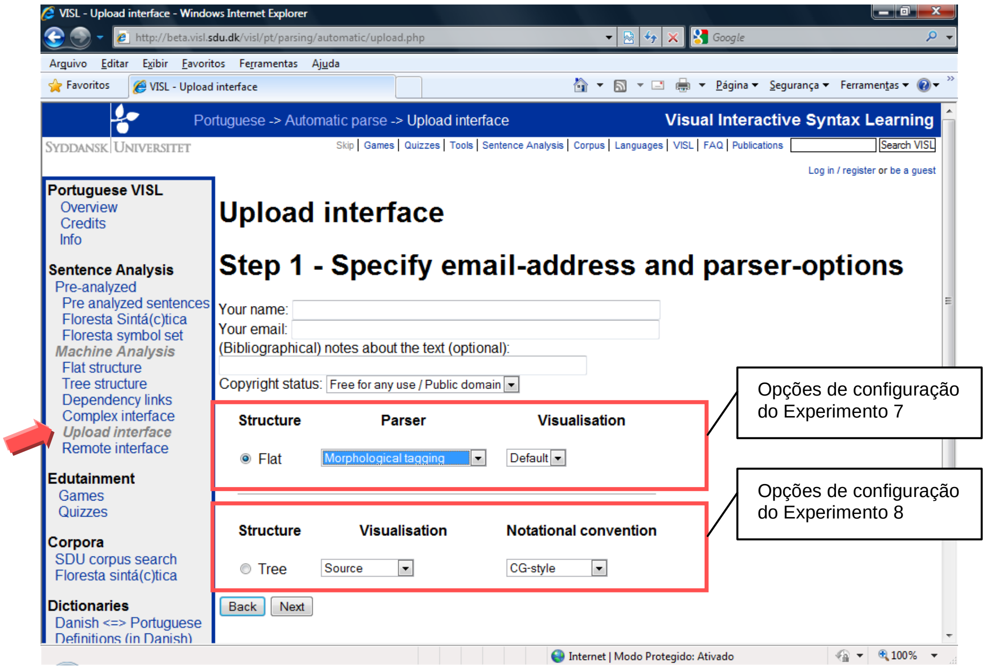
Com o objetivo de comparar o desempenho de um etiquetador baseado em regras com o desempenho dos etiquetadores probabilísticos e híbridos abordados neste trabalho, foram realizados dois experimentos com o etiquetador do analisador gramatical PALAVRAS (BICK, 2000) e as sentenças do Bosque CF 7.4. Para avaliação da acurácia, foi usado o método *20-fold-cross-validation*. Os experimentos foram configurados no ambiente VISL[12] (Figura 5.9), pela sequência de *links Sentence Analysis*, *Machine Analysis*, *Upload interface*, opção que foi selecionada por ser a recomendada pelo VISL para processar a quantidade de texto presente no Bosque CF 7.4.

Os dados etiquetados pelo PALAVRAS no VISL são fornecidos em formato diferente do formato do Bosque CF 7.4 (vide Apêndice F), o que dificultou a extração dos diferentes atributos léxicos para formar etiquetas iguais às do conjunto de etiquetas modificadas descrito

[12] http://beta.visl.sdu.dk/visl/pt/parsing/automatic/

na Tabela 5.1, de forma que se pudesse comparar os resultados, já que um novo *software* teria que ser construído para essa finalidade. Por esse motivo, nestes experimentos, avaliou-se apenas textos com etiquetas simples, que apesar de estarem em formato diferente ao do Bosque CF 7.4, tiveram a extração de etiquetas do arquivo de resultado realizada de forma menos trabalhosa.

FIGURA 5.9 - Tela de configuração do etiquetador e analisador gramatical PALAVRAS.



5.2.5.1 Experimento 7

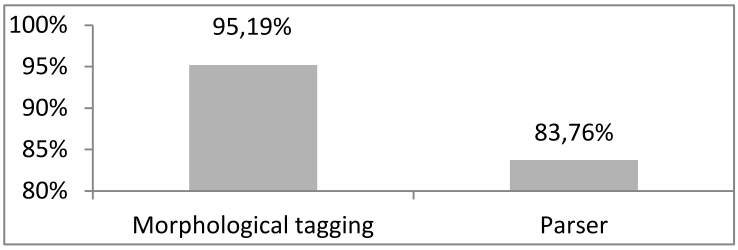
No Experimento 7, na tela *Upload interface*, a etiquetagem morfológica (*morphological tagging*) foi realizada com as opções de configuração: *flat structure*, *morphological tagging* e *default visualization* (Figura 5.9).

5.2.5.2 Experimento 8

No Experimento 8, na tela *Upload interface*, a etiquetagem foi realizada embutida na tarefa de análise gramatical (*parser*), com as opções de configuração: *tree structure*, *source visualization*, *CG-style notational convention* (Figura 5.9).

O Bosque CF 7.4 é um *corpus* etiquetado automaticamente pelo PALAVRAS e revisado manualmente por especialistas, por isso contém menos erros do que a sua versão que foi etiquetada nos Experimentos 7 e 8 desta Seção, cujos resultados são mostrados na Figura 5.10.

FIGURA 5.10 - Acurácia média com o etiquetador do PALAVRAS



5.2.6 Avaliação dos Resultados da Etiquetagem Baseada no Bosque CF 7.4

Com base na avaliação dos resultados dos experimentos, destacam-se as seguintes observações:

1) Na abordagem probabilística, a melhor acurácia média obtida com o conjunto de etiquetas simples foi de 95,63% e a melhor acurácia média obtida com o conjunto de etiquetas modificadas foi de 97,18%, ambas com o uso do etiquetador TreeTagger.
2) Na abordagem baseada em regras, com o etiquetador do PALAVRAS, a acurácia média obtida foi de 95,19% com a etiquetagem morfológica (*morphological tagging*)[13]. A etiquetagem embutida na aplicação do analisador gramatical (*parser*) apresentou uma queda de acurácia de 11,43% em relação à etiquetagem morfológica, em virtude das discordâncias entre etiquetas das palavras do *gold standard* e do resultado, que ocorreram principalmente em casos de artigos que foram etiquetados como pronome determinativo (vide Apêndice F).

[13] Nessa opção, o PALAVRAS atribui, em alguns casos, não uma única etiqueta para uma palavra, mas uma lista das etiquetas mais prováveis para essa palavra e a escolha da etiqueta correta fica a critério do usuário. Nesse experimento, essa escolha do usuário foi substituída pela atribuição da etiqueta correta como sendo a que consta no *corpus* previamente etiquetado – o Bosque CF 7.4 revisado – de forma que todas as ocorrências desse tipo foram consideradas como acerto.

3) A abordagem híbrida permitiu explorar duas estratégias para aumentar a acurácia no processo de etiquetagem: a modificação do conjunto de etiquetas inicial do *corpus* e a aplicação de regras (codificadas manualmente e extraídas de forma automática).
4) A melhor estratégia foi a etiquetagem com o TreeTagger seguida da aplicação de regras obtidas com o µ-TBL, que atingiu valores significativos de 99,75% e 99,72% de acurácia, em média, em conjuntos de testes com etiquetas simples e etiquetas modificadas, respectivamente.
5) O uso do conjunto de etiquetas modificadas apresentou ganho de acurácia em relação ao uso do conjunto de etiquetas simples apenas na etiquetagem probabilística.
6) Observa-se que quanto mais acertos houver na etiquetagem probabilística, menos regras serão necessárias, o que é importante em termos de custo computacional da aplicação.
7) Dependendo do tamanho do *corpus* de treinamento, da capacidade do microcomputador e dos parâmetros informados ao µ-TBL, a extração automática de regras com essa ferramenta pode levar de minutos a dias.
8) Ressalta-se que as regras construídas manualmente usadas nesses experimentos cobrem os casos de ambiguidade mais frequentes e que mais regras ainda poderiam ter sido construídas para melhorar a acurácia na etiquetagem. Para o subconjunto de teste com a melhor acurácia obtida com regras manuais (99,60%), foi construído o máximo de regras possível, servindo como parâmetro de comparação para a acurácia obtida com as regras extraídas automaticamente.
9) Com poucas etiquetas (18) e muitas regras manuais (774), o melhor resultado foi de 95,93% (acurácia média) e 97,21% (acurácia do melhor subconjunto de teste).
10) Com muitas etiquetas (152) e menos regras manuais (281), o melhor resultado foi de 97,77% (acurácia média) e 99,60% (acurácia do melhor subconjunto de teste).
11) O TreeTagger foi a ferramenta que mostrou melhor acurácia na etiquetagem probabilística.
12) É importante a ferramenta de etiquetagem oferecer um recurso de visualização dos erros como, por exemplo, o arquivo em formato *html* gerado após a

aplicação das regras extraídas pelo µ-TBL (vide Figura 4.10 – Capítulo 4), o que permite verificar com maior facilidade se há problemas na anotação previamente realizada no *corpus* de treinamento, o que por sua vez facilita a correção desses problemas.

Apesar dos ótimos resultados alcançados nos experimentos desta Seção, realizados em um ambiente controlado (*corpus* de pequeno tamanho, pouco ruído, gênero jornalístico), esses resultados podem não ser tão bons em novos textos, dado o tamanho reduzido do *corpus* de treinamento (80.078 palavras ou *tokens*) e a ausência de muitos padrões nos dados linguísticos. Assim, na próxima Seção será avaliado o *corpus* Mac-Morpho (LÁCIO-WEB, 2007), um *corpus* que contém maior volume de textos do português brasileiro.

6 Etiquetagem Baseda em Textos do Corpus Mac-Morpho

"Criatividade consiste no total rearranjo do que sabemos com o objetivo de descobrir o que não sabemos." (George Kneller)

Neste capítulo, são apresentadas as três estratégias avaliadas, experimentalmente, em textos do *Corpus* Mac-Morpho e Bosque CF 7.4, as quais foram incorporadas ao modelo de etiquetador proposto: a consulta a um léxico de nomes próprios, a adição de regras codificadas manualmente ao método TBL e o uso de um léxico de grande tamanho para o TreeTagger. São também apresentados os experimentos mais relevantes com os *corpora* Mac-Morpho e Bosque CF 7.4 seguidos da análise de seus resultados. Os recursos etiquetadores x *corpora* foram combinados nas análises:

- Treinamento dos etiquetadores com o *Corpus* Mac-Morpho:
  - Teste com o *Corpus* Mac-Morpho
  - Teste com o Bosque CF 7.4
- A estratégia de modificação no conjunto de etiquetas não pôde ser aplicada ao *corpus* Mac-Morpho, uma vez que esse *corpus* só disponibiliza a categoria gramatical da palavra.

6.1 Estratégias Avaliadas

6.1.1 Consulta a um Léxico de Nomes Próprios

Essa estratégia se refere à consulta realizada pelo algoritmo de etiquetagem a um arquivo léxico contendo cerca de 60.000 nomes próprios para corrigir os erros da etiquetagem probabilística que permaneceram nessa categoria de palavras e, dessa forma, melhorar a acurácia do etiquetador.

Erros com a etiqueta nome próprio (NPROP) são muito frequentes nos resultados da etiquetagem em textos jornalísticos, chegando a cerca de 23% do total de erros nos experimentos desta tese. Para resolver parte desses erros, o algoritmo de etiquetagem proposto consulta um dicionário ou léxico de nomes próprios. Se as palavras do conjunto de teste estão presentes no léxico, então elas são etiquetadas como nomes próprios. Este léxico foi extraído de cerca de 28 milhões de palavras ou *tokens* que compõem o *Corpus* CETENFolha. Ao ser formado esse arquivo léxico, foi verificado se havia palavras que tinham mais de uma etiqueta possível além de nome próprio e, em caso positivo, essas palavras foram removidas do léxico. Foi observado também que o léxico trouxe alguns erros (ruídos) de sua etiquetagem automática original, por exemplo, palavras etiquetadas como nomes próprios que eram nomes comuns no início das frases. Durante o treinamento do etiquetador, erros como estes foram identificados e tais palavras foram retiradas do léxico. É possível inserir cada vez mais nomes próprios nesse léxico, por exemplo, nomes próprios de outros domínios que não o jornalístico.

6.1.2 Adição de Regras ao Método TBL

Durante a revisão dos erros de etiquetagem, foi observado que a adição de regras codificadas manualmente poderia corrigir muitos problemas. Essa estratégia de adição de regras ao modelo TBL também foi usada por Finger (2000). As regras foram construídas de forma semelhante à citada na Seção 5.1.1 e também com base na documentação do conjunto de etiquetas do Mac-Morpho (LÁCIO-WEB, 2008), na qual são descritas regras gramaticais da língua portuguesa para a formação desse conjunto de etiquetas.

Dois módulos de regras foram aplicados na etiquetagem híbrida: um conjunto de regras intermediárias usado para corrigir erros das saídas da etiquetagem probabilística e um conjunto de regras de pós-correção usado para corrigir erros das saídas da etiquetagem com sistema μ-TBL.

6.1.2.1 Regras para a correção de erros em uma etapa intermediária de etiquetagem

Foram codificadas em Java 116 regras intermediárias que cobrem os casos mostrados na Tabela 6.1. As regras corrigem erros da palavra corrente e das palavras seguintes. Alguns exemplos de regras intermediárias são:

- Regra para correção de etiqueta para a palavra “De” ou “de” seguida pela palavra “esta” ou pela palavra “estas”:

  “SE a palavra corrente é ‘De’ ou ‘de’ e a próxima palavra é ‘esta’ ou ‘estas’ etiquetada como PROADJ ou PROSUB, ENTÃO mudar a etiqueta da palavra ‘De’ ou ‘de’ para PREP|+.”

- Regra para correção de etiqueta de uma sequência de palavras, as quais devem ser nomes próprios (NPROP) em vez de nomes comuns (N), por exemplo, a frase: "*As Mil e Uma Noites*":

  "SE a palavra corrente *i* é igual a "(aspas), as palavra *i*+1 e *i*+2 iniciam com letra maiúscula, a palavra *i*+3 é igual a 'e' ou 'de', as palavras *i*+4 e *i*+5 iniciam com letra maiúscula e a palavra *i*+6 é igual a "(aspas), ENTÃO mudar a etiqueta das palavras *i*+1, *i*+2, *i*+3, *i*+4 e *i*+5 para NPROP."

