

177540-4

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

**AVALIANDO UM ROTULADOR
ESTATÍSTICO DE CATEGORIAS
MORFO-SINTÁTICAS PARA A
LÍNGUA PORTUGUESA**

Por

Aline Villavicencio

Dissertação submetida à avaliação, como requisito parcial
para obtenção do grau de
Mestre em Ciência da Computação

Prof. Dra. Rosa Maria Viccari

Orientadora



UFRGS

SABi



05226573

Porto Alegre, outubro de 1995.

UFRGS
INSTITUTO DE INFORMÁTICA
BIBLIOTECA

CIP - CATALOGAÇÃO NA PUBLICAÇÃO

Villavicencio, Aline

Avaliando um Rotulador de Categorias Morfo-Sintáticas para a Língua Portuguesa / Aline Villavicencio. - Porto Alegre: CPGCC da UFRGS, 1995.

136.: il.

Dissertação (mestrado) - Universidade Federal do Rio Grande do Sul, Curso de Pós-Graduação em Ciência da Computação, Porto Alegre, BR - RS, 1995. Orientador: Viccari, Rosa Maria.

1:Rotuladores de Categorias Morfo-Sintáticas. 2:Corpus.
3:Hidden Markov Models. 4:Métodos Estatísticos. I. Viccari, Rosa Maria. II.Título

| UFRGS INSTITUTO DE INFORMÁTICA BIBLIOTECA | | |
|---|-------------------|---------------------|
| N.º CHAMADA 681.3.013(043) V727A | N.º REG.: 4004 | |
| | DATA: 06.02.95 | |
| ORIGEM: D | LATA: 13/12/96 | PREÇO: R\$ 30,00 |
| FUNDO: II | FORN.: II | |

*Inteligência artificial. SBU
Linguística computacional
Rotuladores: categorias morfo-sintáticas
Processamento: língua natural*

ENPq 1.03.01.00-3

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Reitor: Prof. Hélgio Trindade

Pró-Reitor de Pesquisa e Pós-Graduação: Prof. Cláudio Scherer

Diretor do Instituto de Informática: Prof. Roberto Tom Price

Coordenador do CPGCC: Prof. José Palazzo Moreira de Oliveira

Bibliotecária-Chefe do Instituto de Informática: Zita Prates

*"O, wonder!
How many goodly creatures are there here!
How beauteous mankind is!
O BRAVE NEW WORLD
That has such people in't"*

William Shakespeare, "A Tempestade", Ato V

*"There are more things in heaven and earth,
Horatio,
Than are dreant of in your philosophy."*

William Shakespeare, "Hamlet", Ato I

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Sistema de Biblioteca da UFRGS

681.3.013(043)
V727A

4004
INF
1997/177540-7
1997/02/06

AGRADECIMENTOS

Ao meu marido, Fábio, por existir. Passamos juntos por muita coisa, inclusive as nossas dissertações e posso dizer que tudo fica mais agradável quando tu estás junto. Obrigada por toda a ajuda e por todo o amor que me deste.

Ao CNPq pelo auxílio financeiro recebido em forma de bolsa de mestrado.

Aos professores do CPGCC, em especial a minha orientadora, Profa. Rosa, pela amizade, pelo apoio e pelas valiosas sugestões que muito me auxiliaram na elaboração deste trabalho. A tua orientação foi além do âmbito acadêmico, guiando meus passos também no campo da vida.

Ao Prof. Gabriel Lopes por emprestar seus tão vastos conhecimentos de processamento de linguagem natural para que eu pudesse tatear por esta área. Um agradecimento especial ao Nuno, pela amizade, pelo auxílio nas horas de desespero e pelas grandes discussões filosóficas ao longo deste trabalho. E aos grandes amigos que encontrei em Portugal (Sérgio & Tê, Michael & Iara, Victor, Antonio, Kadu, Celson, Paulo & Fernanda). Obrigada a vocês pela confiança, pela amizade e por acreditarem no meu trabalho.

Ao Mano (Fábio Villavicencio) pela ajuda, amizade e belo trabalho apresentados enquanto estivemos pesquisando juntos.

Às integrantes da “tripla dinâmica” Neila e Milene, pelas horas agradáveis que passamos juntas nos dedicando ao estudo do divertimento.

A TODOS os diversos amigos que fiz no Instituto de Informática. Gente, muito obrigada pela amizade e apoio. Vocês ajudaram a criar estas páginas.

Representando minha família, que tanto gosto, envio um agradecimento especial aos meus irmãos caçulinhas: Ricardo e Bianca.

Por fim, dedico este trabalho aos meus pais, Ricardo e Ingrid, e à minha avó, Hilda, que tanto amo. Obrigada por tudo. Vocês me proporcionaram tudo o que consegui até hoje, mas a lição mais importante que aprendi com vocês foi a do AMOR. Vocês são únicos !

SUMÁRIO

| | |
|---|-----------|
| LISTA DE ABREVIATURAS..... | 9 |
| LISTA DE TRADUÇÕES..... | 10 |
| LISTA DE FIGURAS..... | 11 |
| LISTA DE TABELAS..... | 11 |
| LISTA DE FÓRMULAS..... | 13 |
| RESUMO..... | 16 |
| ABSTRACT..... | 18 |
| 1 INTRODUÇÃO..... | 19 |
| 2 CONCEITOS BÁSICOS..... | 23 |
| 2.1 Corpora Disponíveis..... | 23 |
| 2.2 Conjuntos de Rótulos..... | 24 |
| 2.2.1 Definindo o Conjunto de Rótulos..... | 25 |
| 2.2.1.1 Tamanho do Conjunto de Rótulos..... | 26 |
| 2.2.1.2 Especificidade do Conjunto de Rótulos..... | 26 |
| 2.2.1.3 Generalidade do Conjunto de Rótulos..... | 27 |
| 2.3 Ambigüidade Lexical..... | 29 |
| 2.3.1 Classes de Ambigüidade..... | 30 |
| 2.4 Teoria da Informação de Shannon..... | 31 |
| 2.5 Modelo de N-Gramas..... | 32 |
| 2.6 Modelos de Markov..... | 33 |
| 2.6.1 Cadeia de Markov..... | 34 |
| 2.6.2 Hidden Markov Model..... | 36 |
| 2.7 Marcação de Categorias Morfo-Sintáticas..... | 38 |
| 2.7.1 O Problema da Ambigüidade Lexical..... | 39 |
| 2.8 Marcação Manual x Marcação Automática..... | 40 |
| 3. ROTULADORES DE CATEGORIAS MORFO-SINTÁTICAS..... | 41 |
| 3.1 Sistemas Rotuladores Baseados em Regras..... | 41 |
| 3.2 Sistemas Rotuladores Baseados em Métodos Estatísticos..... | 42 |
| 3.3 Sistemas Rotuladores Baseados em Regras X Sistemas Rotuladores Estatísticos..... | 43 |
| 3.3.1 Rotulador de Categorias Morfo-Sintáticas Misto..... | 45 |
| 3.4 Marcação Automática de Categorias Morfo-Sintáticas Usando a Teoria da Informação de Shannon..... | 49 |

| | |
|---|------------|
| 3.5 HMM para Marcação Automática de Categorias Morfo-Sintáticas..... | 52 |
| 3.6 Algoritmos de Treinamento e de Marcação..... | 57 |
| 3.6.1 Frequência Relativa(FR)..... | 58 |
| 3.6.2 Algoritmo de Forward-Backward..... | 60 |
| 3.6.2.1 Forward e Backward Probabilities..... | 60 |
| 3.6.2.2 Funcionamento do Algoritmo..... | 64 |
| 3.6.2.3 Problemas..... | 69 |
| 3.6.3 Algoritmo de Viterbi..... | 74 |
| 3.6.4 Esparsidade dos Dados..... | 77 |
| 3.6.4.1 Refinamento..... | 78 |
| 3.7 Treinamento com Corpus Marcado X Treinamento com Corpus não Marcado..... | 78 |
| 4. IMPLEMENTAÇÃO DO SISTEMA ROTULADOR..... | 82 |
| 4.1 Conjunto de Rótulos..... | 83 |
| 4.2 Classes de Ambigüidade..... | 90 |
| 4.3 O Radiobras Corpus..... | 90 |
| 4.3.1 Preparação do corpus..... | 93 |
| 4.3.2 Marcando o Radiobras Corpus..... | 96 |
| 4.3.3 Corpus de Treinamento e Corpus de Teste..... | 97 |
| 4.4 O Dicionário..... | 97 |
| 4.5 Arquitetura do Sistema Rotulador..... | 98 |
| 4.5.1 Módulo Classificador..... | 99 |
| 4.5.2 Módulo Construtor de HMMs..... | 101 |
| 4.5.3 Módulo de Viterbi..... | 103 |
| 4.5.4 Testes de Funcionamento..... | 104 |
| 4.6 Marcação..... | 105 |
| 4.6.1 Precisão..... | 106 |
| 5. AVALIAÇÃO DO SISTEMA ROTULADOR..... | 108 |
| 5.1 Treinamento com Corpus Acentuado..... | 108 |
| 5.2 Tamanho do Corpus de Treinamento..... | 111 |
| 5.3 Dicionário Fechado x Dicionário Aberto..... | 112 |
| 5.4 Resultados..... | 113 |
| 5.4.1 Corpus de Treinamento Acentuado..... | 114 |
| 5.4.1.1 Usando o Dicionário Aberto..... | 115 |
| 5.4.1.2 O Corpus Acentuado com o Dicionário Aberto..... | 116 |
| 5.4.1.3 Usando o Dicionário Fechado..... | 119 |
| 5.4.1.4 O Corpus Acentuado com o Dicionário Fechado..... | 120 |
| 5.4.1.5 Dicionário Aberto x Dicionário Fechado..... | 123 |