TABELA 6.1 - Regras para correção de erros em uma etapa intermediária de etiquetagem

Etiqueta(s)	**Informação sobre a regra**	**Número de regras**
PREP ou PREP|+	Regras para correção de etiquetas para preposições (PREP) e para preposição com contração (PREP|+), e.g., *de*_PREP|+ *a*_art, *de*_PREP|+ *aquele*_PROADJ, de_prep *outro*_PROADJ.	42
N|DAD, N|DAT, N|HOR e N|TEL	Regras para correção de etiquetas das palavras que têm um formato pré-definido para N|DAD, N|DAT, N|HOR E N|TEL, e.g., 10 x 30, 01/01/1994.	15
NPROP	Regras para correção de etiquetas para sequências de palavras que devem ser etiquetadas como nomes próprios.	16
Outras etiquetas	Regras para correção de etiquetas para ADV, ADV|[, ADJ, ART, IN, KS, KC|[, KC|], N, NUM, PRO-KS, PROPESS, PROSUB, V, VAUX.	43

6.1.2.2 Regras para pós-correção de erros

Após a aplicação das regras do µ-TBL, foi observado que um novo conjunto de regras codificado manualmente poderia resolver os erros listados na Tabela 6.2, os quais são causados por ocorrências de padrões estruturais de baixa frequência ou por ocorrências envolvendo palavras desconhecidas que não puderam ser inferidas pelo sistema. Da mesma maneira usada para codificar as regras intermediárias, foram codificadas em Java 203 regras de pós-correção.

No caso de etiquetas complementares para palavras estrangeiras, é muito difícil para os etiquetadores classificar uma palavra estrangeira desconhecida. Nesses casos, a etiquetagem dessas palavras será considerada correta se o etiquetador atribuir etiquetas correspondentes na língua portuguesa, por exemplo, N, ADJ ou ADV se a etiqueta deveria ser N|EST, ADJ|EST ou ADV|EST, respectivamente.

TABELA 6.2 - Regras para pós-correção de erros

Etiqueta(s)	**Informação sobre a regra**	**Número de regras**
PCP	Regras para correção de etiquetas para particípio passado (PCP)[14].	36
ADJ	Regras para correção de etiquetas para adjetivo (ADJ).	44
N	Regras para correção de etiquetas para nome (N).	43
ART, PREP	Regras para correção de etiquetas da palavra “a” para artigo (ART) ou para preposição (PREP).	16
Outras Etiquetas	Regras para correção de etiquetas para ADV, IN, KC, KS, PDEN, PRO-KS-REL, PROPESS, V|+, VAUX.	64

6.1.3 Uso de um Léxico de Grande Tamanho para o TreeTagger

A partir dos resultados de experimentos com a abordagem híbrida, observa-se que quanto maior a acurácia na etapa probabilística, menos regras precisarão ser geradas, o que melhora o desempenho do etiquetador híbrido, principalmente na fase de treinamento, na qual a extração de regras requer um tempo bem maior do que a extração das probabilidades léxicas. Dessa forma, buscou-se uma estratégia que aumentasse a acurácia na etapa probabilística.

Apesar do *Corpus* CETENFolha conter 28 milhões de palavras e estar disponível para uso, esse *corpus* foi anotado automaticamente e possui muitos erros de etiquetagem. Contudo, as palavras desse *corpus* e suas etiquetas potenciais podem ser inseridas no arquivo léxico do TreeTagger, que deverá ser retreinado. Essa estratégia reduz significativamente o efeito das palavras desconhecidas nos erros de etiquetagem, conforme será explorado nos experimentos deste Capítulo. A possibilidade de usar um léxico na etapa de treinamento de uma maneira similar à descrita por Banko e Moore (2004, p. 558), a qual explora o conhecimento sobre quais etiquetas são possíveis para cada palavra no léxico, eleva o desempenho em termos de acurácia do etiquetador. Dessa forma, foi extraído um léxico do CETENFolha, o qual foi adicionado ao léxico extraído do Mac-Morpho, que resultou em um léxico final contendo cerca de 379.000 palavras diferentes (entradas léxicas) e suas possíveis etiquetas.

[14] No Mac-Morpho, devido à dificuldade em resolver a ambiguidade que pode ocorrer entre uma forma terminada em do(a) dos verbos, que pode ter tanto a função de adjetivo quanto do particípio de um verbo, foi atribuída a etiqueta PCP para ambos os casos independente de exercer uma ou outra função.

6.2 Treinamento do Etiquetador com o Corpus Mac-Morpho

O Mac-Morpho possui 1.221.538 palavras. Em virtude de terem sido encontrados ruídos no *corpus*, este passou por uma revisão em sua anotação. Para os experimentos desta tese, foi usado um conjunto de 496.281 palavras já revisto.

6.2.1 Etiquetagem em Conjuntos de Testes do *Corpus* Mac-Morpho

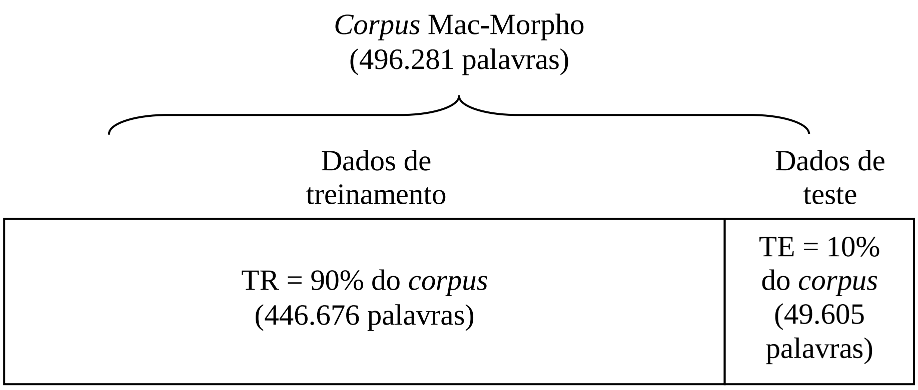
Os experimentos desta Seção tiveram como objetivo avaliar o desempenho dos etiquetadores TreeTagger e µ-TBL, bem como o modelo que aplica as estratégias propostas, quanto à acurácia na abordagens híbrida para a etiquetagem de textos do português brasileiro.

Cinco experimentos são apresentados, todos com etiquetagem híbrida.

6.2.1.1 Métodos de avaliação

O conjunto de textos do *corpus* Mac-Morpho usado neste estudo contém 496.281 palavras. Este conjunto foi dividido em dois conjuntos disjuntos: um conjunto de treinamento, aqui chamado de TR, que contém 90% do *corpus* (446.676 palavras) e um conjunto de teste, aqui chamado de TE, que contém 10% do *corpus* (49.605 palavras), não visto no processo de treinamento, conforme está esquematizado na Figura 6.1.

FIGURA 6.1 - Divisão do *corpus* em conjuntos de treinamento (TR) e teste (TE)



Os conjuntos TE e TR apresentam igual proporção de textos de cada seção jornalística do Mac-Morpho. Na etapa inicial, o arquivo léxico requerido pelo TreeTagger foi extraído de TR e contém 34.986 entradas léxicas e suas possíveis etiquetas.

6.2.1.1.1 Avaliação na etapa probabilística

A avaliação de acurácia na etapa probabilística é realizada de duas maneiras:

1) Pela aplicação do método de validação cruzada com 10 divisões sobre TR, que é tomado como se fosse o *corpus* completo. Para esse propósito, TR é dividido em 10 subconjuntos. Desses, cada nove subconjuntos combinados com o léxico formam um arquivo de parâmetros que é usado para etiquetar seu respectivo subconjunto disjunto ou conjunto de teste da validação cruzada, aqui chamado de TE_{VC}. Portanto, são formados 10 arquivos de parâmetros e a etiquetagem se dá em 10 iterações, produzindo 10 resultados de textos etiquetados (vide comandos no Apêndice G). Em média, cada conjunto de treinamento contém 402.008 palavras e cada TE_{VC} contém 44.668 palavras. Após a aplicação do processo de validação cruzada, os 10 resultados anotados são comparados com seus respectivos subconjuntos pré-anotados. A acurácia de cada resultado, bem como a acurácia média dos 10 resultados é então calculada.
2) Nesses resultados, diz-se que há *overfitting*, em virtude de muitas estruturas da língua não ocorrem no treinamento e de haver poucas palavras desconhecidas na avaliação com validação cruzada, visto que o léxico foi extraído do conjunto TR inteiro. Isto faz com que o modelo de classificação extraído do treinamento não alcance uma acurácia tão boa em dados novos quanto foi com a validação cruzada. Para reduzir o *overfitting*, o segundo caso de avaliação é sobre TE, que possui muitas palavras desconhecidas e novas estruturas linguísticas. Neste caso, utiliza-se um arquivo de parâmetro formado pela combinação de TR com o léxico para a etiquetagem de TE (vide comandos no Apêndice G) e, ao final, compara-se o resultado desta avaliação com o seu respectivo modelo pré-anotado e calcula-se a acurácia.

6.2.1.1.2 Avaliação na etapa baseada em regras

Na etapa baseada em regras, os 10 conjuntos de testes TE_{VC}, após terem sido etiquetados com o TreeTagger, são concatenados em um único arquivo, o qual contém os erros da etiquetagem probabilística e possui o mesmo tamanho (446.676 palavras) e os mesmos textos que TR. O µ-TBL compara as palavras e etiquetas desses dois conjuntos de dados (modelo x resultado) e extrai uma lista ordenada de regras (vide comandos no Apêndice H), as quais são então aplicadas sobre os conjuntos de testes TE_{VC} e sobre TE. Ao final desse processo, são calculadas as taxas de acurácia em cada um dos 10 resultados dos conjuntos de testes TE_{VC}, bem como a acurácia média desses 10 resultados, e a acurácia em TE.

6.2.1.2 Experimento 1

Com este experimento, foram obtidos valores de referência (*baseline*) quanto à acurácia da etiquetagem híbrida com o método TBL sobre os conjuntos de testes do *Corpus* Mac-Morpho (TE_{VC} e TE). Neste experimento, foram realizadas as duas etapas do método TBL: a etiquetagem probabilística seguida da etiquetagem com regras extraídas pelo µ-TBL. Na etapa baseada em regras, o µ-TBL extraiu 806 regras em 09 h 58 min. Na Tabela 6.3, são apresentados os resultados e alguns números importantes no processo de etiquetagem. Os valores relativos aos conjuntos de testes TE_{VC}, com exceção do tamanho do conjunto de treinamento na etiquetagem baseada em regras, são a média das dez iterações do algoritmo de etiquetagem.

TABELA 6.3 - Desempenho da etiquetagem híbrida no *Corpus* Mac-Morpho

Item	**Etiquetagem probabilística com o TreeTagger**		**Etiquetagem baseada em regras com o µ-TBL**	
	Conj TE_{VC}	**Conj TE**	**Conj TE_{VC}**	**Conj TE**
Tamanho conj treinamento	402.008	446.676	446.676	446.676
Tamanho conj teste	44.668	49.605	44.668	49.605
# palavras desconhecidas no conj de teste	2.347	2.808	2.347	2.808
# palavras desconhecidas no conj de erros	10	390	10	388
# erros no conj de teste	1.707	2.524	905	1.806
% acurácia global	**96,18**	**94,91**	**97,97**	**96,36**
% acurácia em palav conhecidas	95,99	95,44	97,89	96,97
% acurácia em palav desconhecidas	99,57	86,11	99,57	86,18

As estatísticas sobre os erros de etiquetagem nos conjuntos de testes TE_{VC}, ao final do processo, foram calculadas e as categorias gramaticais que mais apresentaram erros foram identificadas. Essas categorias com erros são quase sempre as mesmas em ambas as etapas probabilística e baseada em regras, tanto para palavras conhecidas como para desconhecidas.

Na Tabela 6.4, são apresentados os pares de categorias atribuídas com a etiqueta errada e suas respectivas categorias corretas, que tiveram frequência superior a 0,6%, em média, na lista de erros dos conjuntos TE_{VC}. Além dos pares listados na Tabela 6.4, houve 297 outros pares de categorias contendo erros menos frequentes.

TABELA 6.4 - Categorias mais frequentes em erros de etiquetagem no *Corpus* Mac-Morpho

Etiqueta incorreta> Etiqueta correta	**# erros em média**	**%**	**Etiqueta incorreta> Etiqueta correta**	**# erros em média**	**%**
ADJ>N	47	5,08	N>PCP	10	1,10
N>ADJ	41	4,45	PCP>N	10	1,06
N>NPROP	41	4,42	ART>PROSUB	10	1,03
NPROP>N	41	4,38	N>ADV	9	1,00
V>VAUX	36	3,84	PRO-KS-REL>PRO-KS	9	0,98
PREP>PREP|+	33	3,58	PREP|+>ADV	9	0,96
ART>PREP	32	3,45	NPROP>PREP	8	0,91
VAUX>V	31	3,31	PREP>KS	8	0,90
PRO-KS-REL>KS	23	2,50	PRO-KS>PROSUB	8	0,87
PREP>ADV	21	2,22	V>N	8	0,85
PREP>ART	20	2,19	N|AP>N	8	0,82
KS>PRO-KS-REL	18	1,92	KS>ADV	8	0,81
PREP|+>PREP	16	1,69	N>NUM	7	0,79
NUM>N	14	1,55	PREP>PDEN	7	0,77
ADV>KC	14	1,46	PROADJ>PROSUB	7	0,73
N>V	13	1,44	NPROP>ADJ	7	0,70
ADV>PDEN	13	1,36	PRO-KS>PRO-KS-REL	7	0,70
ART>NUM	12	1,30	KC>NPROP	6	0,65
PREP>NPROP	11	1,17	ART>NPROP	6	0,65
PROSUB>PRO-KS	11	1,15	PREP>N	6	0,62
PDEN>ADV	10	1,10	ADJ>ADV	6	0,62

6.2.1.3 Análise dos erros de etiquetagem do Experimento 1

A análise dos erros do Experimento 1 mostrou que o *corpus* de treinamento precisava de revisão. Para realizar essa tarefa, um método semi-automático foi usado para identificar ruídos ou inconsistências no *corpus*. No primeiro estágio, foi examinado um arquivo *html* gerado pelo μ-TBL que lista os erros de etiquetagem que permaneceram no resultado final do processo. Considerando-se o exemplo mostrado na Figura 6.2, em que se vê uma seção desse arquivo, observa-se que os erros listados são falsos erros, uma vez que a palavra "*a*" não é um artigo (ART) e sim uma preposição (PREP) nessas sentenças. Esse tipo de problema ocorreu porque as palavras estavam etiquetadas incorretamente no *corpus* de treinamento. Para resolver esse problema, foram examinadas todas as ocorrências da palavra "*a*" alinhadas em uma coluna de um gerenciador de banco de dados no qual as sentenças do *corpus* foram armazenadas e exploradas em um contexto de tamanho 7. Posteriormente, as anotações dessas ocorrências problemáticas foram corrigidas no *corpus*.

FIGURA 6.2 - Falsos erros visualizados em um arquivo gerado pelo μ-TBL

20 occurrences tagged as PREP that should be ART:

8988:	ele , também terão direito **a**	voz o advogado Luiz Eduardo
28838:	a Casa Branca está conectada **a**	essa rede , disponibilizando publicações
47172:	tais interações eram consideradas fracas **a**	ponto de poderem ser ignoradas
46344:	Sabemos que , de Platão **a**	Bataille , de Sade a

…..

Outros problemas são, por exemplo, os casos mostrados na Figura 6.3, nos quais as palavras do termo "*a respeito de*" são etiquetadas de maneiras diferentes, ora como PREP, ora como PREP|+ ou como N. Na documentação do *corpus*, relatou-se que esse foi dividido e revisado por especialistas diferentes, o que pode ter causado esse problema. Neste trabalho, resolveu-se não modificar o *corpus* de treinamento, mas sim construir regras intermediárias para serem aplicadas na etiquetagem de textos procurando considerar os padrões mais evidentes do *corpus* de treinamento e também estabelecer regras de comparação do modelo com o resultado que contemplassem alguns casos evitando marcá-los como erros.

FIGURA 6.3 - Casos em que as palavras do termo "a respeito de" tem diferentes etiquetas

palav1	**etiq1**	**palav2**	**etiq2**	**palav3**	**etiq3**	**palav4**	**etiq4**	**palav5**	**etiq5**
tendenciosas	ADJ	a	PREP	respeito	N	de	PREP|+	a	ART
.	.	A	PREP	respeito	N	de	PREP|+	a	ART
jogador	N	a	PREP	respeito	PREP	de	PREP	sua	PROADJ
controvérsia	N	a	PREP|+	respeito	PREP|+	de	PREP|+	o	ART
interrogado	PCP	a	PREP|+	respeito	PREP|+	de	PREP|+	a	ART

Muitos ruídos foram identificados e os mais evidentes foram removidos do *corpus* de treinamento. Além dos falsos erros em diversas categorias de palavras, outros problemas identificados na lista de erros foram, por exemplo, palavras estrangeiras que deveriam receber as etiquetas ADJ|EST, ADV|EST, N|EST ou NPROP|EST e foram etiquetadas como ADJ, ADV, N ou NPROP, respectivamente; acronismos ou abreviaturas que foram algumas vezes etiquetadas como N e outras vezes como NPROP; e palavras que deveriam ter sido etiquetadas como símbolo de moeda corrente (CUR) e foram etiquetadas como N ou NPROP.

Em resumo, padrões estruturais de baixa frequência, ruídos e casos em que o etiquetador requer informações que estão além do nível morfossintático da linguagem são as causas de erros mais comuns.

6.2.1.4 Experimentos 2 e 3

No Experimento 2, foi realizada a etiquetagem híbrida incluindo duas estratégias – consulta ao léxico de nomes próprios e etiquetagem com os conjuntos de regras adicionais – no conjuntos de teste TE do *Corpus* Mac-Morpho. Devido à redução do número de erros de etiquetagem após a aplicação das regras intermediárias, o tempo de aprendizado automático das regras pelo µ-TBL foi reduzido de 09 h 58min para 08 h 32 min e foram extraídas 757 regras.

Nos erros que permaneceram, encontrou-se que N etiquetado como NPROP (7,45%), N etiquetado como ADJ (5.99%), ADJ etiquetado como N (5.68%), VAUX etiquetado como V (4.92%), V etiquetado como VAUX (4.76%), and NPROP etiquetado como N (4.22%) foram os erros mais frequentes. Eles representaram 33,02% dos erros, enquanto que os outros erros estão distribuídos em entre várias categorias, com frequências abaixo de 2,3%.

No Experimento 3, foi realizada a etiquetagem híbrida incluindo três estratégias – consulta ao léxico de nomes próprios, etiquetagem com os conjuntos de regras adicionais e o uso de um léxico de grande tamanho – no conjuntos de teste TE do *Corpus* Mac-Morpho. Do Experimento 2 para este experimento, o uso de um léxico de grande tamanho para treinamento do TreeTagger melhorou a acurácia global em 0,66%. O ganho foi mais significativo na etapa probabilística, na qual a acurácia aumentou em 0,99%. O sistema µ-TBL extraiu 764 regras em 09 h 04 min.

Nos erros restantes, encontrou-se que V etiquetado como VAUX (6,50%), N etiquetado como NPROP (5,47%), N etiquetado como ADJ (5,26%), NPROP etiquetado como N (4,64%), ADJ etiquetado como N (3,92%), VAUX etiquetado como V (3,72%), ADV etiquetado como PREP (2,89%), e KS etiquetado como PREP (2,68%) foram os erros mais frequentes. Eles representaram 35,08% dos erros, enquanto os outros errors estão distribuídos dentre várias categorias com frequências menores do que 2,5%.

Na Tabela 6.5, apresenta-se os resultados médios dos Experimentos 2 e 3 sobre o conjunto de teste não visto TE do *Corpus* Mac-Morpho.

TABELA 6.5 - Desempenho da etiquetagem híbrida sobre o conjunto de teste TE do Mac-Morpho após as estratégias propostas terem sido incluídas

Item	**Etiquet. Probab.**	**Léxico Nomes Próprios**	**Etiquetagem baseada em regras**		
			Interm	**µ-TBL**	**Pós**
Tamanho do conj. de treinamento	446.676	-	-	446.676	-
Tamanho do conj. de teste	49.605	49.605	49.605	49.605	49.605
# palavras desconhecidas no conj. de teste	2.808	2.808	2.808	2.808	2.808
	Experimento 2 Léxico do TreeTagger com 35.000 entradas léxicas				
# palavras desconhecidas no conj. de erros	387	319	260	258	181
# erros no conj. de teste	2.472	2.394	2.088	1.531	1.241
% acurácia global	**95,02**	**95,17**	**95,79**	**96,91**	**97,50**
% acurácia em palav. conhecidas	95,54	95,57	96,09	97,28	97,73
% acurácia em palav. desconhecidas	86,22	88,64	90,74	90,81	93,55
	Experimento 3 Léxico do TreeTagger com 379.000 entradas léxicas				
# palavras desconhecidas no conj. de erros	56	45	43	44	38
# erros no conj. de teste	2.000	1.980	1.767	1.216	969
% acurácia global	**95,97**	**96,01**	**96,44**	**97,55**	**98,05**
% acurácia em palav. conhecidas	95,85	95,87	96,32	97,50	98,01
% acurácia em palav. desconhecidas	98,01	98,40	98,47	98,43	98,65

6.2.2 Etiquetagem do Bosque CF 7.4

6.2.2.1 Método de avaliação

Nesta Seção, o etiquetador treinado com o *corpus* Mac-Morpho usado na Seção anterior para etiquetar TE é usado para etiquetar o Bosque CF 7.4. Foram realizados dois experimentos descritos a seguir.

6.2.2.2 Experimentos 4 e 5

No Experimento 4, foi realizada a etiquetagem híbrida incluindo duas estratégias – consulta ao léxico de nomes próprios e etiquetagem com os conjuntos de regras adicionais – no conjuntos de teste formado pelo Bosque CF 7.4. Em virtude dos conjuntos de etiquetas dos *corpora* Mac-Morpho e Bosque CF 7.4 serem diferentes, para comparar os resultados foi necessário mapear o conjunto de etiquetas do Bosque CF 7.4 para o conjunto de etiquetas do Mac-Morpho.

Nos erros restantes, encontrou-se que ADV etiquetado como PREP (16,33%), N etiquetado como ADJ (11,34%), ADV etiquetado como KS (9,58%), ART etiquetado como PREP (5,67%), PREP etiquetado como ART (5,20%), PROPESS etiquetado como KS (4,39%), e ADV etiquetado como ADJ (4,18%) foram os erros mais frequentes. Eles representaram 56,69% dos erros, enquanto os outros erros estão distribuídos dentre várias categorias com frequências abaixo de 4%.

No Experimento 5, foi realizada a etiquetagem híbrida incluindo três estratégias – consulta ao léxico de nomes próprios, etiquetagem com os conjuntos de regras adicionais e o uso de um léxico de grande tamanho – no conjunto de teste formado pelo Bosque CF 7.4.

Nos erros restantes, encontrou-se que ADV etiquetado como PREP (17,47%), N etiquetado como ADJ (9,92%), ADV etiquetado como KS (9,92%), PROPESS etiquetado como KS (5,68%), PREP etiquetado como ART (5,54%), and ART etiquetado como PREP (4,96%) foram os erros mais frequentes. Eles representaram 53,49% dos erros, enquanto os outros erros estão distribuídos dentre várias categorias com frequências abaixo de 4%.

Na Tabela 6.6, apresenta-se os resultados médios dos Experimentos 4 e 5 sobre o conjunto de teste formado pelo Bosque CF 7.4.

TABELA 6.6 - Desempenho da etiquetagem híbrida no Bosque CETENFolha 7.4 após as estratégias propostas terem sido incluídas

Item	Etiquet. Probab.	Léxico Nomes Próprios	Etiquetagem baseada em regras		
			Interm	µ-TBL	Pós
Tamanho do conj. de treinamento	446.676	-	-	446.676	-
Tamanho do conj. de teste	80.522	80.522	80.522	80.522	80.522
# palavras desconhecidas no conj. de teste	6.609	6.609	6.609	6.609	6.609
	Experimento 4 Léxico do TreeTagger com 35.000 entradas léxicas				
# palavras desconhecidas no conj. de erros	965	965	964	960	766
# erros no conj. de teste	2.441	2.441	2.421	2.220	1.963
% acurácia global	**96,97**	**96,97**	**96,99**	**97,24**	**97,56**
% acurácia em palav. conhecidas	98,00	98,00	98,03	98,30	98,38
% acurácia em palav. desconhecidas	85,40	85,40	85,41	85,47	88,41
	Experimento 5 Léxico do TreeTagger com 379.000 entradas léxicas				
# palavras desconhecidas no conj. de erros	309	309	309	309	261
# erros no conj. de teste	1.708	1.708	1.691	1.498	1.393
% acurácia global	**97,88**	**97,88**	**97,90**	**98,14**	**98,27**
% acurácia em palav. conhecidas	98,11	98,11	98,13	98,39	98,47
% acurácia em palav. desconhecidas	95,32	95,32	95,32	95,32	96,05

6.2.2.3 Discussão

Para o português brasileiro, o estado da arte dos etiquetadores tem atingido acurácia abaixo de 97%. Pesquisa correlata foi apresentada por Finger (2000), que também propôs a adição de regras manuais à abordagem TBL de Brill para melhorar a acurácia na etiquetagem morfossintática do Tycho Brahe, um *corpus* histórico do Português, e atingiu acurácia global de 95,45%. Outra pesquisa sobre a etiquetagem do português brasileiro que se tem conhecimento foi realizada por Santos, Milidiú e Rentería (2008) que propôs um método de aprendizado baseado em transformação guiado por entropia (ETL), uma estratégia de aprendizado de máquina para acelerar o aprendizado baseado em transformação. Além de melhorar o desempenho da etiquetagem, ETL atingiu seu melhor resultado com uma acurácia de 96,75% para o *Corpus* Mac-Morpho. Em relação a essas duas pesquisas, nesta tese, foi obtida uma melhoria significativa na acurácia de etiquetagem baseada no *corpus* Mac-Morpho. Além disso, são focados os problemas que emergiram durante o processo de etiquetagem, relacionados aos métodos aplicados, ao *corpus* de treinamento e à linguagem.

Os valores de acurácia obtidos nos diferentes trabalhos não devem ser comparados diretamente, uma vez que usam diferentes conjuntos de treinamento e teste, bem como diferentes conjuntos de etiquetas. Usou-se um método estatístico baseado no processo de Bernoulli (WITTEN; FRANK, 2005, p. 150) para se obter uma estimativa sobre as taxas de acurácia alcançadas e, dessa forma, comparar a acurácia alcançada por Santos, Milidiú e Rentería (2008), que usaram conjunto de treinamento de 1 milhão de palavras, conjunto de teste de 200.000 palavras e conjunto de 22 etiquetas do *corpus* Mac-Morpho, com a acurácia obtida nesta pesquisa (98,05%), que usou conjunto de treinamento revisado de 446.676 palavras, conjunto de teste de 49.605 palavras e conjunto de 78 etiquetas do mesmo *corpus*. Como resultado, tem-se que, com um nível de confiança de 98%, o primeiro alcança acurácia em um intervalo de 96,66% a 96,84%, enquanto que, nesta pesquisa, a acurácia se encontra entre 97,90% a 98,19%. A abordagem TBL foi explorada aqui em uma combinação de ferramentas que usam métodos bem sucedidos de etiquetagem desenvolvidos ao longo dos anos e aumenta seus desempenhos com as três estratégias propostas, tentando extrair benefícios dos recursos disponíveis para a língua portuguesa. Os resultados obtidos são satisfatórios, uma vez que elevam os valores de acurácia do estado da arte e esclareceram muitos problemas que exigem atenção em uma aplicação baseada em *corpus*.

Nas Figuras 6.4 e 6.5, são ilustrados, graficamente, os resultados finais dos Experimentos 2 e 3 e dos Experimentos 4 e 5, respectivamente, sobre os três conjuntos de testes usados: conjuntos de teste TE_{VC}, conjunto de teste TE do Mac-Morpho e conjunto de teste do Bosque CF 7.4.