| | |
|---|------------|
| 5.4.2 Corpus de Treinamento Desacentuado..... | 125 |
| 5.4.2.1 Usando o Dicionário Aberto..... | 125 |
| 5.4.2.2 O Corpus Desacentuado com o Dicionário Aberto..... | 126 |
| 5.4.2.3 Usando o Dicionário Fechado..... | 130 |
| 5.4.2.4 O Corpus Desacentuado com o Dicionário Fechado..... | 131 |
| 5.4.3 Dicionário Aberto x Dicionário Fechado..... | 134 |
| 5.4.4 Corpus Acentuado x Corpus Desacentuado..... | 136 |
| 5.4.5 Conclusões a Respeito dos Resultados..... | 140 |
| 6. CONCLUSÕES E TRABALHOS FUTUROS..... | 143 |
| 6.1 Trabalhos Futuros..... | 145 |
| 7. BIBLIOGRAFIA..... | 146 |

LISTA DE ABREVIATURAS

| | |
|------------|------------------------------------|
| PLN | Processamento de Linguagem Natural |
| HMM | “Hidden Markov Model” |
| POS Tagger | “Part-of-Speech Tagger” |

LISTA DE TRADUÇÕES

| | |
|---------------------------------|--|
| Part-of-Speech Tagging | Marcação de Categorias Morfo-Sintáticas |
| Part-of-Speech Tagger | Sistema Rotulador de Categorias Morfo-Sintáticas |
| Bigrams | Bigramas |
| Trigrams | Trigramas |
| N-grams | N-gramas |
| Noisy Channel | Canal com Ruído |
| Stochastic Methods | Métodos Estocásticos |
| Relative Frequency Algorithm | Algoritmo de Frequência Relativa |
| Tag | Rótulo |
| Tag Set | Conjunto de Rótulos |

LISTA DE FIGURAS

| | |
|---|-----|
| Figura 2.1 - Canal com Ruído | 31 |
| Figura 2.2- Cadeia de Markov | 35 |
| Figura 2.3 - HMM Com Dois Estados | 37 |
| Figura 2.4 - Árvore da Seqüência "aab" | 38 |
| Figura 2.5 - Marcação da sentença 2.5 | 39 |
| Figura 2.6 - Possíveis alinhamentos para a sentença 2.5 | 39 |
| Figura 2.7 - Alinhamento correto da sentença 2.5 | 39 |
| Figura 3.1 - Sistema Rotulador | 41 |
| Figura 3.2 - Treinamento do Rotulador Inicial | 45 |
| Figura 3.3 - Marcando com o Rotulador Inicial | 46 |
| Figura 3.4 - Marcando com o Rotulador Inicial | 47 |
| Figura 3.5 - Segunda Fase | 48 |
| Figura 3.6 - Marcação através do modelo de Shannon | 49 |
| Figura 3.7 - HMM Com Três Estados | 54 |
| Figura 3.8 - HMM Com Três Estados (outra visualização) | 54 |
| Figura 3.9 - HMM de Primeira Ordem Para Marcação Morfo-Sintática | 56 |
| Figura 3.10 - HMM de Segunda Ordem para Marcação Morfo-Sintática | 56 |
| Figura 3.11 - Ciclo de Aquisição e Utilização do Conhecimento de um Rotulador | 57 |
| Figura 3.12 - HMM Com Dois Estados | 61 |
| Figura 3.13 - HMM Inicial | 66 |
| Figura 3.14 - HMM Após a Primeira Iteração | 67 |
| Figura 3.15 - HMM Após a Segunda Iteração | 68 |
| Figura 3.16 - HMM Após a Terceira Iteração | 69 |
| Figura 3.17 - Segundo HMM Inicial | 70 |
| Figura 3.18 - HMM Após a Primeira Iteração | 71 |
| Figura 3.19 - HMM Após a Segunda Iteração | 72 |
| Figura 3.20 - Gráfico do Melhor Local x Melhor Global | 73 |
| Figura 3.21 - HMM Com Dois Estados | 74 |
| Figura 3.22 - Árvore da Seqüência "aab" usando o Algoritmo de Viterbi | 75 |
| Figura 3.23 - Árvore da Seqüência "aab" | 76 |
| Figura 4.1 - Boletim da Radiobras | 91 |
| Figura 4.2 - Notícia da Radiobras | 92 |
| Figura 4.3- Final do Boletim da Radiobras | 93 |
| Figura 4.4 - Boletim Após o Processo de Retirada de Frases que não Apresentem Estrutura Sintática | 94 |
| Figura 4.5 - Boletim Após o Processo de Troca | 95 |
| Figura 4.6 - Classes de Ambigüidade das Palavras | 96 |
| Figura 4.7 - Marcação das Palavras | 96 |
| Figura 4.8 - Palavras no Corpus Marcado | 97 |
| Figura 4.9 - Corpus Ordenado | 98 |
| Figura 4.10 - Duas Entradas do Dicionário | 98 |
| Figura 4.11 - A Arquitetura do Sistema Rotulador | 99 |
| Figura 4.12 - O Módulo Classificador | 100 |
| Figura 4.13 - Palavras Analisadas pelo Classificador | 100 |
| Figura 4.14 - A Arquitetura do Construtor | 102 |

| | |
|---|-----|
| Figura 4.15 - A Arquitetura do Classificador | 104 |
| Figura 4.16 - Marcação de um Corpus | 106 |
| Figura 5.1 - Corpus Acentuado | 109 |
| Figura 5.2 - Corpus Desacentuado | 110 |
| Figura 5.3 - Dicionário Aberto: Tamanho do Corpus de Treinamento Acentuado x Precisão | 116 |
| Figura 5.4 - Dicionário Aberto: Palavras Ambíguas e Palavras Desconhecidas no Corpus Acentuado | 118 |
| Figura 5.5 - Dicionário Aberto: Rótulos por Palavra no Corpus de Teste Acentuado | 118 |
| Figura 5.6 - Dicionário Aberto: Palavras Desconhecidas no Corpus Acentuado | 119 |
| Figura 5.7 - Dicionário Fechado: Tamanho do Corpus de Treinamento Acentuado x Precisão | 120 |
| Figura 5.8 - Dicionário Fechado: Palavras Ambíguas e Palavras Desconhecidas no Corpus Acentuado | 122 |
| Figura 5.9 - Dicionário Fechado: Rótulos por Palavra no Corpus de Teste Acentuado | 122 |
| Figura 5.10 - Classes de Ambigüidade no "Corpus" Acentuado - Dicionário Aberto x Dicionário Fechado | 124 |
| Figura 5.11 - Precisão no Corpus Acentuado - Dicionário Aberto x Dicionário Fechado | 124 |
| Figura 5.12 - Dicionário Aberto: Tamanho do Corpus de Treinamento Desacentuado x Precisão | 126 |
| Figura 5.13 - Dicionário Aberto: Palavras Ambíguas e Desconhecidas no Corpus Desacentuado | 129 |
| Figura 5.14 - Dicionário Aberto: Rótulos por Palavra no Corpus de Teste Desacentuado | 129 |
| Figura 5.15 - Dicionário Aberto: Palavras Desconhecidas no Corpus Desacentuado | 129 |
| Figura 5.16 - Dicionário Fechado: Tamanho do Corpus de Treinamento x Precisão | 130 |
| Figura 5.17 - Dicionário Fechado: Palavras Ambíguas no Corpus Desacentuado | 133 |
| Figura 5.18 - Dicionário Fechado: Rótulos por Palavra no Corpus de Teste Desacentuado | 133 |
| Figura 5.19 - Classes de Ambigüidade - Dicionário Aberto x Dicionário Fechado | 135 |
| Figura 5.20 - Corpus Acentuado - Dicionário Aberto x Dicionário Fechado | 135 |
| Figura 5.21 - Precisão: Corpus Acentuado x Corpus Desacentuado | 138 |
| Figura 5.22 - Variação da Precisão | 138 |
| Figura 5.23 - Classes de Ambigüidade: Corpus Acentuado x Corpus Desacentuado | 139 |