FIGURA 6.4 - Acurácia média de etiquetagem após a aplicação das estratégias de consulta ao léxico de nomes próprios e aplicação de regras adicionais sobre os conjuntos de testes TE_{VC} e TE do Mac-Morpho (MM) e conjunto de teste formado pelo Bosque CF 7.4 (CF 7.4)

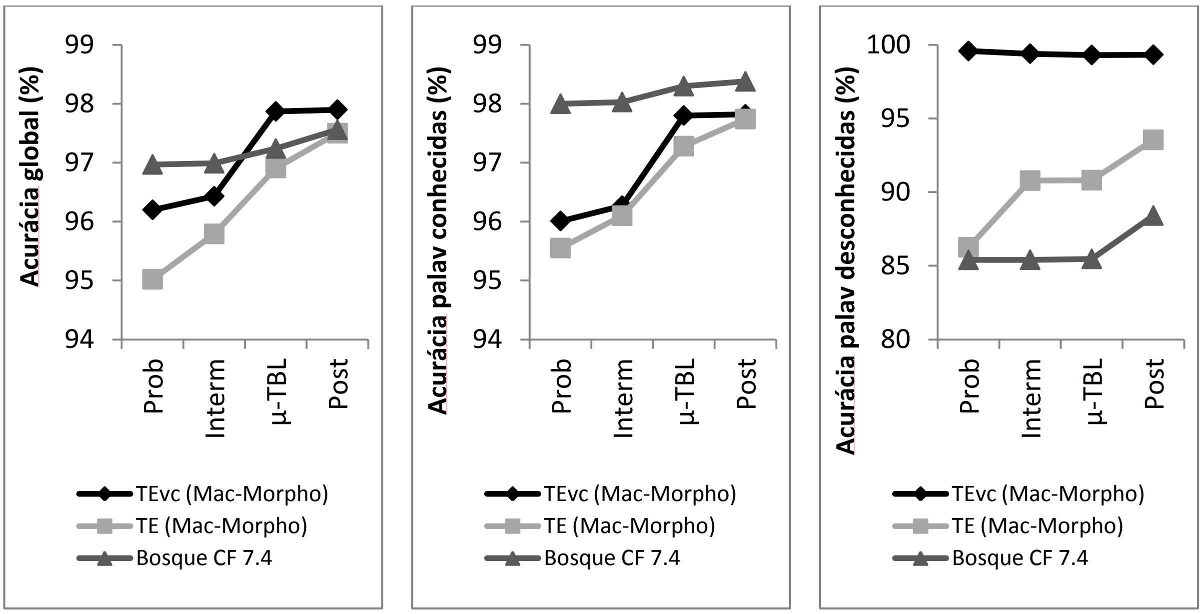
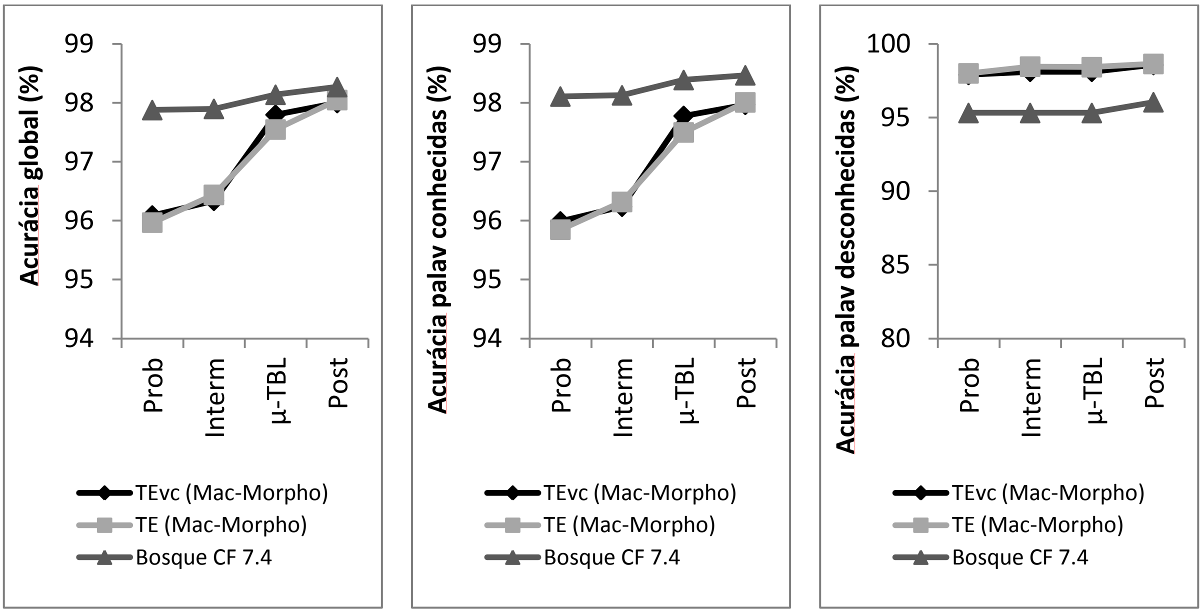


FIGURA 6.5 - Acurácia média de etiquetagem após a aplicação das estratégias de consulta ao léxico de nomes próprios, aplicação de regras adicionais sobre os conjuntos de teste e uso de um léxico de grande tamanho sobre os conjuntos de testes TE_{VC} e TE do Mac-Morpho (MM) e conjunto de teste formado pelo Bosque CF 7.4 (CF 7.4)



6.2.3 Avaliação dos Resultados da Etiquetagem Baseada no *Corpus* Mac-Morpho

Com os resultados aqui relatados, são discutidas as seguintes questões:

1) Com relação à abordagem híbrida, realizada em "*pipeline*" com as etapas probabilísticas e baseada em regras, qual o desempenho do etiquetador em cada uma delas?

   Considerando-se o conjunto de teste TE (não visto no treinamento) do Mac-Morpho e o conjunto de teste formado pelo Bosque CF 7.4, respectivamente, após as três estratégias terem sido aplicadas, o primeiro alcançou acurácia de 95,97% na etapa probabilística e 98,05% após a etapa baseada em regras, o último alcançou acurácia de 97,88% na etapa probabilística e 98,27% após a etapa baseada em regras. As Tabelas 6.5 e 6.6 apresentam os resultados com mais detalhes.

   No Mac-Morpho, o ganho de acurácia de uma etapa para a outra foi de 2,08%, enquanto no Bosque CF 7.4, o ganho foi de 0,39%, o que mostra que o Mac-Morpho requereu mais o uso das regras para atingir a acurácia de 98%.

Comparando o desempenho dos dois *corpora* na etapa probabilística, a acurácia para o Bosque CF 7.4 foi 1,91% mais elevada do que a atingida para o Mac-Morpho. No Bosque CF 7.4, foram observados padrões mais regulares, pouco ruído e um pequeno conjunto de etiquetas, itens que influenciaram a acurácia probabilística ter sido mais elevada.

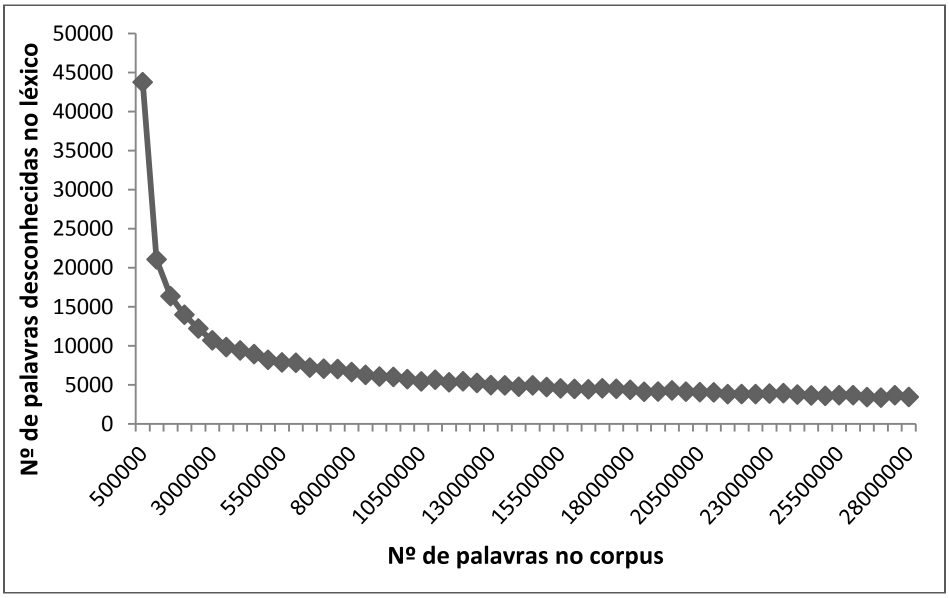
2) Qual a influência do tamanho do *corpus* de treinamento na acurácia de etiquetagem de textos em português?

Foi observado, pela comparação da etiquetagem do conjunto TE e dos conjuntos de testes TE_{VC} do Mac-Morpho, que a adição de maior quantidade de dados ao conjunto de treinamento influencia a acurácia da etiquetagem, uma vez que melhora o arquivo de parâmetros do etiquetador probabilístico por aumentar o léxico e por adicionar mais padrões estruturais da linguagem. Essa ação reduz o número de palavras desconhecidas e faz com que ocorrências esparsas de uma palavra se tornem mais frequentes, aumentando as probabilidades lexiciais de etiquetar uma palavra corretamente.

3) Qual a influência das palavras desconhecidas na acurácia da etiquetagem?

A percentagem de palavras desconhecidas nos conjuntos de testes TE_{VC} e TE do Mac-Morpho foram similares (5,25% e 5,96%, respectivamente) na etapa probabilística. Considerando-se o exemplo do Experimento 2, no qual a diferença é de que nos conjuntos de testes TE_{VC} as palavras desconhecidas não estavam presentes no conjunto de treinamento, mas estavam presentes no léxico do TreeTagger, enquanto que no conjunto de teste TE, as palavras desconhecidas não estavam nem no conjunto de treinamento e nem no léxico. Isto levou o etiquetador a terminar com 16 erros (0,036%) em média para a etiquetagem de palavras desconhecidas nos TE_{VC}, enquanto que em TE, havia 181 erros (0,36%). Portanto, a influência da acurácia das palavras desconhecidas para o Mac-Morpho foi um ganho de 1,18% para os TE_{VC} na etapa probabilística e 0,4% após a etapa baseada em regras, quando não havia palavras desconhecidas no léxico. No Experimento 3, com o uso de um grande léxico no treinamento do TreeTagger, a influência da acurácia das palavras desconhecidas para o Mac-Morpho foi um ganho de 0,12% para os TE_{VC} na etapa probabilística e um ganho de 0,05% para TE após a etapa baseada em regras quando não havia palavras desconhecidas no léxico. Assim, a acurácia para TE foi melhor após a etapa baseada em regras do que para os TE_{VC}. Desta forma, a influência das palavras desconhecidas foi significativamente reduzida em termos de acurácia de etiquetagem. Aumentar o léxico causou ganhos de 5,1% e 7,64% na acurácia de palavras desconhecidas para o Mac-Morpho e o Bosque CF 7.4, respectivamente.

FIGURA 6.6 - Número de novas palavras no léxico a cada acréscimo de 500 mil palavras



4) A que tamanho de *corpus* a acurácia das palavras desconhecidas não é mais relevante?

   Para responder a esta pergunta, usou-se o *Corpus* CETENFolha, que tem cerca de 28 milhões de palavras, e identificou-se as suas palavras diferentes com as suas etiquetas potenciais, que formaram um léxico contendo 378.019 palavras diferentes. Então, foi observada a distribuição dessas 378.019 palavras a cada 500.000 palavras. Nas primeiras 500.000 palavras, havia 43.789 palavras desconhecidas no *corpus* (11,4 palavras por palavra desconhecida), e nas últimas 500.000, quando o *corpus* continha 28 milhões de palavras, havia 3.461 palavras desconhecidas (144,5 palavras por palavra desconhecida). Este processo é ilustrado na Figura 6.6 e se observa que a curva é mais acentuada até quando o *corpus* possui cerca de 2 milhões de palavras. Acima desse tamanho, o número de palavras desconhecidas é inferior a 0,05% do léxico extraído do *corpus* de 28 milhões de palavras em cada adição de 500.000

palavras. Como foi mostrado aqui, palavras desconhecidas estarão sempre presentes nos sistemas de etiquetagem, mas foi observado neste estudo que os seus efeitos podem ser reduzidos pelo uso de um léxico maior no treinamento do TreeTagger. É importante que os etiquetadores tenham capturado os padrões estruturais da linguagem.

5) Quais questões linguísticas estão relacionadas aos problemas encontrados no processo de etiquetagem morfossintática?

   Foi identificado que os conjuntos de testes do Mac-Morpho e Bosque CF 7.4 apresentaram problemas diferentes no que tange às categorias gramaticais ao se usar o *Corpus* Mac-Morpho para treinar os etiquetadores.

   Nos conjuntos de testes do Mac-Morpho, as ocorrências de maior dificuldade estão relacionadas ao seguinte:

   - No caso de nomes próprios, não se conseguiu resolver os seguintes problemas: erros causados pela confusão de um substantivo no início de uma frase com um nome próprio, os erros causados pela presença de uma palavra no início de uma frase (por exemplo, uma palavra que deve ser etiquetada como um artigo e é erroneamente etiquetada como um substantivo próprio), seguido por um substantivo próprio polilexical, os erros causados quando há preposições ou conjunções entre dois substantivos próprios formados por várias palavras, e os erros causados porque muitas palavras que começam com uma letra maiúscula são marcadas como substantivos e outras como nomes próprios.
   - Os erros causados entre adjetivos e substantivos, nas observações desta pesquisa, são mais difíceis de resolver. O etiquetador probabilístico falha em atribuir a etiqueta correta com base nas probabilidades contextuais em algumas ocorrências, pois extrapolam o nível morfossintático da língua. Por exemplo, na sequência de palavras “o crítico Roberto ...”, a palavra crítico poderia ser etiquetada como um substantivo ou adjetivo no mesmo contexto. Quanto ao desempenho do TBL, observou-se que as regras extraídas são mais lexicalizadas, isto é, se referem a pares específicos de palavras e etiquetas e não apresentam um bom desempenho em palavras desconhecidas ou palavras que têm uma pontuação menor do que a configurada para o algoritmo TBL.

- Foi investigada a desambiguização de nomes, adjetivos e particípio passado pelo uso de análise de sufixo. Para a língua portuguesa, notou-se que o comprimento do sufixo para definir a etiqueta correta varia bastante e deve estar acima de cinco caracteres. Frequentemente, mesmo o tamanho total da palavra é insuficiente. Por exemplo, a palavra “empregado” pode ser tanto um nome quanto um particípio e tem quase a mesma probabilidade no *corpus*. Nesses casos, a palavra é desambiguizada pelo contexto.
- Um problema frequente é relacionado às etiquetas de verbo auxiliar e verbo, mas não foi possível abordar esse problema porque essas categorias contêm ruído.
- Outro problema é desambiguizar a palavra “a” como artigo ou preposição. Em um estudo prévio (Domingues et al., 2008), foi testado o uso de atributos léxicos como gênero, número, pessoa, tempo e modo verbais na configuração das etiquetas, o que melhorou a acurácia da desambiguização de palavras. Correntemente, o *Corpus* Mac-Morpho não contém esse tipo de informação.
- Nos conjuntos de testes do Bosque CETENFolha 7.4, os seguintes pares de categorias foram mais difíceis de desambiguizar: advérbio e preposição (tipicamente a palavra *como*), advérbio e conjunção subordinativa (*quando* e *como* são as mais frequentes), pronome pessoal e conjunção subordinativa (tipicamente a palavra *se*), e similarmente ao Mac-Morpho, nome e adjetivo, e preposição e artigo na desambiguização da palavra *a*.

6) Quanto ruído está presente no *corpus*?