LISTA DE TABELAS

| | |
|--|-----|
| Tabela 2.1 - Rótulos de Morfo-Sintáticos usados para a Sentença 2.1 | 25 |
| Tabela 2.2 - Informação Lexical | 29 |
| Tabela 3.1 - Padrões de Rótulos | 48 |
| Tabela 3.2 : Bigramas X Trigramas | 52 |
| Tabela 3.3 - Probabilidades Lexicais | 59 |
| Tabela 3.4 - Forward Probabilities | 62 |
| Tabela 3.5 - Backward Probabilities | 64 |
| Tabela 3.6 - Primeira Iteração | 67 |
| Tabela 3.7 - Segunda Iteração | 68 |
| Tabela 3.8 - Terceira Iteração | 69 |
| Tabela 3.9 - Primeira Iteração | 71 |
| Tabela 3.10 - Segunda Iteração | 72 |
| Tabela 4.1 - Conjunto de Rótulos Usado | 85 |
| Tabela 5.1 - Testes Efetuados | 114 |
| Tabela 5.2 - Resultados no Corpus Acentuado, Dicionário Aberto | 115 |
| Tabela 5.3 - Dicionário Aberto: Classes de Ambigüidade do Corpus Acentuado | 117 |
| Tabela 5.4 - Resultados no Corpus Acentuado, Dicionário Fechado | 120 |
| Tabela 5.5 - Dicionário Fechado: Classes de Ambigüidade do Corpus Acentuado | 121 |
| Tabela 5.6 - Corpus Acentuado: Dicionário Aberto x Dicionário Fechado | 123 |
| Tabela 5.7 - Resultados no Corpus Desacentuado, Dicionário Aberto | 125 |
| Tabela 5.8 - Dicionário Aberto: Classes de Ambigüidade do Corpus Desacentuado | 126 |
| Tabela 5.9 - Resultados no Corpus Desacentuado, Dicionário Fechado | 130 |
| Tabela 5.10 - Dicionário Fechado: Classes de Ambigüidade do Corpus Desacentuado | 132 |
| Tabela 5.11 - Corpus Desacentuado: Dicionário Aberto x Dicionário Fechado | 134 |
| Tabela 5.12 - Corpus Acentuado x Corpus Desacentuado | 137 |

LISTA DE FÓRMULAS

| | | |
|--------------|---|----|
| Fórmula 2.1 | $\hat{E} = \max_E P(E I) = \max_E P(E)P(I E)$ | 29 |
| Fórmula 3.1 | $\max P(T W) = \max_T P(T)P(W T)$ | 47 |
| Fórmula 3.2 | $P(T) = P(t_1, t_2, \dots, t_n) \approx \prod_{i=1}^n P(t_i t_{i-1})$ | 48 |
| Fórmula 3.3 | $P(T) = P(t_1, t_2, \dots, t_n) \approx \prod_{i=1}^n P(t_i t_{i-2}, t_{i-1})$ | 48 |
| Fórmula 3.4 | $P(W T) = P(w_1, w_2, \dots, w_n t_1, t_2, \dots, t_n) \approx \prod_{i=1}^n P(w_i t_i)$ | 48 |
| Fórmula 3.5 | $P(t_n w_{1,n-1}, t_{1,n-1}) = P(t_n t_{n-1})$ | 52 |
| Fórmula 3.6 | $P(w_n w_{1,n-1}, t_{1,n}) = P(w_n t_n)$ | 52 |
| Fórmula 3.7 | $P(w_{1,n}) = \sum_{t_{1,n+1}} \prod_{i=1}^n P(w_i t_i) P(t_{i+1} t_i)$ | 53 |
| Fórmula 3.8 | $P(t_n w_{1,n-1}, t_{1,n-1}) = P(t_n t_{n-2}, t_{n-1})$ | 53 |
| Fórmula 3.9 | $P(w_{1,n}) = \sum_{t_{1,n+1}} \prod_{i=1}^n P(w_i, t_i) P(t_{i+1} t_{i-1}, t_i)$ | 53 |
| Fórmula 3.10 | $P(w_i t_i) = \frac{F(w_i, t_i)}{F(t_i)}$ | 57 |
| Fórmula 3.11 | $P(t_i t_{i-1}) = \frac{F(t_{i-1}, t_i)}{F(t_i)}$ | 57 |
| Fórmula 3.12 | $\stackrel{def}{i}(t) = P(w_{1,t-1}, S_t = s^i), t > 1$ | 58 |
| Fórmula 3.13 | $\stackrel{def}{i}(1) = \begin{cases} 1,0 & \text{se } i=1 \\ \text{senão} & 0 \end{cases}$ | 58 |
| Fórmula 3.14 | $P(w_{1,n}) = \sum_{i=1}^{\sigma} P(w_{1,n}, S_{n+1} = s^i) = \sum_{i=1}^{\sigma} \alpha_i (n+1)$ | 58 |

- Fórmula 3.15
$$j(t+1) = \sum_{i=1}^{\sigma} \alpha_i(t) P(s^i \xrightarrow{w_t} s^j)$$
 58
- Fórmula 3.16
$$\beta_i(t) \stackrel{def}{=} P(w_{t,n} | S_t = s^i)$$
 60
- Fórmula 3.17
$$\beta_1(1) = P(w_{1,n} | S_1 = s^1) = P(w_{1,n})$$
 60
- Fórmula 3.18
$$\beta_i(t-1) = \sum_{j=1}^{\sigma} P(s^i \xrightarrow{w_{t-1}} s^j) \beta_j(t)$$
 60
- Fórmula 3.19
$$\beta_i(n+1) = P(\varepsilon | S_{n+1} = s^i) = 1$$
 60
- Fórmula 3.20
$$P_e(s^i \xrightarrow{w^k} s^j) = \frac{C(s^i \xrightarrow{w^k} s^j)}{\sum_{l=1, m=1}^{\sigma, \omega} C(s^i \xrightarrow{w^m} s^l)}$$
 62
- Fórmula 3.21
$$C(s^i \xrightarrow{w^k} s^j) = \sum_{s_{1,n+1}} P(s_{1,n+1} | w_{1,n}) \eta(s^i \xrightarrow{w^k} s^j, s_{1,n}, w_{1,n})$$
 62
- Fórmula 3.22
$$i(1) = s^i$$
 72
- Fórmula 3.23
$$i(t+1) = \sigma_j(t) \circ s^i, \quad j = \arg \max_{k=1}^{\sigma} P(\sigma_k(t)) P(s^k \xrightarrow{w_t} s^i)$$
 72
- Fórmula 4.1
$$p(W, T) = \Pi p(t_i w_i | t_{i-1}) = \Pi p(t_i C_i | t_{i-1})$$
 98
- Fórmula 4.2
$$p(t_i C_i | t_{i-1}) = \frac{f(t_{i-1}, t_i C_i)}{f(t_{i-1})}$$
 98
- Fórmula 4.3
$$p(t_i C_i | t_{i-1}) = \varepsilon \frac{f(t_{i-1}, t_i C_i)}{f(t_{i-1})} + (1 - \varepsilon) \frac{1}{N_T N_C}$$
 98

RESUMO

O Processamento de Linguagem Natural (PLN) é uma área da Ciência da Computação, que vem tentando, ao longo dos anos, aperfeiçoar a comunicação entre o homem e o computador. Várias técnicas têm sido utilizadas para aperfeiçoar esta comunicação, entre elas a aplicação de métodos estatísticos.

Estes métodos têm sido usados por pesquisadores de PLN, com um crescente sucesso e uma de suas maiores vantagens é a possibilidade do tratamento de textos irrestritos.

Em particular, a aplicação dos métodos estatísticos, na marcação automática de “corpus” com categorias morfo-sintáticas, tem se mostrado bastante promissora, obtendo resultados surpreendentes.

Assim sendo, este trabalho descreve o processo de marcação automática de categorias morfo-sintáticas. Inicialmente, são apresentados e comparados os principais métodos aplicados à marcação automática: os métodos baseados em regras e os métodos estatísticos. São descritos os principais formalismos e técnicas usadas para esta finalidade pelos métodos estatísticos. É introduzida a marcação automática para a Língua Portuguesa, algo até então inédito.

O objetivo deste trabalho é fazer um estudo detalhado e uma avaliação do sistema rotulador de categorias morfo-sintáticas, a fim de que se possa definir um padrão no qual o sistema apresente a mais alta precisão possível. Para efetuar esta avaliação, são especificados alguns critérios: a qualidade do “corpus” de treinamento, o seu tamanho e a influência das palavras desconhecidas.

A partir dos resultados obtidos, espera-se poder aperfeiçoar o sistema rotulador, de forma a aproveitar, da melhor maneira possível, os recursos disponíveis para a Língua Portuguesa.

PALAVRAS-CHAVE: Rotuladores de Categorias Morfo-Sintáticas, Corpus, Hidden Markov Models

TITLE: "Evaluating a Stochastic Part-of-Speech Tagger for the Portuguese Language."

ABSTRACT

Natural Language Processing (NLP) is an area of Computer Sciences, that have been trying to improve communication between human beings and computers. A number of different techniques have been used to improve this communication and among them, the use of stochastic methods.

These methods have successfully being used by NLP researchers and one of their most remarkable advantages is that they are able to deal with unrestricted texts.

Namely, the use of stochastic methods for part-of-speech tagging has achieving some extremely good results.

Thus, this work describes the process of part-of-speech tagging. At first, we present and compare the main tagging methods: the rule-based methods and the stochastic ones. We describe the main stochastic tagging formalisms and techniques for part-of-speech tagging. We also introduce part-of-speech tagging for the Portuguese Language.

The main purpose of this work is to study and evaluate a part-of-speech tagger system in order to establish a pattern in which it is possible to achieve the greatest accuracy. To perform this evaluation, several parameters were set: the corpus quality, its size and the relation between unknown words and accuracy.

The results obtained will be used to improve the tagger, in order to use better the available Portuguese Language resources.

KEYWORDS: Part-of-Speech Taggers, Corpus, Hidden Markov Models.

1. INTRODUÇÃO

A área do Processamento de Linguagem Natural, PLN, tem como objetivo o desenvolvimento de ferramentas que possibilitem uma comunicação mais natural entre o homem e o computador. E que meio mais natural do que a própria língua que o homem utiliza para se comunicar com os seus semelhantes?

Os estudos nesta área tentam possibilitar que, algum dia, o homem possa expressar as suas necessidades ao computador, da mesma forma que as expressa aos seus semelhantes. Em outras palavras, fazer com que o homem possa interagir com o computador sem necessitar de conhecimento algum que não a sua própria língua, seja na forma falada ou na escrita.

Muitos avanços foram alcançados até o momento. Contudo, devido à complexidade das línguas naturais em geral, se torna muito difícil construir uma ferramenta que consiga capturá-las integralmente. Este problema, normalmente, é resolvido com a delimitação de um subconjunto da língua e o tratamento exclusivo dele. Ou seja, a maioria das ferramentas de PLN existentes realiza uma análise bem detalhada, mas trata somente de um determinado domínio restrito da língua. Isto significa que a ferramenta somente poderá ser utilizada para textos que tenham uma determinada estrutura especificada pela ferramenta.

Entretanto, nos últimos anos, com o avanço cada vez mais rápido dos computadores e da tecnologia relacionada, pôde-se observar um aumento muito grande na quantidade de informações disponíveis. Diariamente, transitam pelo mundo, trilhões de unidades de informação.

Graças a este avanço, textos eletrônicos são facilmente obtidos. Estão disponíveis, por exemplo, textos clássicos de grandes escritores, como Shakespeare. Há também informativos eletrônicos com notícias, como o da Radiobras. Isto sem mencionar as publicações científicas, como artigos e teses disponíveis também em formato eletrônico. Outra tecnologia que está facilitando o acesso a textos é o uso dos CD-ROMs. Pode-se ter, por exemplo, textos de jornais como “A Folha de São Paulo”

ou “Zero Hora”. Isto, além de inúmeras revistas e enciclopédias como a “Neo Interativa” e a “Encarta”.

Com todos estes recursos, dispõe-se de textos com milhões, bilhões de palavras, com estruturas lingüísticas das mais variadas. Segundo Church, em [CHU93] “... *nunca houve tantos textos disponíveis como agora*”. Face a tal quantidade de informações, há a necessidade de ferramentas que tratem de textos com tal magnitude. Precisa-se de ferramentas que não apresentem restrições de domínio; ferramentas que possam tratar de textos “reais” (ou irrestritos).

Desta necessidade, renasceu o pensamento empírico que esteve em evidência nos anos 50, que sugeria a construção de ferramentas que pudessem tratar tais quantidades de textos. Este pensamento sugere ainda que, para a análise lingüística, sejam usados métodos empíricos e estatísticos. Uma frase de Firth, pesquisador de renome nos anos 50, citado em [CHU93], pode resumir claramente a filosofia defendida por esta corrente de pensadores: “*Você conhecerá uma palavra por suas companhias*”. Ou seja, uma palavra é tratada não apenas pelo que ela representa, mas também pelo que as palavras ao seu redor representam. É introduzido o conceito de contexto, que dita que o tratamento a ser dado a uma palavra depende do tratamento dado à sua vizinhança. E, seguindo o pensamento empírico, o tratamento destas palavras é feito usando métodos estatísticos.

Com o uso dos métodos estatísticos, é possível fazer a análise de textos irrestritos. Estes textos são denominados assim, pois não são apresentadas restrições, nem em relação ao seu formato, nem em relação ao seu tamanho. Deste modo, pode-se analisar textos com milhões de palavras, textos científicos, jornalísticos, livros, etc.

Outra vantagem destes métodos é o fato de não se precisar definir um domínio específico para a análise, como geralmente é feito quando se trabalha com a língua natural. Assim, textos com as mais variadas estruturas lingüísticas podem ser tratados sem restrições.

Muitas áreas estão sendo beneficiadas pelo uso de métodos estatísticos; entre elas as áreas de aquisição de conhecimento lexical, construção de gramáticas e

tradução automática. Há uma gama muito grande de aplicações a serem pesquisadas, como, por exemplo, o reconhecimento de restrições semânticas entre palavras, a identificação de classes de subcategorização ou um “parsing” parcial.

Uma aplicação que tem apresentado resultados bastante promissores é a marcação automática de categorias morfo-sintáticas (“Part-of-Speech Tagging”) de textos escritos. Estão sendo construídos sistemas rotuladores (“Part-of-Speech Taggers”), capazes de realizar esta marcação automática. Estes sistemas utilizam métodos estatísticos para poderem fazer tratamento de textos irrestritos. Recebem, como entrada, um texto e analisando-o, atribuem, para cada uma das palavras do texto, o rótulo da categoria morfo-sintática correspondente. Para tanto, executam a resolução de ambigüidades lexicais e o tratamento de palavras desconhecidas. Estes sistemas têm conseguido atingir uma precisão surpreendentemente alta utilizando apenas recursos modestos [CHU93].

Contudo, até o momento, não havia sido projetado nenhum sistema deste tipo para a Língua Portuguesa. Resolveu-se, então, construir um. Foram realizados o projeto, a implementação e os testes de um sistema deste tipo, em conjunto com o Grupo de Processamento de Linguagem Natural, da Universidade Nova de Lisboa, em Portugal. No entanto, como os recursos existentes para a Língua Portuguesa são, ainda, muito escassos, deve-se tentar utilizá-los da melhor forma possível. Portanto, decidiu-se fazer uma avaliação de certos aspectos do sistema, considerados importantes. Esta avaliação, teve por objetivo encontrar um padrão de comportamento que resultasse na precisão máxima do sistema. O primeiro aspecto avaliado foi a influência da qualidade do “corpus” de treinamento na precisão. Como segundo aspecto, foi verificada a relação existente entre o tamanho do “corpus” de treinamento e a precisão obtida na marcação de um texto. Por último, foi verificada a relação entre as palavras desconhecidas e a precisão do sistema. São estas avaliações que constituem o objetivo principal deste trabalho.

Inicialmente, no capítulo 2, serão apresentados os “corpora”¹ mais frequentemente utilizados em pesquisas na área. Todavia, o enfoque principal ficará

¹ Denomina-se corpus a um conjunto de textos escritos em uma língua. Corpora é o plural de corpus.

com a apresentação do que foi utilizado neste trabalho: o Radiobras Corpus. Alguns dos conjuntos de rótulos utilizados na literatura serão apresentados; serão, também, discutidos os critérios que devem ser adotados na construção de um conjunto de rótulos. Será introduzido o conceito de ambigüidade lexical e os problemas que isto acarreta. Após, serão explicados os formalismos utilizados: a Teoria da Informação de Shannon, o Modelo de N-Gramas e os “Hidden Markov Models”. Por último, será feita uma introdução à marcação automática de categorias morfo-sintáticas, explicando a sua utilização, sendo também discutidas as dificuldades encontradas para se fazer esta marcação em uma língua como o Português.

No capítulo 3, serão apresentadas e comparadas duas abordagens para a construção de sistemas rotuladores de categorias morfo-sintáticas: a abordagem estatística e a baseada em regras. Após, será explicada em detalhes a abordagem estatística usando “Hidden Markov Models” (HMMs), com suas vantagens e desvantagens, bem como os algoritmos usados por esta técnica.

O capítulo 4 será dedicado à descrição detalhada do sistema rotulador utilizado neste trabalho; cada passo necessário à sua construção será descrito. O sistema rotulador implementado, segue a abordagem estatística descrita por Church [CHU88] e utiliza o conceito de classes de ambigüidade, como sugerido por [CUT92].

No quinto capítulo, a avaliação do sistema será apresentada, com base nos três critérios anteriormente descritos. Cada um dos experimentos será detalhadamente descrito, bem como os resultados encontrados nos mesmos.

Por fim, no capítulo 6, estão as conclusões deste trabalho e sugestões para futuras experiências e aprimoramentos.

2. CONCEITOS BÁSICOS

2.1 Corpora Disponíveis

Denomina-se “corpus” a uma coleção de textos escritos em uma determinada língua (Português, Inglês, Chinês, etc). A este “corpus” pode-se atribuir uma marcação lingüística qualquer (categoria morfo-sintática, seqüência de fonemas, etc). Pode-se, por exemplo, associar a cada uma das palavras do “corpus”, o rótulo da categoria morfo-sintática correspondente. Outro exemplo seria relacionar a cada palavra a seqüência de fonemas que a formam.

O interesse na utilização de “corpora” marcados teve um grande crescimento nos últimos anos, principalmente por pesquisadores das áreas de lingüística e lingüística computacional. Este fato se deve à facilidade proporcionada para observação da ocorrência de fenômenos lingüísticos em um “corpus” marcado.