Na análise dos resultados do Experimento 1, foi observado que os erros mais frequentes foram causados por presença de padrões estruturais de baixa frequência, ruído, e dificuldade na etiquetagem devido a desambiguização de uma palavra extrapolar o nível morfossintático da linguagem. Para os casos de baixa frequência, foram aplicadas as três estratégias descritas na Seção 6.1. Os casos fora do nível morfossintático da linguagem serão resolvidos em futuros experimentos. Para os erros restantes, a presença de ruído (ou discordâncias com a documentação do conjunto de etiquetas) impediram que regras fossem codificadas. Esses casos precisam ser revisados com a ajuda de especialistas humanos. A quantificação da presença de ruído é uma questão em aberto, visto que requer muito esforço para resolvê-la.

7) Qual o impacto do ruído?

Apesar de não se ter medido esse valor, é muito importante reduzir o impacto do ruído (tipicamente de 1% para 3%) com vistas a quebrar a barreira dos 97% de acurácia de etiquetagem. Foi observado que uma das principais contribuições para a alta acurácia é o desenvolvimento de um grande *corpus* com baixo nível de ruído. Neste estudo, a presença de ruído nos *corpora* foi o gargalo para uma aplicação de etiquetagem baseada em *corpus* do Português com alta acurácia.

7 Etiquetagem em Textos de Gêneros Diferentes

"... For many NLP tasks, however, we are confronted with new domains in which labeled data is scarce or non-existent. In such cases, we seek to adapt existing models from a resource-rich source domain to a resource-poor target domain. ..."
(BLITZER et al., 2006, p. 120).

Neste capítulo, o etiquetador proposto, que foi, originalmente, desenvolvido para anotar textos jornalísticos, é avaliado na anotação de textos do domínio científico. Buscou-se solucionar o problema da etiquetagem de palavras desconhecidas no domínio alvo, bem como ajustar o etiquetador para obter bom desempenho nos dois gêneros textuais.

7.1 Metodologia dos Experimentos

O método para etiquetar textos de diferentes gêneros com o uso de dados anotados e de dados não anotados é chamado de método semi-supervisionado de adaptação de domínio, conforme se explica em Domingues e Favero (2011). Nos experimentos deste capítulo, os dados do domínio fonte (*Corpus* Mac-Morpho) são dados jornalísticos anotados e os dados do domínio alvo (*Corpus* Selva Científica) são dados científicos não anotados, significando que os padrões sobre as sequências de palavras destes últimos não são usados no treinamento do etiquetador. Apesar dos dados do domínio alvo serem provenientes de um *corpus* anotado, a informação de anotação só é usada com o propósito de comparação entre o modelo e o resultado para medir a acurácia do etiquetador.

A metodologia deste estudo de caso apresenta seis etapas: 1) identificação e remoção de ruídos do *corpus* Selva Científica, 2) etiquetagem do *corpus* Selva, 3) avaliar os erros de etiquetagem e ajustar o etiquetador, 4) etiquetar o Selva depois dos ajustes e reavaliar os erros restantes, 5) etiquetar os *corpora* de textos jornalísticos depois dos ajustes e 6) e avaliar os resultados da etiquetagem em textos de gêneros diferentes. Estas etapas são ainda parte da abordagem para desenvolver um etiquetador de alta acurácia para o português brasileiro.

7.2 Identificação e Remoção de Ruídos do *Corpus* Selva Científica

Por ser um *corpus* anotado automaticamente e parcialmente revisado por especialistas, a presença de ruído era esperada no Selva. Nessa etapa, o *corpus* foi anotado com o etiquetador do estado da arte proposto nesta tese; os erros foram examinados e muitos falsos erros foram identificados. Por exemplo, as palavras “bromo”, “cobre”, “demanda”, “deriva”, “desperta”, dentre outras foram corretamente etiquetadas como nome (n), enquanto no Selva elas estavam etiquetadas como verbo finito (v-fin). Esse tipo de erro foi corrigido no Selva para que se pudesse ter um *corpus* de boa qualidade para a etapa de métrica da acurácia do etiquetador. Este processo levou cerca de cinco meses.

7.3 Etiquetagem do Corpus Selva Científica

7.3.1 Método de Avaliação

Nos experimentos deste estudo, o etiquetador foi treinado com os conjuntos de parâmetros de 402.008 palavras em média do *Corpus* Mac-Morpho (vide Seção 6.2.1.1.1). Foi realizado o método de validação cruzada com os 10 arquivos de parâmetros, cada um extraído de diferentes subconjuntos de treinamento contendo 90% das palavras de TR. O Selva Científica foi então etiquetado em 10 iterações, cada uma utilizando um dos 10 arquivos de parâmetros. Ao final, foi calculada a acurácia média das 10 iterações do processo de etiquetagem.

7.3.2 Experimentos 1 e 2

Para os experimentos com o Selva Científica, foi tomada como linha básica a taxa de acurácia alcançada no Experimento 3 da etiquetagem com o *Corpus* Mac-Morpho (Capítulo 5), de 98,05% no conjunto de teste do Mac-Morpho.

No Experimento 1, o *corpus* de teste formado pelas 141.361 palavras do Selva Científica foi etiquetado com a versão corrente do etiquetador, que usa no treinamento um léxico do TreeTagger de 379.335 palavras.

No Experimento 2, o *corpus* de teste formado por 141.361 palavras do Selva Científica foi etiquetado com uma versão atualizada do etiquetador. Nessa versão, para minimizar os erros de etiquetagem em palavras desconhecidas, a solução foi baseada na estratégia que tira proveito do léxico usado no treinamento do TreeTagger. Nesse léxico foram inseridas as palavras do Selva e suas etiquetas possíveis (obtidas de uma etiquetagem prévia) no léxico do TreeTagger, que passou a conter 422.411 entradas léxicas. O etiquetador foi retreinado, na etapa probabilística do modelo, com o novo arquivo de parâmetros do TreeTagger e, na etapa baseada em regras, as regras já existentes foram aplicadas para a correção de erros. Na Tabela 7.1, são mostrados os resultados dos Experimentos 1 e 2.

TABELA 7.1 - Desempenho do etiquetador em textos do *Corpus* Selva Científica

Item	**Etiquet. Probab.**	**Léxico Nomes Próprios**	**Etiquetagem baseada em regras**		
			Interm	**µ-TBL**	**Pós**
Tamanho do conj. de treinamento	446.676	-	-	446.676	-
Tamanho do conj. de teste	141.361	141.361	141.361	141.361	141.361
# palavras desconhecidas no conj. de teste	21.849	21.849	21.849	21.849	21.849
	Experimento 1 Léxico do TreeTagger com 379.335 entradas léxicas				
# palavras desconhecidas no conj. de erros	1.882	1.881	1.513	1.516	1.481
# erros no conj. de teste	6.144	6.140	5.226	5.192	5.076
% acurácia global	**95,65**	**95,66**	**96,30**	**96,33**	**96,41**
% acurácia em palav. conhecidas	96,43	96,44	96,89	96,92	96,99
% acurácia em palav. desconhecidas	91,38	91,39	93,07	93,06	93,22
	Experimento 2 Léxico do TreeTagger com 422.411 entradas léxicas				
# palavras desconhecidas no conj. de erros	723	722	389	412	405
# erros no conj. de teste	4.940	4.936	4.053	4.036	3.946
% acurácia global	**96,51**	**96,51**	**97,13**	**97,15**	**97,21**
% acurácia em palav. conhecidas	96,47	96,47	96,93	96,97	97,04
% acurácia em palav. desconhecidas	96,69	96,70	98,22	98,11	98,15

7.4 Avaliação dos Erros de Etiquetagem e Ajustes no Etiquetador

Em virtude das taxas de acurácia global dos Experimentos 1 e 2 ficarem abaixo da já estabelecida como linha básica, os erros da etiquetagem foram analisados para verificar a sua causa. Os principais problemas encontrados são apresentados e exemplificados a seguir, seguidos da solução utilizada para resolvê-los:

- diferenças entre o *corpus* de treinamento e o *corpus* de teste na itemização (*tokenization*) de palavras. Por exemplo, a expressão “todos os” forma um *token* (todos_os) no *Corpus* Selva e é etiquetada como pronome determinativo (pron-det), enquanto no Mac-Morpho, essas duas palavras são etiquetadas separadamente (todos/PROADJ os/ART). Isto fez com que algumas regras extraídas pelo µ-TBL atribuíssem a etiqueta errada às palavras. Foram incluídas exceções às regras para os *tokens* que estavam causando esse problema;

- erros na lista de etiquetas possíveis para algumas palavras presentes no léxico do TreeTagger causaram a presença de ruídos no *corpus*. Por exemplo, a palavra “largo” tinha NPROP como uma etiqueta possível, o que causava erros. A etiqueta NPROP foi eliminada da lista de etiquetas possíveis para “largo.” Se essa palavra for considerada um nome próprio no contexto de um título de livro, por exemplo, em que algumas palavras iniciam com letras minúsculas, uma regra apropriada para esses casos foi codificada;
- o pronome pessoal “se” etiquetado como conjunção subordinativa aparecia como o erro de etiquetagem mais frequente nos resultados dos Experimentos 1 e 2. No Mac-Morpho, o verbo antes de um pronome pessoal separado por hífen, por exemplo, “utiliza-se”, é itemizado sem o hífen (utiliza/V|+ se/PROPESS), o que cria dificuldades na desambiguização da palavra “se”. Uma regra foi inicialmente codificada apenas para verbos que ocorreram no Mac-Morpho. No Selva, o verbo é itemizado incluindo o hífen (utiliza-/v-fin se/pron-pers). Isto permitiu a codificação de uma regra de pós-correção mais abrangente que inspeciona a letra final da palavra seguida por um hífen, como é mostrado na Figura 7.1. Esse ajuste na regra reduziu os erros com o pronome “se”, significativamente, de 342 para 12 ocorrências;

FIGURA 7.1 - Regra para corrigir a etiqueta para o pronome pessoal “se”

```
/*Regra: Corrigir a etiqueta para o pronome pessoal “se”
 */
SE a palavra corrente termina com: a-, e- (e é diferente
   de se-), i-, o-, u-, m-,r-, s-, z-,á-, ê-,
SE a próxima palavra é se ou se-,
ENTÃO as palavras se ou se- são etiquetadas como um
   pronome pessoal PROPESS.
```

- Muitos erros no *Corpus* Selva foram causados pela ausência de mapeamento entre os conjuntos de etiquetas dos *corpora* envolvidos para algumas etiquetas, já que não foi possível identificar esse mapeamento nos experimentos com o Mac-Morpho e CETENFolha. Por exemplo, palavras tais como “isto,” “isso,” etc., são etiquetadas como PROSUB no Mac-Morpho e como “pron-indp” no

Selva. Assim, elas foram etiquetadas como PROSUB e então foram contadas como erros. Foi feita uma modificação nas regras de mapeamento dessas palavras para considerar a etiqueta PROSUB correta, visto que foi adotado o conjunto de etiquetas do Mac-Morpho como modelo.

7.5 Etiquetagem do Selva Científica depois dos Ajustes

7.5.1 Experimentos 3 e 4

No Experimento 3, o etiquetador foi retreinado com o léxico de 379.335 palavras, em virtude dos ajustes descritos na Seção 7.4. A acurácia foi recalculada para a etiquetagem do *corpus* de teste Selva Científica. Os resultados são mostrados na Tabela 7.2 e indicam uma melhora na acurácia da etiquetagem em textos científicos.

No Experimento 4, o etiquetador foi retreinado com o léxico de 422.411 palavras, em virtude dos ajustes descritos na Seção 7.2.2.1. A acurácia foi recalculada para a etiquetagem do *corpus* de teste Selva Científica. Os resultados são mostrados na Tabela 7.2 e indicam uma melhora na acurácia da etiquetagem em textos científicos.

7.5.1.1 Avaliação dos resultados

Os erros que resultam dos Experimentos 3 e 4 mostram os problemas de etiquetagem que ainda não puderam ser resolvidos. Na Tabela 7.3, são apresentados os erros mais frequentes de acordo com as categorias gramaticais. Na Tabela 7.4, de acordo com as palavras mais frequentes.

TABELA 7.2 - Desempenho em textos do *Corpus* Selva Científica após os ajustes no etiquetador

Item	**Etiquet. Probab.**	**Léxico Nomes Próprios**	**Etiquetagem baseada em regras**		
			Interm	**μ-TBL**	**Pós**
Tamanho do conj. de treinamento	446.676	-	-	446.676	-
Tamanho do conj. de teste	141.361	141.361	141.361	141.361	141.361
# palavras desconhecidas no conj. de teste	21.835	21.835	21.835	21.835	21.835
	Experimento 3 Léxico do TreeTagger com 379.335 entradas léxicas				
# palavras desconhecidas no conj. de erros	1.490	1.490	1.121	1.124	1.064
# erros no conj. de teste	5,370	5,370	4.464	4.425	3.532
% acurácia global	**96,20**	**96,20**	**96,84**	**96,87**	**97,50**
% acurácia em palav. conhecidas	96,75	96,75	97,20	97,24	97,93
% acurácia em palav. desconhecidas	93,18	93,18	94,87	94,85	95,13
	Experimento 4 Léxico do TreeTagger com 422.411 entradas léxicas				
# palavras desconhecidas no conj. de erros	660	660	327	345	329
# erros no conj. de teste	4.505	4.505	3.631	3.599	2.746
% acurácia global	**96,81**	**96,81**	**97,04**	**97,45**	**98,07**
% acurácia em palav. conhecidas	96,78	96,78	97,24	97,28	97,98
% acurácia em palav. desconhecidas	96,98	96,98	98,50	98,42	98,49

Os principais obstáculos nos experimentos de etiquetagem, nos quais o etiquetador foi treinado com o *Corpus* Mac-Morpho e testado com o *Corpus* Selva Científica, incluem, primeiro, a desambiguização de nomes e adjetivos. Ao serem comparados o *modelo* e *resultado*, observa-se que apesar do Selva Científica conter a etiqueta "*n-adj*" em seu conjunto de etiquetas, ainda é possível encontrar no corpus, mesmo após a revisão, vários casos onde ocorre dúvida, que se encontram etiquetados como nome ou como adjetivo. Segundo, a etiquetagem da palavra "que" como um pronome relativo ou uma conjunção subordinativa é outro problema crítico que necessita ser explorado de forma mais aprofundada, uma vez que o etiquetador não consegue obter um modelo claro do *corpus* de treinamento (Mac-Morpho) para etiquetar essa palavra corretamente. Terceiro, a classificação da palavra "a" como preposição ou artigo é outro grande problema para o etiquetador. Como foi apresentado na Seção 5.1.2 e nos experimentos da Seção 5.2, a acurácia nesses casos aumenta com a inclusão de maior quantidade de atributos léxicos nas etiquetas tais como gênero, número, pessoa, tempo e modo verbais, no entanto, o *corpus* Mac-Morpho não possui esses atributos.