Brill [BRI93b] cita algumas aplicações de “corpora” marcados na área da lingüística, como:

- o uso de “corpora” marcados para o estudo de elipses do Sintagma Verbal (SV);
- o estudo da marcação sintática de um “corpus”, ajudando no desenvolvimento de uma teoria sobre como o ser humano resolve ambigüidades sintáticas.

Já, na lingüística computacional, “corpora” marcados têm sido usados, entre outros propósitos, para fazer o treinamento de rotuladores de categorias morfo-sintáticas, como o desenvolvido por Church [CHU88].

Muito esforço está sendo feito no sentido de se conseguir reunir mais “corpora”. Entidades como a “Association for Computational Linguistics” Data Collection Initiative” (ACL/DCI), a “European Corpus Initiative” (ECI), ICAME, o “British National Corpus” (BNC), o “Linguistic Data Consortium” (LDC), o “Consortium for Lexical Research” (CLR), “Eletronic Dictionary Research”, entre

outras, têm colocado a disposição seus “corpora” e, com isto, permitem que as pesquisas, na área, tenham prosseguimento.

Contudo, ainda não é muito grande o número de “corpora” marcados disponíveis. Na Língua Inglesa, pode-se citar alguns “corpora”, tais como o “Brown Corpus”, o “Penn Treebank”, o “Wall Street Journal”, o “LOB Corpus”, o “Birmingham Corpus” e o “SUSANNE” [CHU93]. Para realizar este trabalho, está-se usando dois dos “corpora” existentes na Língua Portuguesa: o Lusa Corpus (Portugal) e o Radiobras Corpus (Brasil).

2.2 Conjuntos de Rótulos

Quando se deseja fazer a marcação de um “corpus” com os rótulos de categorias morfo-sintáticas, deve-se, primeiramente, definir quais os rótulos que serão utilizados. Um rótulo de categoria morfo-sintática pode indicar a categoria morfológica (substantivo, adjetivo,...) a qual a palavra pertence. Porém, além disto, pode indicar uma classificação mais específica. Um exemplo disto é o caso do rótulo **BEZ** (verbo “to be”, no presente, terceira pessoa do singular), que faz parte do conjunto de rótulos do Brown Corpus [CHU88]. Este rótulo, além de indicar que a palavra é um verbo, indica qual é o verbo (“to be”), o tempo (presente), a pessoa (3ª) e o número (singular).

Pode-se dizer que, as palavras pertencentes a um determinado rótulo, compartilham algumas propriedades sintáticas. Alguns dos rótulos têm uma delimitação bem clara dos seus componentes, enquanto outros são mais amplos. Como exemplo, pode-se comparar o rótulo **AT** com **JJ**, ambos provenientes do Brown Corpus. O rótulo dos artigos, **AT**, é um conjunto fechado, que tem como elementos todos os artigos conhecidos na Língua Inglesa. Já **JJ**, o rótulo que designa os adjetivos, é um conjunto aberto, pois, por maior que seja, é muito difícil que contenha todos os adjetivos conhecidos.

Além dos rótulos referentes às palavras de uma sentença, a pontuação (como ponto final, reticências, vírgula, etc.) também tem seu rótulo apropriado. Abaixo, são apresentadas a sentença 2.1 e a marcação feita para ela, usando um

pequeno conjunto de rótulos, apresentado na tabela 2.1. Este conjunto de rótulos é um subconjunto do conjunto de rótulos usado para marcar o Radiobras Corpus, e será apresentado na seção 4.1.

Sentença 2.1- Nós tínhamos feito um acordo para a divisão do patrimônio.

Nós PPR tínhamos VTER feito VPP um ART acordo N para PREP a
ART divisão N do CONT patrimônio N . PTO

Tabela 2.1 - Rótulos de Morfo-Sintáticos usados para a Sentença 2.1

| Rótulos | Significado |
|----------------|---------------------------------|
| ART | Artigo |
| CONT | Contração de Preposição |
| PPOA | Pronome Pessoal do Caso Oblíquo |
| PPR | Pronome Pessoal do Caso Reto |
| PREP | Preposição |
| PTO | Ponto Final |
| N | Substantivo |
| VPP | Verbo no Particípio |
| VTD | Verbo Transitivo Direto |
| VTER | Verbo Ter |

2.2.1 Definindo o Conjunto de Rótulos

A definição do conjunto de rótulos a ser usado é uma questão complexa que varia de língua para língua. Ela envolve a análise de conhecimentos lingüísticos e de questões de performance. Entre os itens a serem analisados, pode-se citar:

- o tamanho do conjunto de rótulos;
- a quantidade de informação lingüística necessária para realizar uma determinada tarefa - pode-se definir distinções como número, gênero, etc.;
- o nível da análise - o conjunto de rótulos não deve incluir distinções que não possam ser resolvidas a este nível de análise.

Nas seções seguintes, serão descritos cada um destes itens.

2.2.1.1 Tamanho do Conjunto de Rótulos

O primeiro item - o tamanho do conjunto de rótulos - está ligado a questões de performance. O objetivo é fazer com que o processo de marcação (“tagging”) seja tão eficiente quanto possível.

Desta forma, quando se usa um modelo estatístico de marcação automática, deve-se ter o cuidado de evitar que o conjunto de rótulos seja grande, pois o número de rótulos existentes está relacionado ao número de parâmetros do sistema rotulador que devem ser estimados. E, quanto maior for o número de parâmetros, maior deverá ser o tamanho do “corpus” usado para estimá-los.

Serão citados, a seguir, alguns dos conjuntos de rótulos encontrados na literatura, para dar uma idéia da enorme variação no que diz respeito ao tamanho destes conjuntos. Para a Língua Inglesa, tem-se o “PennTreebank” com 48 rótulos diferentes, que é baseado no “Brown Corpus”, que tem, ao todo, 87 rótulos. O “LOB Corpus” tem 135 rótulos, o “SUSANNE”, 425 e o usado por Cutting et al [CUT92], 38.

Chanod e Tapanainen [CHD95] definem 88 rótulos diferentes para o Francês, sendo que se pode obter um total de 353 seqüências possíveis e se forem consideradas as seqüências com **clíticos**, 6525. Para o Chinês, Chang e Chen [CHA93] apresentam um total de 57, sendo 46 denominados regulares e 11, especiais.

2.2.1.2 Especificidade do Conjunto de Rótulos

Este aspecto está relacionado com a capacidade, que o conjunto de rótulos tem, de fazer as distinções lingüísticas necessárias para uma determinada finalidade. Isto implica uma definição bastante clara da tarefa a ser executada e dos requisitos básicos necessários para a execução de tal tarefa. A seguir, deve-se modelar estes requisitos básicos no conjunto de rótulos.

Isto envolve a definição de um conjunto de rótulos mais específico ou menos específico. Assim, caso seja necessário, pode-se definir, por exemplo, características como gênero, número, pessoa, modo, tempo, grau, etc, se as mesmas

forem relevantes ao resultado final. Além disto, deve-se ter em mente que cada língua tem características diferentes que devem ou não ser consideradas.

Pode-se subdividir as línguas de acordo com o grau de inflexão apresentado. Há as línguas altamente flexionadas como o Grego ou o Húngaro, as que têm um grande número de inflexões como o Português ou o Francês, e as línguas com menos inflexões como a Língua Inglesa, por exemplo. Note-se, contudo, que, para as línguas altamente flexionadas, se torna impossível construir um sistema rotulador simples, se todas as inflexões possíveis na língua forem representadas no conjunto de rótulos, pois neste caso haveria um número muito grande de rótulos. Já uma língua, como o Português ou o Francês, apresenta um grau de complexidade bem mais reduzido, sendo flexionada em gênero, número e pessoa, entre outros. Uma Língua mais simples ainda, como o Inglês, é flexionada apenas na pluralidade dos substantivos e em algumas propriedades verbais. Foram feitas algumas experiências sobre este assunto por pesquisadores da área como Chanod e Tapanainen [CHD95], para o Francês e por Elworthy [ELW95], para o Francês, Inglês e Sueco.

2.2.1.3 Generalidade do Conjunto de Rótulos

O conjunto de rótulos deve ser o mais completo possível. Deve-se tentar modelar todas as características da língua que são relevantes para a aplicação em questão. Porém, deve-se estudar cuidadosamente quais os atributos pertinentes a este nível de análise. Não se deve sobrecarregar o sistema rotulador com considerações que, possivelmente, não poderão ser resolvidas por ele. Além disto, deve-se ter sempre em mente que os sistemas rotuladores têm apenas um conhecimento restrito da *vizinhança* (bigramas: uma palavra; e trigramas: duas palavras, discutido na seção 2.5).

Outro aspecto relacionado com o conjunto de rótulos é o problema da ambigüidade. Nas experiências feitas por Elworthy [ELW95] sobre a ambigüidade para o Francês, o autor demonstra que informações de gênero devem ser evitadas, porque, além de não auxiliarem na resolução da ambigüidade, ainda a aumentam. Foi constatado que onde havia informação relativa a gênero, o grau de ambigüidade das palavras aumentava. Nas experiências de precisão com palavras desconhecidas, foi

novamente constatado que a presença do gênero contribuía para a diminuição desta precisão. A maior precisão foi obtida com os seguintes fatores:

- 36 rótulos no conjunto de rótulos;
- não uso de informação de gênero;
- não uso de informação de número;
- não uso de informação de pessoa;
- tratar os verbos “avoir” e “etre” de maneira diferenciada dos demais verbos.