TABELA 7.3 - Erros mais frequentes por categorias gramaticais

Frequência	Etiqueta incorreta	Etiqueta correta
332	N	ADJ
242	ADJ	N
217	PRO-KS-REL	KS
148	ART	PREP
131	PREP	ART
115	N	NUM
75	ART	PROPESS
70	KS	PRO-KS-REL
64	ART	NUM
59	PRO-KS	ART
56	N	PCP
52	PCP	N
44	KS	PREP
43	V	N
42	PREP	N

TABELA 7.4 - Erros mais frequentes por palavras

Frequência	Palavra	Etiqueta incorreta	Etiqueta correta
217	que	PRO-KS-REL	KS
148	a	ART	PREP
129	a	PREP	ART
70	que	KS	PRO-KS-REL
59	o	PRO-KS	ART
39	como	KS	PREP
39	o	ART	PROPESS
38	um	ART	NUM
30	se	PROPESS	KS

7.6 Etiquetagem do Mac-Morpho e Bosque CF 7.4 depois dos Ajustes

7.6.1 Experimentos 5 e 6

Nos Experimentos 5 e 6, o etiquetador foi reavaliado para os textos jornalísticos do Mac-Morpho (49.605 palavras) e do Bosque CF 7.4 (80.522 palavras), respectivamente, após os ajustes nos experimentos com o Selva. Foi usada a versão do etiquetador treinado com um léxico do TreeTagger de 422.411 palavras (incluindo as palavras do Selva). Nos resultados, há uma pequena melhora na taxa de acurácia em ambos os *corpora*, conforme se observa na Tabela 7.5.

TABELA 7.5 - Desempenho em textos do *Corpus* Mac-Morpho e Bosque CF 7.4 após os ajustes no etiquetador

Exp.	# palavras desconhecidas no conj. de teste	# palavras desconhecidas no conj. de erros	# erros no conj. de teste	Acurácia global (%)	Acurácia em palavras conhecidas (%)	Acurácia em palavras desconhecidas (%)
5	2.957	38	961	**98,06**	98,00	99,00
6	6.823	210	1.367	**98,30**	98,43	96,93

7.7 Avaliação dos Resultados da Etiquetagem em Textos de Gêneros Diferentes

Neste capítulo, foi avaliada a proposta de um modelo híbrido de etiquetagem de alta precisão em textos de dois gêneros diferentes: científico e jornalístico. A avaliação foi feita em duas etapas. Em primeiro lugar, anotou-se o conjunto de textos científicos com o etiquetador proposto em seu estado da arte, que alcançou acurácia de 97,21%, inferior à acurácia alcançada para os textos jornalísticos (acima de 98%). A avaliação dos erros de etiquetagem mostrou que a maioria das ocorrências foram causadas por: presença de palavras desconhecidas, que foi reduzida quando estas palavras foram incluídas no léxico do TreeTagger; pelas diferenças entre *corpora* na itemização de palavras; pela presença de ruído no léxico de treinamento; pela presença de novas estruturas linguísticas envolvendo palavras cujas etiquetas não tinham sido mapeadas de um *corpus* para outro e apareciam como falsos erros. Para resolver esses problemas, foram feitas modificações para melhorar a acurácia de etiquetagem em textos do Português, incluindo o ajuste na itemização de palavras nos textos de entrada, eliminação do ruído no léxico usado pelo TreeTagger, refinamento e adição de regras codificadas manualmente e revisão e ajuste do mapeamento das etiquetas dos vários *corpora* envolvidos no processo de etiquetagem. Em segundo lugar, após essas modificações para o gênero científico, os conjuntos de testes dos dois gêneros foram reavaliados. Os resultados foram positivos, com taxas de acerto de 98,07% para o Selva, 98,06% para o Mac-Morpho e 98,30% para o Bosque CF 7.4. O estudo mostra que esta abordagem combinada, guiada pelas questões críticas e os ajustes adicionais apresentados aqui, funciona bem tanto para o gênero científico quanto para o gênero jornalístico. Este resultado levou-nos a avançar cerca de 1 a 2% de acurácia em relação à acurácia dos etiquetadores do estado da arte (96-97% em média para o Português), validando assim o método proposto para o desenvolvimento de uma ferramenta mais confiável.

8 Considerações Finais

Nesta tese, foi avaliada uma proposta de um modelo híbrido de etiquetagem de alta precisão levando-se em consideração as questões críticas: 1) tamanho do *corpus* de treinamento, 2) presença de ruído, 3) conjunto de etiquetas, 4) gênero textual, 5) abordagem de etiquetagem e 6) presença de palavras desconhecidas nos textos a serem etiquetados.

Dentre as estratégias propostas e avaliadas, três estratégias foram adaptadas em uma abordagem baseada em TBL para melhorar a acurácia na etiquetagem do Português brasileiro: o uso de um léxico de nomes próprios, a adição de regras codificadas manualmente e o uso de um léxico de grande tamanho para treinamento do TreeTagger, combinadas em uma arquitetura que integra as saídas (textos etiquetados) de duas ferramentas de uso livre, o TreeTagger e o μ-TBL, com os módulos adicionados ao modelo. Foram etiquetados três conjuntos de testes – conjunto de teste do *Corpus* Mac-Morpho, Bosque CF 7.4 e Selva Científica – com acurácia global de 98,06%, 98,30% e 98,07%, respectivamente. A acurácia em palavras desconhecidas alcançou seu valor mais alto de 99% no conjunto de teste do Mac-Morpho após os ajustes no etiquetador nos experimentos com o Selva Científica. Esses resultados são bastante satisfatórios para as aplicações de etiquetagem do estado da arte.

Os principais problemas observados nos experimentos foram causados pela presença de ruído no *corpus* de treinamento e pela dificuldade na desambiguação de nomes próprios, nomes, adjetivos e palavras desconhecidas. Por outro lado, conclui-se que: é importante usar um etiquetador probabilístico robusto na etapa inicial do método TBL, que permita que sejam adicionadas palavras em seu arquivo léxico e que possa ser retreinado com facilidade; a visualização de erros e a análise desses erros em seus contextos são muito úteis para detectar os problemas de etiquetagem; o tamanho do *corpus* tem uma grande influência na melhoria da acurácia do etiquetador; mesmo que se tenha um *corpus* muito grande para treinar o etiquetador, as palavras desconhecidas sempre estarão presentes, mas etiquetadores probabilísticos que extraem informações de um léxico de palavras e suas etiquetas possíveis podem reduzir a influência desse problema. Grandes *corpora*, anotados automaticamente, mesmo contendo ruído, podem ser bastante úteis para fornecer este léxico porque na maioria das ocorrências, as palavras são etiquetadas com suas etiquetas possíveis. Estas conclusões são relevantes porque representam passos em direção a uma acurácia de 99%. Esta tese propõe um modelo de etiquetador para o Português brasileiro, mas sua arquitetura pode ser adaptada para outras línguas.

Para trabalhos futuros, é importante desenvolver um recurso primário para a etiquetagem morfossintática de textos: um grande *corpus* (de um a dois milhões de palavras) com o mínimo de ruído possível. Propõe-se também que o etiquetador seja avaliado em textos de outros gêneros, por exemplo, textos de gêneros diversificados disponíveis na *Web*, com o uso de técnicas de adaptação de domínio, para tornar o etiquetador cada vez mais preciso em domínios irrestritos.

Referências

AFONSO, S. **Árvores deitadas**: descrição do formato e das opções de análise na Floresta Sintá(c)tica. 2006. Disponível em: <http://www.linguateca.pt/documentos/Afonso2006ArvoresDeitadas.pdf>. Acesso em: 12 mar. 2006.

AIRES, R. V. X. **Results in NILC´s Taggers**. 2000a. Disponível em: <http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html>. Acesso em: 02 mai 2007.

AIRES, R. V. X. **Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil.** São Paulo, 2000b. Tese (Mestrado) - Instituto de Ciências Matemáticas de São Carlos, Universidade de São Paulo.

ALLEN, J. **Natural Language Understanding**. 2nd edition. Redwood City, CA: Benjamin/Cummings Publishing Company, 1995.

ALUÍSIO, S. M.; PELIZZONI, J. M.; MARCHI, A. R.; OLIVEIRA, L. H.; MANENTI, R.; MARQUIAFÁVEL, V. An account of the challenge of tagging a reference corpus of Brazilian Portuguese. In: MAMEDE, N.J.; BAPTISTA, J.; TRANCOSO, I.; NUNES, M.G.V. (Eds.). **Lecture Notes in Computer Science**: v. 2721. Computational Processing of the Portuguese Language. Heidelberg: Springer-Verlag, 2003. p.110-117. DOI: 10.1007/3-540-45011-4_17

BANKO, M.; MOORE, R. C. Part of speech tagging in context. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 20., 2004. **Proceedings**… [S.l. : S.n.], 2004. p.556-561. DOI: 10.3115/1220355.1220435

BELLEGARDA, J. R. Part-of-Speech Tagging by Latent Analogy. **IEEE Journal of Selected Topics in Signal Processing**, v.4, n.6, p. 985-993, 2010. DOI: 10.1109/JSTSP.2010.2075970

BERBER SARDINHA, T. **Linguística de Corpus.** Barueri, SP: Manole, 2004. 410 p.

BICK, E. **The Parsing System PALAVRAS**: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework. [S.l.]: Aarhus University Press, 2000.

BLITZER, J. **Domain adaptation of natural language processing systems**. [S.l.]: ProQuest Digital Dissertations, 2007. (AAT 3309400)

BLITZER, J.; MCDONALD, R.; PEREIRA, F. (2006). Domain adaptation with structural correspondence learning. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2006. **Proceedings…[S.l]**, 120-128.

BRANTS, T. TnT – a statistical part-of-speech tagger. In: APPLIED NATURAL LANGUAGE PROCESSING CONFERENCE ANLP, 16., 2000. **Proceedings…** Seatle, WA: [S.n.], 2000.

BRILL, E. Some advances in transformation-based part of speech tagging. In: PROCEEDINGS OF THE TWELFTH NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI-94). Cambridge, Massachusetts: AAAI Press/MIT Press, 1994.

BRILL, E. Transformation-Based Error-Driven Learning and Natural Language Processing: a case study in part of Speech Tagging. **Computational Linguistics**, v.21, n.4, p. 543-565, 1995.

BRILL, E. Unsupervised learning of disambiguation rules for part of speech tagging. In: **Natural Language Processing Using Very Large Corpora**. Dordrecht: Kluwer Academic Press, 1997. p. 27-42.

BRILL, E.; WU, J. Classifier Combination for Improved Lexical Disambiguation. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL'98). **Proceedings…** Montreal, Canada: [S.n.], 1997.v.1.p.191-195.

CHARNIAK, E.; HENDRICKSON, C.; JACOBSON, N.; PERKOWITZ, M. Equations for part-of-speech tagging". In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 11., 1993. **Proceedings…** [S.l.]: Menlo Park, CA, 1993. p. 784-789.

CHURCH, K. A. Stochastic Parts Program and Noum Phrase Parser for Unrestricted Text. In: CONFERENCE ON APPLIED NATURAL LANGUAGE PROCESSING, 2., 1988. **Proceedings…** [S.l.]: ANLP 88, 1988. p. 136-143.

DAELEMANS, W.; ZAVREL, J.; BERCK, P.; GILLIS, S. MBT: a Memory-Based Part of Speech Tagger-Generator. In: WORKSHOP ON VERY LARGE CORPORA, 14., 1996. **Proceedings…** Copenhagen: ACL SIGDAT, 1996. p.14-27.

DOMINGUES, M. L.; MEDEIROS, I. P.; FAVERO, E. L. O desenvolvimento de um etiquetador morfossintático com alta acurácia para o português. In: TAGNIN, S. E. O.; VALE, O. A. (Eds.). **Avanços da Linguística de Corpus no Brasil.** São Paulo, SP: Humanitas, 2008. p. 267-286.

DOMINGUES, M. L.; FAVERO, E. L. Domain Adaptation in Part-of-Speech Tagging. In: BANDYOPADHYAY, S.; NASKAR, S.K.; EKBAL, A. (Eds.). **Emerging Applications of Natural Language Processing**: Concepts and new research. IGI Global, 2011. (no prelo).

EKBAL, A.; HAQUE, R.; BANDYOPADHYAY, S. Bengali Part of Speech Tagging using Conditional Random Field. In: INTERNATIONAL SYMPOSIUM OF NATURAL LANGUAGE PROCESSING, 7., 2007. **Proceedings…** [S.l : S.n.], 2007. p.131-136.

FINGER, M. Técnicas de otimização da precisão empregadas no etiquetador Tycho Brahe apud NUNES, M.G.V. (Ed.). In: ENCONTRO PARA O PROCESSAMENTO COMPUTACIONAL DA LÍNGUA PORTUGUESA ESCRITA E FALADA, 5., 2000. Atibaia, SP. **Anais**… Atibaia, SP: ICMC/USP, 2000. p. 141-154.

GASPERIN, C.; MAZIERO, E.; ALUÍSIO, S. M. Challenging choices for text simplification. In: PARDO, T. A. S.; BRANCO, A.; KLAUTAU, A.; VIEIRA, R.; LIMA, V. L. S. (Eds.). **Lecture Notes in Computer Science**: v. 6001/2010. Computational Processing of the Portuguese Language. Heidelberg: Springer-Verlag, 2010. p.40-50. DOI: 10.1007/978-3-642-12320-7_6

GIMÉNEZ, J.; MÁRQUEZ, L. SVMTool: a general POS tagger generator based on Support Vector Machines. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC'04), 4., 2004. **Proceedings…** Lisboa, Portugal: [S.n.], 2004.

GREENE, B. B.; RUBIN, G. M. **Automatic grammatical tagging of English**. Technical Report. [S.l.]: Brown University, Providence, RI, 1971.

HALÁCSY, P.; KORNAI, A.; ORAVECZ, C.; TRÓN, V.; VARGA, D. Using a morphological analyzer in high precision POS tagging of Hungarian. In: CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC). ELRA, 5., 2006. **Proceedings…** [S.l. : S.n.], 2006. p. 2245-2248.

HARB, M. P. A. A.; BRITO, S. R.; SILVA, A. S.; FAVERO, E. L.; TAVARES, O. L.; FRANCÊS, C. R. L. AmAm: ambiente de aprendizagem multiparadigmático. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO. Rio de Janeiro: NCE-IM-UFRJ, 2003.