E os resultados obtidos foram os seguintes:

- **grau de ambigüidade das palavras do texto: 56,08%;**
- **precisão com palavras ambíguas: 96,34%;**
- **precisão com palavras desconhecidas: 52,59%.**

Chanod e Tapanainen acrescentam ainda alguns dados a este estudo como, por exemplo, experiências com o uso de informações relativas aos tempos verbais. É recomendado que não se usem tais informações porque se quer evitar a sobrecarga do sistema rotulador com qualquer tipo de informação que não se possa resolver neste nível da análise. Além disto, após a análise, este tipo de informação pode ser facilmente recuperado através de consultas ao dicionário. Pois, caso o tempo verbal não seja ambíguo, nenhuma informação foi perdida e, no caso de haver tal ambigüidade, de qualquer forma, o rotulador não a resolveria de maneira confiável.

Há ainda o problema do modo verbal, que não é usado porque há muitos casos de ambigüidade entre o presente do indicativo e o presente do subjuntivo que também só podem ser resolvidos com uma análise mais ampla da sentença. Há problemas de ambigüidade também com a informação de pessoa verbal, por exemplo, nos casos em que a primeira pessoa do singular é igual a terceira pessoa do singular.

Frente aos problemas acima citados, observa-se a necessidade de analisar as vantagens e as desvantagens de especificar determinadas características no conjunto de rótulos. Como cada língua tem seus requisitos específicos, cabe aqui

salientar a necessidade de se realizarem estudos, como os apresentados em [ELW95], voltados para a Língua Portuguesa.

2.3 Ambigüidade Lexical

Quando se faz a marcação de um “corpus”, um dos principais problemas que costuma ocorrer é o da ambigüidade das palavras ou ambigüidade lexical. Por ambigüidade lexical se entende o fato de uma palavra poder ter mais de um rótulo associado a ela. Isto pode ser visto na tabela a seguir, que mostra os possíveis rótulos ou “tags” para a palavra “a”:

Tabela 2.2 - Informação Lexical

| | | |
|---|------|---------------------------------|
| a | ART | Artigo |
| | PREP | Preposição |
| | PPOA | Pronome Pessoal do Caso Oblíquo |

Erroneamente, se pensa que a ambigüidade lexical é restrita apenas a umas poucas palavras e que este não é um problema importante. Contudo, a maioria das palavras de nossa língua têm mais de um rótulo associado a elas. Merialdo, em seu estudo para a Língua Inglesa [MER94], constatou que, no “corpus” utilizado, quase metade das palavras tem um único rótulo e cerca de 25% das palavras têm somente dois rótulos possíveis.

Pode-se ter uma idéia da extensão do problema da ambigüidade lexical observando-se o que Church [CHU93] afirma sobre este problema:

“Muitas pessoas, que não trabalham com Linguística Computacional, têm uma forte intuição de que a ambigüidade lexical não é exatamente um problema. É muito comum acreditarem que a maioria das palavras tem apenas um rótulo referente a sua categoria morfo-sintática, e que as poucas exceções como “table” são facilmente resolvidas através do contexto na maioria dos casos. Por outro lado, muitos especialistas em Linguística Computacional acreditam que a ambigüidade lexical é um grande problema; é dito que praticamente todas as “content words”

podem ser usadas como substantivos, verbos ou adjetivos e que nem sempre o contexto local é adequado para resolver a ambigüidade.”

Felizmente, na maioria dos casos, esta ambigüidade PODE SER e É resolvida através de uma análise do contexto local em que a palavra ocorre. Analisando-se a Sentença 2.2:

Sentença 2.2: Deu o presente a ele.

pode-se ver pela vizinhança das palavras ambíguas as melhores opções de rótulos para cada uma delas: neste caso, “a” é uma preposição.

2.3.1 Classes de Ambigüidade

Como já foi explicado, uma mesma palavra pode ter vários rótulos diferentes associados a ela, ocasionando o que se denomina de ambigüidade lexical. Por exemplo, a palavra “avanço” pode ser tanto um substantivo quanto um verbo (“avançar”). O mesmo ocorre com a palavra “repouso” (substantivo “repouso” e verbo “repousar”). Neste caso, porque não juntá-las em uma única classe chamada SUBSTANTIVO-VERBO, que contém todas as palavras que podem ser tanto substantivos quanto verbos? A este tipo de classe se dá o nome de *classe de ambigüidade*. Quando há várias palavras diferentes com o mesmo conjunto de rótulos associado, estas palavras formam uma classe de ambigüidade. Porém, somente um dos rótulos desta classe estará correto de acordo com o contexto em que a palavra aparecer.

O conceito de classes de ambigüidade, também denominadas classes de equivalência de palavras, foi apresentado pela primeira vez por Kupiec, citado em [CUT92], e, posteriormente, utilizado em muitos trabalhos, tais como [CUT92] e [CHA93]. Uma grande vantagem da utilização deste conceito é a diminuição no número de parâmetros a serem estimados no modelo utilizado. Isto é muito importante, pois como explicado anteriormente, o número de parâmetros a ser estimado depende do número de rótulos. Além disto, o uso de classes de ambigüidade não diminui significativamente a precisão. E, segundo [CHA93], o sistema rotulador apresenta mais robustez quando utiliza este conceito.

No trabalho descrito em [CUT92], todas as palavras são representadas pelas suas classes de ambigüidade, com exceção das mais comuns, que são representadas individualmente, uma vez que há dados suficientes para produzir estimativas robustas. Assim, o vocabulário do “Brown Corpus”, de 50.000 palavras pode ser reduzido para aproximadamente 400 classes de ambigüidade.

Além disto, com poucas classes de ambigüidade, pode-se conseguir uma grande abrangência, pois é muito improvável que, a adição de novas palavras ao dicionário, vá resultar na necessidade de incluir novas classes de ambigüidade e conseqüentemente na reformulação do sistema rotulador.

2.4 Teoria da Informação de Shannon

A Teoria da Informação de Shannon também é conhecida como Teoria da Comunicação. Ela foi originariamente desenvolvida para modelar a comunicação por um meio onde houvesse ruídos (“noisy channel”), tal como a linha telefônica. Entretanto, esta teoria pode também ser transportada para outros domínios, tais como marcação automática de categorias morfo-sintáticas.

Esta teoria supõe a existência de um canal ou meio com ruído. Neste meio, tem-se, como entrada, uma seqüência de texto (E) correta e, como saída, a mesma seqüência de texto (I) só que com erros, corrompida.



Figura 2.1 - Canal com Ruído

Supondo o exemplo da linha telefônica: têm-se duas pessoas conversando ao telefone. A pessoa 1 falou pelo telefone a sentença E e a pessoa 2 ouviu a sentença corrompida I.

A questão que se apresenta aqui é como obter a sentença correta E a partir da sentença ouvida I? Uma opção é tentar deduzir todas as possíveis sentenças

de entrada E e descobrir qual é a mais provável, a qual se denomina \hat{E} . Esta será a que tiver o resultado mais alto em $P(E | I)$:

$$\hat{E} = \max_E P(E|I) = \max_E P(E)P(I|E)$$

Fórmula 2.1

onde a função \max tem como retorno o argumento que apresentar o maior valor. As probabilidades usadas $P(E)$ e $P(I|E)$ são, respectivamente, a probabilidade a priori e a probabilidade conjunta. $P(E)$ é a probabilidade de se ter E como entrada do canal de ruído. No exemplo da linha telefônica, é a probabilidade da pessoa 1 falar a sentença E . $P(I|E)$ é a probabilidade conjunta de I aparecer na saída do canal quando E é apresentado na entrada deste. Esta probabilidade terá um valor alto se I tiver semelhanças com E . Caso contrário, esta probabilidade será baixa. No exemplo, esta é a probabilidade da pessoa 1 falar “intenção” e a pessoa 2 ouvir “invenção”.

Como já foi explicado, a Teoria da Informação de Shannon foi originalmente pensada e modelada para simular o ato da comunicação usando um meio que contivesse ruídos. Mas, devido ao sucesso que obteve, esta teoria influenciou toda uma série de trabalhos, e, graças a ela, muitas conquistas foram obtidas nesta área. É ela que serve de base para a construção do sistema rotulador, apresentado no capítulo 4.

2.5 Modelo de N-Gramas

O problema da ambigüidade lexical, como foi explicado anteriormente, é uma constante no processamento de linguagem natural, especialmente quando se trata com textos de domínio irrestrito. Os seres humanos, de um modo geral, resolvem a ocorrência de uma palavra ambígua quando analisam o contexto no qual a palavra está inserida.

Por exemplo, a palavra “a” pode ser, entre outras coisas, um artigo ou uma preposição. Quando analisada isoladamente, é impossível dizer qual destes dois rótulos se deve escolher. Contudo, quando se analisa a sentença:

Sentença 2.3: A gaivota planava majestosa.

vê-se que o rótulo mais adequado é artigo. Mas na sentença:

Sentença 2.4: Vamos levá-la a um restaurante na cidade.

o “a” deve ser rotulado como preposição, tendo em conta os rótulos das palavras vizinhas (verbo e pronome oblíquo).