HEIKKILÄ, J. A TWOL-based lexicon and feature system for English. In: KARLSSON, F.; VOUTILAINEN, A.; HEIKKILÄ, J.; ANTTILA, A. (Eds.). **Constraint Grammar**: a Language-Independent System for Parsing Unrestricted Text. Berlim: Mouton de Gruyter, Berlin, 1995. p.103-131.

HUANG, F.; YATES, A. Exploring representation-learning approaches to domain adaptation. 2010. Disponível em: <http://www.cis.temple.edu/~yates/papers/2010-danlp-lvlms-for-domain-adaptation.pdf>. Acesso em: 02 fev. 2011.

ITAI, A.; SEGAL, E. A Corpus Based Morphological Analyzer for Unvocalized Modern Hebrew. In: WORKSHOP MACHINE TRANSLATION FOR SEMITIC LANGUAGES: ISSUES AND APPROACHES, 9., 2003. Disponível em:<http://www.mt-archive.info/MTS-2003-Itai.pdf>. Acesso em: jun 2007.

JACKSON, P.; MOULINIER, I. **Natural language processing for online applications; text retrieval, extraction and categorization.** Amsterdam ; Philadelphia: John Benjamins Publ., 2002. 225p.

JIANG, J. **Domain Adaptation in natural language processing**. 2008. Disponível em: <http://hdl.handle.net/2142/10870>. Acesso em: jan 2011.

JURAFSKY, D.; MARTIN, J. **Speech and Language Processing:** an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Upper Saddle. River, NJ: Prentice-Hall, 2000. 934 p.

KARLSSON, F.; VOUTILAINEN, A.; HEIKKILÄ, J.; ANTTILA, A. (Eds.). **Constraint Grammar**: a Language-Independent System for Parsing Unrestricted Text. Berlim: Mouton de Gruyter, 1995.

KEMPE, A. **A Probabilistic Tagger and an Analysis of Tagging Errors**. Research Report. Stuttgart: Universität Stuttgart; Institut fur Maschinelle Sprachverarbeitung, 1993.

KINOSHITA, J.; SALVADOR, L. N.; MENEZES, C. E. D. CoGrOO: a Brazilian-portuguese grammar Checker based on the CETENFOLHA Corpus. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION. 15., 2006. Genoa, Italy: LREC, 2006.

KINOSHITA, J.; SALVADOR, L. N.; MENEZES, C. E. D.; SILVA, W. D. C. M. CoGrOO - An OpenOffice grammar checker. In: INTERNATIONAL CONFERENCE ON

INTELLIGENT SYSTEMS DESIGN AND APPLICATIONs, 17., 2007. Washington, DC: IEEE Computer Society, 2007. p.525-530. DOI: 10.1109/ISDA.2007.145

KLEIN, S.; SIMMONS, R. F. A computational approach to grammatical coding of English words. **Journal of the ACM**, v. 10, n.3, p. 334-347, 1963.

KEPLER, F. N. **Modelagem de Contextos para Aprendizado Automático Aplicado à Análise Morfossintática.** São Paulo, 2010. 107f. Tese (Doutorado em Ciências) – Instituto de Matemática e Estatística, Universidade de São Paulo.

KEPLER, F. N; FINGER, M. Comparing two Markov methods for part-of-speech tagging of Portuguese. In: SICHMAN, J.S.; COELHO, H.; REZENDE, S. O. (Eds.). Lecture Notes in Computer Science: v. 4140/2006. Advances in Artificial Intelligence – IBERAMIA-SBIA 2006. Heidelberg: Springer Berlin, 2006, p. 482-491. DOI: 10.1007/11874850_52

LÁCIO-WEB. **Lácio-Web manuals**: compilação de corpus do português do Brasil e implementação de ferramentas para análises linguísticas. 2007. Disponível em: <http://www.nilc.icmc.usp.br/lacioweb/manuais.htm>. Acesso em: 20 jul 2007.

LAFFERTY, J.; MCCALLUM, A.; PERREIRA, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: 18TH INTERNATIONAL CONF. ON MACHINE LEARNING, 2001. **Proceedings…** [S.l.:S.n.].

LAGER, T. The µ-TBL system: logic programming tools for transformation-based learning. In: INTERNATIONAL WORKSHOP ON COMPUTATIONAL NATURAL LANGUAGE LEARNING, 3., 1999. **Proceedings…** [S.l.:S.n.], p.33-42. Disponível em: <http://www.ling.gu.se/~lager/Mutbl/bibliography.html>. Acesso em: jun 2007.

LAGER, T. **The µ-TBL Homepage**: tools for Transformation-Based Learning, 2007. Disponível em: <http://www.ling.gu.se/~lager/mutbl.html>. Acesso em: jun 2007.

LINGUATECA. CETENFolha. In: **LINGUATECA.** 2007. Disponível em: <http://www.linguateca.pt/CETENFolha/index_info.html>. Acesso em: 12 fev. 2007.

LINGUATECA. Material que compõe a Floresta Sintá(c)tica. In: **LINGUATECA.** Disponível em:<http://www.linguateca.pt/Floresta/material.html>. Acesso em: 12 fev. 2009.

LUGER, G. F. **Inteligência Artificial**: estruturas e estratégias para a solução de problemas complexos. Porto Alegre: Bookmann, 2004..

MANNING, C. D.; SCHUTZE, H. **Foundations of Statistical Natural Language Processing.** Cambridge, Massachusetts: The MIT Press, 1999.

MASON, O. **QTAG**. Disponível em: <http://www.softpedia.com/get/System/File-Management/OM-QTag.shtml>. Acesso em: 12 out. 2011.

MASON, O.; TUFIS, D. Probabilistic Tagging in a Multi-lingual Environment: Making an English Tagger Understand Romanian, In: THIRD TELRI EUROPEAN CONFERENC, 3., 1997. **Proceedings…** Montecatini, Italy: [S.n.], 1997.

NUGUES, P. **An Introduction to Language Processing with Perl and Prolog**: an Outline of Theories, Implementation, and Application with Special Consideration of English, French, and German. Berlim, Heidelberg: Springer, 2006.

NUNES, M.G.V. O Processamento de Línguas Naturais: para quê e para quem? **1ª. Escola Brasileira de Linguística Computacional**. São Paulo: FFCLCH-USP, 2007.

OTHERO, G.; MENUZZI, S. **Linguística computacional**: teoria e prática. São Paulo: Parábola Editorial, 2005.

POPOVIĆ, M.; NEY, H. Towards Automatic Error Analysis of Machine Translation Output. **Computational Linguistics**, 2011. (no prelo).

RATNAPARKHI, A. A maximum entropy model for part-of-speech tagging. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING EMNLP-96. **Proceedings**… Philadelphia; [S.n.], 1996.

RICH, E.; KNIGHT, K. **Inteligência Artificial**. São Paulo: Makron Books, 1993, 722p.

SAMUELSSON, C.; VOUTILAINEN, A. Comparing a Linguistic and a Stochastic Tagger. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS AND 8TH CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 35., 1997. **Proceedings...** Madrid: ACL, 1997, p. 246-253.

SANTOS, C. N. dos; MILIDIÚ, R. L.; RENTERÍA, R. Portuguese part-of-speech tagging using entropy guided transformation learning. In: TEIXEIRA, A.; LIMA, V.L.S.; OLIVEIRA, L.C.; QUARESMA, P. (Eds.). **Lecture Notes in Computer Science: v. 5190. Computational Processing of the Portuguese Language**. Heidelberg: Springer-Verlag, 2008. p. 143-152. DOI: 10.1007/978-3-540-85980-2_15

SANTOS, D. Corporizando algumas questões. In: TAGNIN, S. E. O.; VALE, O. A. (Eds.). **Avanços da Linguística de Corpus no Brasil**. São Paulo, SP: Humanitas, 2008, p. 41-66.

SCHMID, H. Probabilistic part-of-speech tagging using decision trees, In: CONFERENCE ON NEW METHODS IN LANGUAGE PROCESSING. **Proceedings…** Manchester, UK: [S.n.], 1994a. p. 44-49.

SCHMID, H. Part-of-speech tagging with neural networks. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS. **Proceedings…** Kyoto, Japan: [S.n.], 1994b. p. 172-176.

SCHMID, H.. Improvements in part-of-speech tagging with an application to German, In: ACL SIGDAT-WORKSHOP. **Proceedings**… [S.l. : S. n.], March, 1995.

SILVA, J; BRANCO, A; CASTRO, S; REIS, R. Out-of-the-box robust parsing of Portuguese. In: PARDO, T. A. S.; BRANCO, A.; KLAUTAU, A.; VIEIRA, R.; LIMA, V. L. S. (Eds.). **Lecture Notes in Computer Science**: v. 6001/2010. Computational Processing of the Portuguese Language. Heidelberg: Springer-Verlag, 2010. p.75-85. DOI: 10.1007/978-3-642-12320-7_10

SILVA, W. C.; MARTINS, L. E. G. PARADIGMA: Uma Ferramenta de Apoio à Elicitação e Modelagem de Requisitos Baseada em Processamento de Linguagem Natural. In: 11TH. WORKSHOP ON REQUIREMENTS ENGINEERING. **Proceedings…**, [S.l. : S. n.], 2008. p. 140-151. Disponível em: <http://wer.inf.puc-rio.br/WERpapers/artigos/artigos_WER08/silva.pdf>. Acesso em: fev 2011.

SUTTON, C.; ROHANIMANESH, L.; MCCALLUM, A. Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data. In: 21ST INTERNATIONAL CONFERENCE ON MACHINE LEARNING. **Proceedings…**, Banff, Canada, 2004.

TOUTANOVA, K.; KLEIN, D.; MANNING, C. D.; SINGER, Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON HUMAN LANGUAGE TECHNOLOGY, 1., **Proceedings…**, [S.l. : S. n.], 2003.p.173-180. DOI: 10.3115/1073445.1073478

TSURUOKA, Y.; TSUJII, J. Bidirectional inference with the easiest-first strategy for tagging sequence data. In: CONFERENCE ON HUMAN LANGUAGE TECHNOLOGY AND EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING. **Proceedings…** [S.l. : S. n.], 2005, p. 467-474. DOI: 10.3115/1220575.1220634

TUFIS, D. Using a large set of EAGLES-compliant morpho-syntactic descriptors as a tagset for probabilistic tagging. In: INTERNATIONAL CONFERENCE ON LANGUAGES RESOURCES AND EVALUATION. **Proceedings…** Athens: LREC, 2000. p. 1105-1112.

TUFIS, D.; MASON, O. Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger, In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES & EVALUATION, 1., 1998. **Proceedings…** Granada, Spain: LREC, 1998. p. 589-596.

UMANSKY-PESIN, S.; REICHART, R.; RAPPOPORT, A. A multi-domain web-based algorithm for POS tagging of unknown words. In: 23rd INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS. **Proceedings…**, [S.l : S. n.], 2010, p. 1274-1282. Disponível em: <http://www.aclweb.org/anthology/C10-2146>. Acesso em: dez 2010.

VAN DEN BOSCH, A.; MARSI, E.; SOUDI, A. **Memory-based morphological analysis and part-of-speech tagging of Arabic**. 2007. Disponível em: <http://ilk.uvt.nl/downloads/pub/papers/P3-C11-Jan07.pdf>. Acesso em: 12 ago 2007.

VITERBI, A. J. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. **IEEE Transactions on Information Theory**. [S.l], v. 13, n. 2, p. 260-269, 1967.

VISL. Tree structure. In: **Portuguese VISL.** 2011. Disponível em: <http://beta.visl.sdu.dk/visl/pt/parsing/automatic/trees.php>. Acesso em: 03 out 2011.

VOUTILAINEN, A. Morphological disambiguation, In: KARLSSON, F.; VOUTILAINEN, A.; HEIKKILÄ, J.; ANTTILA, A. (Eds.). **Constraint Grammar**: a Language-Independent System for Parsing Unrestricted Text. Berlin: Mouton de Gruyter, 1995a. p. 165-284.

VOUTILAINEN, A. A syntax-based part of speech analyser. In: **EACL 7**, 1995b. p. 157-164.

WILSON, G.; HEYWOOD, M. Use of a Genetic Algorithm in Brill's Transformation-Based Part-of-Speech Tagger. In: CONFERENCE ON GENETIC AND EVOLUTIONARY COMPUTATION GECCO. New York: ACM, 2005. p. 2067-2073

WITTEN, I. H.; FRANK, E. **Data Mining***:* practical machine learning tools and techniques. 2nd Edition. San Francisco: Morgan Kaufmann, 2005.

Apêndices

Apêndice A – Treinamento e Etiquetagem com o QTag

TREINAMENTO: CONSTRUÇÃO DO ARQUIVO DE RECURSOS LÉXICOS

O arquivo de recursos léxicos é gerado pelo seguinte comando em Java executado a partir do *prompt* de comando ou por um arquivo de lotes:

```
java QTag.ResourceCreator corpus_de_treinamento arquivo_de_recursos_léxico
```

Por exemplo, se o *corpus* se chama "treinamento.txt" e o *corpus* é em português, o comando a ser processado é exemplificado abaixo:

```
java QTag.ResourceCreator treinamento.txt QTag-português
```

Com esse comando, dois arquivos serão criados: um léxico ("QTag-portugues.lex") e uma matriz de transição ("QTag-português"), ambos requeridos para o processo de etiquetagem.

ETIQUETAGEM DE SENTENÇAS

As sentenças em um dado "arquivo_teste.txt" são etiquetadas, por exemplo, com os recursos léxicos de "QTag-português" com o comando:

```
java -jar QTag.jar QTag-português arquivo_teste.txt
```

Apêndice B – Treinamento e Etiquetagem com o TreeTagger

TREINAMENTO: CONSTRUÇÃO DO ARQUIVO DE RECURSOS LÉXICOS

O Treetagger é processado no ambiente Windows a partir de um arquivo executável. O arquivo de recursos léxicos ou arquivo de parâmetros é gerado a partir do seguinte comando no *prompt* de comando ou em um arquivo de lotes executado no diretório onde o *software* se encontra instalado:

```
train-tree-tagger <arquivo léxico> <arquivo de classes abertas> <arquivo de treinamento> <arquivo de parâmetros> -st . -st ! -st ?
```

Por exemplo:

```
train-tree-tagger léxico.txt classes_abertas.txt treinamento.txt português.par -st . -st ! -st ?
```

O parâmetro *–st* indica as etiquetas usadas em final de sentença. Essas etiquetas podem ser substituídas pela etiqueta default "*SENT*" representando todos os sinais de pontuação de final de sentença. No comando acima, os finais de sentença foram definidos com os sinais: ponto, exclamação e interrogação. Na criação do arquivo de recursos léxicos, alguns outros parâmetros podem ser informados ao sistema como, por exemplo, o número de palavras que precedem o contexto de etiquetagem (*–cl*). Neste estudo, foi utilizado o valor *default* 2, correspondente ao *trigrama*.