Para explorar o conceito de **contexto**, ou **vizinhança**, quando se faz a marcação automática (ou “tagging” automático) de uma sentença, usa-se o modelo de **n-gramas** [MER93, CUT92]. Com o modelo de n-gramas, é possível explorar este conceito de *vizinhança*. O modelo de n-gramas define que, para cada palavra, deverão ser analisadas as ‘n-1’ palavras vizinhas.

Mas quanto contexto deverá ser analisado para resolver um caso de ambigüidade? Quantas palavras vizinhas? A maioria dos problemas de ambigüidade lexical podem ser solucionados com informações apenas sobre a vizinhança de precedência mais próxima (1 ou 2 palavras precedentes). Os modelos mais utilizados são os de bigrama (n=2) e trigrama (n=3) [CUT92, SCH94b, SCH95, CHU88] que apresentam bons resultados, visto que conseguem capturar uma parte importante da vizinhança e produzem resultados confiáveis. Ou seja, os modelos de bigrama e trigrama podem perfeitamente se ajustar a estas necessidades. No modelo de bigrama um rótulo depende apenas do rótulo anterior (2-1=1). Já o modelo de trigramas especifica que um rótulo depende dos dois rótulos que o precedem imediatamente (3-1=2).

No entanto, deve-se analisar cuidadosamente, para cada caso, a relação de custo x benefício de uma vizinhança maior ou menor. Quanto maior o “n”, maior a precisão, porém, maior o custo.

2.6 Modelos de Markov

Pode-se definir um Modelo de Markov [CHK93, CUT92, KEM94] como um processo estocástico reconhecedor ou um gerador de uma linguagem específica. Reconhecedor quando aceita uma seqüência de símbolos de entrada como

pertencente a uma determinada linguagem. Gerador quando cria seqüências de símbolos a partir de uma linguagem específica.

Os Modelos de Markov podem ser basicamente de dois tipos: Cadeia de Markov e “Hidden Markov Model” (HMM). Para que o HMM, que é utilizado neste trabalho para construir o sistema rotulador descrito no capítulo 4 possa ser melhor entendido, será dada um breve introdução à Cadeia de Markov.

2.6.1 Cadeia de Markov

Uma Cadeia de Markov é constituída por um conjunto de estados interligados por um conjunto de transições. Além disto, tem definido um alfabeto de símbolos de saída.

Cada um dos estados tem associado a ele um ou mais símbolos do alfabeto, que serão emitidos quando se seguir a transição correspondente a ele.

Associada a cada uma das transições que partem de um estado, há uma probabilidade que rege a passagem deste estado origem a um estado destino. Observa-se aqui que o resultado da soma das probabilidades que partem de um estado deve ser igual a 1, figura 2.2.

As cadeias de Markov têm as seguintes características:

- um estado inicial da cadeia que é por onde se começa a percorrer a cadeia;
- um estado final, representado por círculos concêntricos;
- ao sair de um estado e percorrer uma transição até outro estado, emite-se um símbolo de saída.

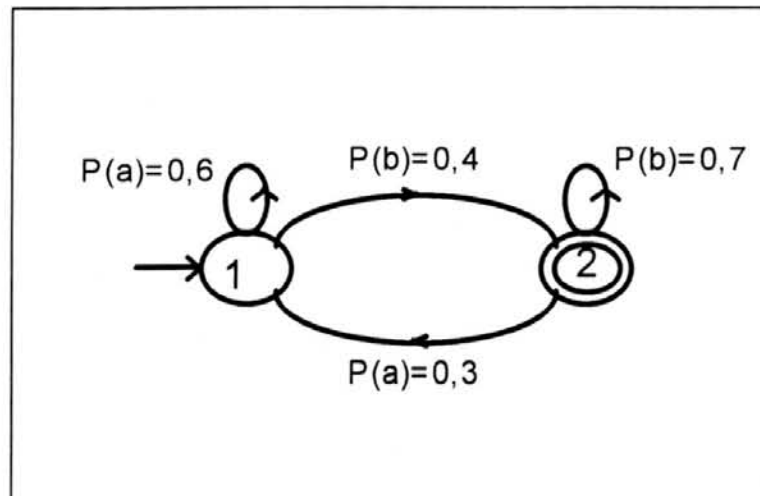


Figura 2.2- Cadeia de Markov

Na cadeia de Markov apresentada na figura 2.2, pode-se distinguir os seguintes componentes:

- um conjunto formado pelos estados 1 e 2 onde 1 é o estado inicial e 2 é o final;
- as transições: $1 \xrightarrow{a} 1, 1 \xrightarrow{b} 2, 2 \xrightarrow{a} 1, 2 \xrightarrow{b} 2$;
- as probabilidades²:

$$P(1 \xrightarrow{a} 1) = 0,6$$

$$P(1 \xrightarrow{b} 2) = 0,4$$

$$P(2 \xrightarrow{a} 1) = 0,3$$

$$P(2 \xrightarrow{b} 2) = 0,7$$

- o alfabeto de símbolos de saída: {a,b}.

A probabilidade de reconhecer uma seqüência de símbolos é dada pelo produto das probabilidades associadas às transições percorridas, durante o reconhecimento desta seqüência.

Assim, a probabilidade resultante do reconhecimento da seqüência 'aaabbab', na cadeia de Markov da figura 3.1, após percorrer a seqüência de estados 11112212, é:

²A expressão $P(1 \xrightarrow{a} 1) = 0,6$ tem a seguinte leitura: a probabilidade da transição do estado 1 para o estado 1 com a emissão do símbolo 'a' é igual a 0,6.

$$P(1 \xrightarrow{a} 1) = 0,6$$

$$P(1 \xrightarrow{a} 1) = 0,6$$

$$P(1 \xrightarrow{a} 1) = 0,6$$

$$P(1 \xrightarrow{b} 2) = 0,4$$

$$P(2 \xrightarrow{b} 2) = 0,7$$

$$P(2 \xrightarrow{a} 1) = 0,3$$

$$P(1 \xrightarrow{b} 2) = 0,4$$

com $P(aaabbab) = 0,0072576$.

A Cadeia de Markov pode ser vista como um autômato finito determinístico, ou seja, dada uma saída, é sempre possível determinar a seqüência de estados que gerou tal saída, bem como a probabilidade associada, como se pôde observar no exemplo acima.

A seguir, será apresentado um tipo especial não-determinístico de Modelo de Markov: o “Hidden Markov Model”.

2.6.2 Hidden Markov Model

Um “Hidden Markov Model”, figura 2.3, é uma generalização de Cadeias de Markov, onde se tem mais que uma transição com o mesmo símbolo partindo do mesmo estado. Desta forma, partindo-se de um estado origem, com um mesmo símbolo, pode-se chegar a vários estados destino diferentes. Por exemplo, na figura 2.3, o símbolo “a” está na transição do estado 1 para o estado 1 e na transição do estado 1 para o 2. Devido a este fato, para uma determinada seqüência de símbolos, pode-se ter várias seqüências de estados (caminhos) que resultem nesta mesma saída.

A definição formal de um “Hidden Markov Model” inclui os seguintes componentes:

S , que é um conjunto finito de estados³;

s^1 , que é o estado inicial, onde $s^1 \in S$;

³ A notação usada define que c^i denota o i -ésimo elemento do conjunto C ; por outro lado c_i se refere ao símbolo de saída no tempo i .

T , que é o conjunto de transições e

W que é o conjunto de símbolos de saída (ou alfabeto).

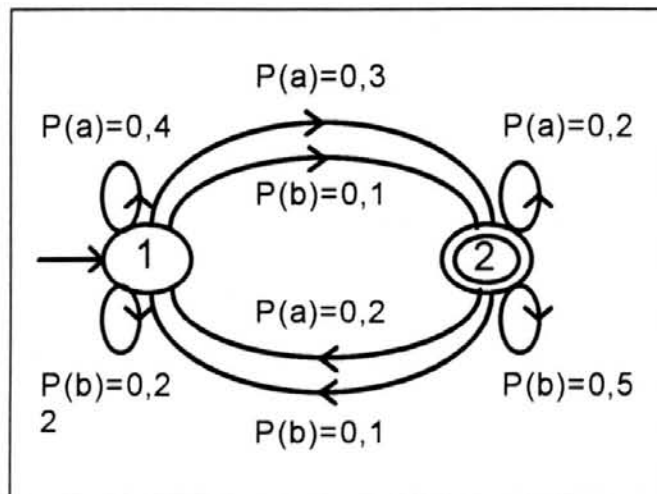


Figura 2.3 - HMM Com Dois Estados

No exemplo acima, tem-se:

$$S = \{1,2\};$$

$$s^1=1;$$

$$T = \{P(1 \xrightarrow{a} 1) = 0,4; P(1 \xrightarrow{b} 1) = 0,2; P(1 \xrightarrow{a} 2) = 0,3; P(1 \xrightarrow{b} 2) = 0,1; \\ P(2 \xrightarrow{a} 2) = 0,2; P(2 \xrightarrow{b} 2) = 0,5; P(2 \xrightarrow{a} 1) = 0,2; P(2 \xrightarrow{b} 1) = 0,1\}$$

$$W = \{a,b\}.$$

Devido à complexidade dos HMMs, quando comparados às Cadeias de Markov, torna-se mais complicado determinar o caminho que foi percorrido na tentativa de reconhecer uma dada seqüência de símbolos. Uma vez que se pode ter muitos caminhos percorridos para uma dada seqüência de símbolos, a probabilidade desta seqüência é o resultado da soma das probabilidades de todos os caminhos que nela resultam. A probabilidade de um caminho, por sua vez, é o produto de todas as transições do caminho.