ETIQUETAGEM DE SENTENÇAS

As sentenças de "arquivo_teste" são etiquetadas, por exemplo, com os recursos léxicos de "português.par" e apresentadas no arquivo "resultado.txt" com o comando:

```
tree-tagger português.par arquivo_teste.txt resultado.txt –token
```

Apêndice C – Informações Estatísticas sobre o Treinamento e Teste com o μ-TBL

```
Learning ...
Loading data: data/treino1... done! Size is 76098.
Loading algorithm: algorithms/brill ... done!
Loading templates: templates/brill_templates ... done!
98    1.00   tag:'v-fin'>prop <- wd:'São_Paulo'@[0]
56    0.92   tag:prp>adv <- wd:','@[1]
32    1.00   tag:'pron-det'>'pron-indp' <- wd:o_que@[0]
…
3     1.00   tag:n>prop <- wd:'Copa_do_Mundo'@[0]
3     1.00   tag:n>prop <- wd:'Igreja_Católica'@[0]
3     1.00   tag:n>prop <- wd:'Polícia_Civil'@[0]
3     1.00   tag:'v-inf'>'conj-s' <- wd:do_que@[0]
2721 rule(s) for feature(s) [tag]
Saving the result ...

Evaluation teste1...
Loading templates: templates/brill_templates ... done!
Loading data: data/teste1 ... done! Size is 3980.

DATA STATISTICS:
            Corpus Size: 3980
         Number of Tags: 3980
 Number of Correct Tags: 3788
       Number of Errors: 192
                 Recall: 95.2%
              Precision: 95.2%
                F-Score: 95.2%
Number of Tags per Word: 1.000
```

Apêndice D – Conjunto de Etiquetas Modificadas para o Bosque CF 7.4

adj
adjFP
adjFS
adjM/FS
adjMP
adjMS
adv
artFP
artFS
artMP
artMS
artNS
conj-c
conj-s
ec
in
nFP
nFS
nM/FP
nM/FS
nMP
nMR
nMS
nNM/FS
nSS
num
ppFS
pron-detFP
pron-detFS
pron-detM/FS
pron-detMP
pron-detMS
pron-indp-FF
pron-indpFP
pron-indpFS
pron-indpM/FS
pron-indpM/FS/P
pron-indpMP
pron-indpMS
pron-indpMS/P
pron-pers1SNOM
pron-pers1PNOM
pron-pers1PNOM/PIV
pron-pers2PNOM
pron-pers3SNOM
pron-pers3SNOM/PIV
pron-pers3PNOM
pron-persF1SACC
pron-persF1SDAT
pron-persF1SPIV
pron-persF3PACC
pron-persF3PNOM/PIV
pron-persF3S/PACC
pron-persF3SACC
pron-persF3SDAT
pron-persM/F1PACC
pron-persM/F1PDAT
pron-persM/F1SACC
pron-persM/F1SDAT
pron-persM/F1SPIV
pron-persM/F3PACC
pron-persM/F3S/PACC
pron-persM/F3S/PDAT
pron-persM/F3SACC
pron-persM/F3SDAT
pron-persM1PACC
pron-persM1PDAT
pron-persM1PNOM/PIV
pron-persM1SACC
pron-persM1SDAT
pron-persM1SPIV
pron-persM3PACC
pron-persM3PDAT
pron-persM3PNOM
pron-persM3PNOM/PIV
pron-persM3S/PACC
pron-persM3SACC
pron-persM3SDAT
pron-persM3SNOM/PIV
pron-persM3SPIV
propFP
propFS
propM/FS
propMP
propMS
prp
pt
v-finCOND1/3S
v-finCOND1S
v-finCOND3P
v-finCOND3S
v-finFUT1/3SSUBJ
v-finFUT1PIND
v-finFUT1PSUBJ
v-finFUT1SIND
v-finFUT3PIND
v-finFUT3PSUBJ
v-finFUT3SIND
v-finFUT3SSUBJ
v-finIMP2S
v-finIMPF1/3SIND
v-finIMPF1/3SSUBJ
v-finIMPF1PIND
v-finIMPF1PSUBJ
v-finIMPF1SIND
v-finIMPF1SSUBJ
v-finIMPF3PIND
v-finIMPF3PSUBJ
v-finIMPF3SIND
v-finIMPF3SSUBJ
v-finMQP1/3SIND
v-finMQP1SIND
v-finMQP3PIND
v-finMQP3SIND
v-finPR1/3SSUBJ
v-fin-PR1/3SSUBJ
v-finPR1PIND
v-finPR1PSUBJ
v-finPR1SIND
v-finPR1SSUBJ
v-finPR2PIND
v-finPR2SSUBJ
v-finPR3PIND
v-finPR3PSUBJ
v-finPR3S
v-finPR3SIND
v-finPR3SINDVFIN
v-finPR3SSUBJ
v-finPS/MQP3PIND
v-fin-PS/MQP3PIND
v-finPS1/3SIND
v-finPS1PIND
v-finPS1SIND
v-finPS2SIND
v-finPS3PIND
v-finPS3SIND
v-finPS3SSUBJ
v-ger
v-inf
v-inf/1/3S
v-inf1P
v-inf1S
v-inf3P
v-inf3S
v-inf-MS
v-pcp
v-pcpFP
v-pcpFS
v-pcpMP
v-pcpMS
v-pcpVPCPFS
vpVPCPFS

Apêndice E – Arquivos de Classes Abertas para o TreeTagger

ETIQUETAS SIMPLES DO BOSQUE CF 7.4, PRESENTES NO ARQUIVO DE CLASSES ABERTAS PARA O TREINAMENTO DO TREETAGGER

adv	v-fin
adj	v-inf
n	v-ger
prop	v-pcp

ETIQUETAS MODIFICADAS DO BOSQUE CF 7.4, PRESENTES NO ARQUIVO DE CLASSES ABERTAS PARA O TREINAMENTO DO TREETAGGER

Adv	v-finPR3PSUBJ	v-finIMPF1SIND
Adj	v-finPR3S	v-finIMP2S
adjFS	v-finPS3SIND	v-finMQP1SIND
adjFP	v-finPS3PIND	v-finMQP3SIND
adjMS	v-finPS3SSUBJ	v-finMQP3PIND
adjMP	v-finPS1SIND	v-finPS/MQP3PIND
adjM/FS	v-finPS1PIND	v-finPR1/3SSUBJ
nMP	v-finPS2SIND	v-finCOND1/3S
nFP	v-finPR3SSUBJ	v-finPS1/3SIND
nMS	v-finFUT1PIND	v-finMQP1/3SIND
nFS	v-finFUT1SIND	v-finIMPF1/3SSUBJ
nSS	v-finFUT1PSUBJ	v-finIMPF1/3SIND
nM/FS	v-finFUT3SIND	v-finFUT1/3SSUBJ
propFS	v-finFUT3PIND	v-fin-PR1/3SSUBJ
propFP	v-finFUT3SSUBJ	v-inf
propMS	v-finFUT3PSUBJ	v-inf3S
propMP	v-finIMPF1PIND	v-inf3P
propM/FS	v-finIMPF3SIND	v-inf/1/3S
v-finPR1SIND	v-finCOND1S	v-inf1P
v-finPR1PIND	v-finCOND3P	v-ger
v-finPR1SSUBJ	v-finCOND3S	v-pcp
v-finPR1PSUBJ	v-finIMPF3PIND	v-pcpFS
v-finPR2PIND	v-finIMPF1SSUBJ	v-pcpMP
v-finPR2SSUBJ	v-finIMPF1PSUBJ	v-pcpMS
v-finPR3SIND	v-finIMPF3PSUBJ	v-pcpFP
v-finPR3PIND	v-finIMPF3SSUBJ	v-pcpVPCPFS

Apêndice F – Etiquetagem Sintática no Bosque CF 7.4 e no Ambiente VISL

BOSQUE CF 7.4 – Exemplo da etiquetagem sintática no Bosque Cf 7.4 para a sentença “Schwarzenegger está no HBO”

```
CETENFolha n=13 cad="TV Folha" sec="clt-soc" sem="94a"
CF13-2 Schwarzenegger está no HBO
A1
STA:fcl
=SUBJ:prop('Schwarzenegger' M S)        Schwarzenegger
=P:v-fin('estar' PR 3S IND)       está
=ADVS:pp
==H:prp('em' <sam->)     em
==P<:np
===>N:art('o' <-sam> <artd> M S) o
===H:prop('HBO' M S)    HBO
```

VISL (TREE) – Exemplo da etiquetagem sintática pelo PALAVRAS com a opção “Tree structure” para a sentença “Schwarzenegger está no HBO”

```
54. Schwarzenegger está no HBO
STA:fcl
=SUBJ:prop("Schwarzenegger" M/F S)      Schwarzenegger
=P:v-fin("estar" &lt;fmc&gt; PR 3S IND VFIN)      está
&lt;SA:pp
=H:prp("em" &lt;sam-&gt;)        em
=P&lt;:np
==&gt;N:pron-det("o" &lt;artd&gt; &lt;-sam&gt; DET M S)  o
==H:prop("HBO" M S)    HBO
```

VISL (MOPHOLOGICAL TAGGING) – Exemplo da etiquetagem sintática pelo PALAVRAS com a opção *morphological tagging* para a sentença “Schwarzenegger está no HBO”

```
Schwarzenegger  [Schwarzenegger] PROP M/F S/P
está        [estar] V PR 3S IND VFIN
em          [em] <sam-> PRP
o           [o] <-sam> <artd> DET M S
HBO         [HBO] PROP M/F S/P
```

Apêndice G – Comandos para Treinamento e Teste do TreeTagger

Comandos em um arquivo de lotes para treinamento do TreeTagger com os 10 arquivos de treinamento (tr01_90pcentoTR.txt a tr10_90pcentoTR.txt), que resultam em 10 arquivos de parâmetros (01.par a 10.par) para a etiquetagem de seus 10 arquivos de testes disjuntos:

```
train-tree-tagger lexTR.txt clas_abertas.txt tr01_90pcentoTR.txt 01.par -st . -st ! -st ? -st ;
train-tree-tagger lexTR.txt clas_abertas.txt tr02_90pcentoTR.txt 02.par -st . -st ! -st ? -st ;
train-tree-tagger lexTR.txt clas_abertas.txt tr03_90pcentoTR.txt 03.par -st . -st ! -st ? -st ;
train-tree-tagger lexTR.txt clas_abertas.txt tr04_90pcentoTR.txt 04.par -st . -st ! -st ? -st ;
train-tree-tagger lexTR.txt clas_abertas.txt tr05_90pcentoTR.txt 05.par -st . -st ! -st ? -st ;
train-tree-tagger lexTR.txt clas_abertas.txt tr06_90pcentoTR.txt 06.par -st . -st ! -st ? -st ;
train-tree-tagger lexTR.txt clas_abertas.txt tr07_90pcentoTR.txt 07.par -st . -st ! -st ? -st ;
train-tree-tagger lexTR.txt clas_abertas.txt tr08_90pcentoTR.txt 08.par -st . -st ! -st ? -st ;
train-tree-tagger lexTR.txt clas_abertas.txt tr09_90pcentoTR.txt 09.par -st . -st ! -st ? -st ;
train-tree-tagger lexTR.txt clas_abertas.txt tr10_90pcentoTR.txt 10.par -st . -st ! -st ? -st ;
```

Comandos em um arquivo de lotes para etiquetagem com o TreeTagger dos 10 arquivos de testes (teste01_10pcentoTR.txt a teste10_10pcentoTR.txt):

```
tree-tagger 01.par teste01_10pcentoTR.txt res01.txt -token
tree-tagger 02.par teste02_10pcentoTR.txt res02.txt -token
tree-tagger 03.par teste03_10pcentoTR.txt res03.txt -token
tree-tagger 04.par teste04_10pcentoTR.txt res04.txt -token
tree-tagger 05.par teste05_10pcentoTR.txt res05.txt -token
tree-tagger 06.par teste06_10pcentoTR.txt res06.txt -token
tree-tagger 07.par teste07_10pcentoTR.txt res07.txt -token
tree-tagger 08.par teste08_10pcentoTR.txt res08.txt -token
tree-tagger 09.par teste09_10pcentoTR.txt res09.txt -token
tree-tagger 10.par teste10_10pcentoTR.txt res10.txt -token
```

Comando para treinamento do TreeTagger com o conjunto de treinamento TR.txt, que resulta em um arquivo de parâmetros (11.par) para ser usado na etiquetagem do conjunto de teste TE (não visto no treinamento do processo de validação cruzada):

```
train-tree-tagger lexTR.txt clas_abertas.txt TR.txt 11.par -st . -st ! -st ? -st ;
```

Comando para etiquetagem com o TreeTagger do arquivo de teste TE:

```
tree-tagger 11.par input_testeTE.txt res11.txt -token
```

Apêndice H – Comandos para Extrair Regras com o μ-TBL

Conteúdo do arquivo *regrasmutbl.script* com os comandos para extrair regras com o μ-TBL.

```
%mutbl.exe -f regrasmutbl.script > regrasmutbl.txt
%inicio: 15:48h de 16/03/2009
%fim: 01:10h de 17/02/2009.
%duracao: 09:58h

set training_data='data/CorpusMMTT/treinamento'.
set algorithm='algorithms/brill'.
set templates='templates/brill_templates'.
set score_threshold=3.
set accuracy_threshold=0.6.
set verbosity=2.

echo("Learning ...").
learn_rule_seq.
echo("Saving the result ...").
save_stack(regrasaprendidas).
echo(' ').
echo("Evaluation teste_01...").
echo(' ').
set test_data='data/CorpusMMTT/teste01'.
test_rule_seq.

...
echo(' ').
echo("Evaluation teste_10...").
echo(' ').
set test_data='data/CorpusMMTT/teste10'.
test_rule_seq.
echo(' ').

echo(' ').
echo('TESTES COM O CONJUNTO NÃO VISTO: TE').
echo(' ').

echo(' ').
echo("Evaluation TE...").
echo(' ').
set test_data='data/CorpusMMTT/te'.
test_rule_seq.
echo(' ').

write_html_error_data.
echo('templates usadas...').
print_used_templates.
echo(' ').
echo('templates não usadas...').
print_unused_templates.

echo(' ').
echo('estatísticas das templates').
template_statistics.
echo(' ').
echo('estatísticas do treinamento').
learning_statistics.
```