Por exemplo, dado o HMM da figura 2.3 e a seqüência 'aab', têm-se os caminhos que aparecem na figura 2.4.

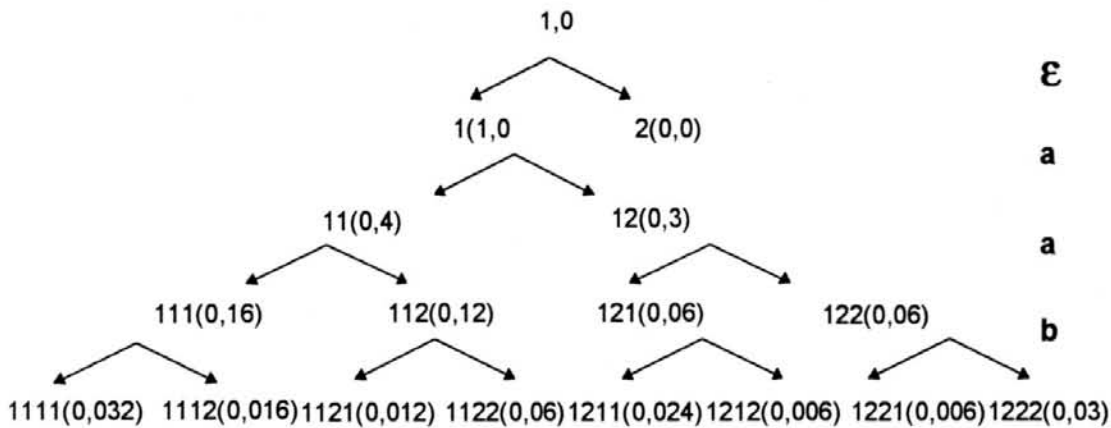


Figura 2.4 - Árvore da Sequência "aab"

Ou seja: 1111, 1112, 1121, 1122, 1211, 1212, 1221, 1222. Este é um problema de crescimento exponencial. Se tem oito possíveis caminhos para uma sequência de apenas 3 símbolos. Além disto, as probabilidades variam de 0,006 a 0,06, o que significa a diferença entre uma certeza de 0,6% e uma de 6%.

No caso de se tratar de uma sequência muito grande, torna-se complicado calcular todos os possíveis caminhos, visto que se trata de um problema exponencial e um grande número de caminhos resultantes é gerado. Para contornar este problema, pode-se utilizar o algoritmo de Viterbi [CHK93, MER94], explicado no capítulo 3, que determina o caminho mais provável para uma dada sequência. Ainda no capítulo 3, serão explicados outros algoritmos que podem ser aplicados aos HMMs.

2.7 Marcação de Categorias Morfo-Sintáticas

Pode-se definir a marcação de uma sentença como a ação de atribuir, para cada palavra de uma sentença, um rótulo referente a sua categoria morfo-sintática [MAG94]. Este rótulo de categoria morfo-sintática é atribuído com base no contexto em que a palavra aparece dentro desta sentença. Assim, dada uma sequência de palavras e um conjunto de rótulos, associa-se a cada palavra o seu respectivo rótulo. Este é o processo usado para fazer a marcação de um "corpus".

Numa definição mais formal, tem-se uma sentença W formada por um conjunto de palavras w_1, w_2, \dots, w_n aos quais é atribuída uma sequência de rótulos T , formada por t_1, t_2, \dots, t_n . O par formado por (W, T) constitui-se de um alinhamento: cada palavra w_i está relacionada com o rótulo t_i correspondente no alinhamento.

2.7.1 O Problema da Ambigüidade Lexical

Nos casos em que ocorre ambigüidade lexical com algumas das palavras de uma sentença, conforme o que foi explicado em 2.3, pode-se ter mais de um único alinhamento possível para a sentença. Por exemplo, considerando a sentença 2.5 e os rótulos da tabela 2.2 tem-se:

Sentença 2.5: Deu o presente a ele .

| | | | | | |
|-----|------|----------|------|-----|-----|
| Deu | o | presente | a | ele | . |
| VTD | ART | N | ART | PPR | PTO |
| | PPOA | | PPOA | | |
| | | | PREP | | |

Figura 2.5 - Marcação da sentença 2.5

Onde os possíveis alinhamentos são:

| | | | | | |
|-----|------|---|------|-----|-----|
| VTD | ART | N | ART | PPR | PTO |
| VTD | PPOA | N | ART | PPR | PTO |
| VTD | ART | N | PPOA | PPR | PTO |
| VTD | PPOA | N | PPOA | PPR | PTO |
| VTD | ART | N | PREP | PPR | PTO |
| VTD | PPOA | N | PREP | PPR | PTO |

Figura 2.6 - Possíveis alinhamentos para a sentença 2.5

| | | | | | |
|-----|-----|----------|------|-----|-----|
| Deu | o | presente | a | ele | . |
| VTD | ART | N | PREP | PPR | PTO |

Figura 2.7 - Alinhamento correto da sentença 2.5

Logo, há seis alinhamentos diferentes para uma sentença pequena: apenas cinco palavras.

Pode-se perceber que, mesmo que uma sentença tenha muitos alinhamentos possíveis, somente um será correto de acordo com o contexto. Apenas um deles terá a seqüência correta de rótulos que correspondam às palavras da sentença. Todos os demais serão incorretos de acordo com o contexto.

Sendo assim, o procedimento de “tagging” deve saber selecionar o alinhamento correto dentre todos os possíveis para uma dada sentença. E, para medir a precisão com que este procedimento realiza a seleção dos alinhamentos, há, dois critérios muito citados na literatura:

- precisão a nível de sentença: este critério mede o número de sentenças corretamente rotuladas. Para que um alinhamento de uma sentença seja considerado correto, a todas as palavras desta sentença deve ser atribuído o rótulo correto;
- precisão a nível de palavra: neste critério, é calculado o número de palavras corretamente rotuladas. Este é o critério usado pela maioria dos trabalhos citados na literatura.

Observa-se aqui que o último critério sempre produz resultados maiores do que o primeiro. Isto porque, como já foi explicado, todas as palavras de uma sentença tem que ter seu rótulo correto para que o alinhamento da sentença seja correto.

2.8 Marcação Manual x Marcação Automática

Para que os estudos e análises feitos com base nos “corpora” marcados tenham validade e reflitam bem a realidade, é preciso usar “corpora” relativamente grandes. Por relativamente grande, se entende um “corpus” de um milhão ou mais de palavras, que é o tamanho comumente utilizado. Por exemplo, o “Brown Corpus” marcado [CHU93] possui aproximadamente 1.000.000 de palavras. Os trabalhos feitos por [CHU88] e [SCH94a] usam, respectivamente, 1.000.000 e 2.000.000 de palavras.

Entretanto, a marcação manual de um “corpus” desta magnitude, se torna quase impraticável, visto ser uma tarefa muito extenuante.

Para auxiliar nesta tarefa, foram criados programas que fazem a marcação automática dos “corpora”. Estes sistemas são chamados de *Rotuladores de Categorias Morfo-Sintáticas* e serão explicados no capítulo seguinte.

3. ROTULADORES DE CATEGORIAS MORFO-SINTÁTICAS

Rotuladores de categorias morfo-sintáticas são programas que têm como entrada uma seqüência de palavras. Após uma análise destas, produzem como saída a seqüência de rótulos correspondentes, figura 3.1.



Figura 3.1 - Sistema Rotulador

Isto significa que o objetivo de um rotulador é o de encontrar a seqüência mais provável de rótulos (ou “tags”) que corresponda à seqüência de palavras dada. Desta forma, eles podem ser usados para realizar a marcação automática de um “corpus”.

Na seqüência de palavras usada como entrada para o rotulador, podem haver palavras que sejam ambíguas e palavras desconhecidas a ele. O sistema rotulador deve saber lidar com estes tipos de problema.

Há duas principais filosofias que regem a construção de sistemas rotuladores: uma se baseia na intuição dos lingüistas, através de regras e a outra se baseia nos padrões encontrados nos dados analisados, utilizando métodos estatísticos.

3.1 Sistemas Rotuladores Baseados em Regras

Na abordagem baseada em regras, são construídas regras a partir de abstrações feitas pelos lingüistas sobre os paradigmas e os sintagmas da linguagem, gerando, como resultado, uma gramática. A construção da gramática é feita manualmente, exigindo dedicação e muito esforço por parte dos lingüistas.

Este é o método tradicionalmente usado na construção de sistemas rotuladores de categorias morfo-sintáticas. Uma vez que estes sistemas rotuladores são construídos de acordo com o pensamento do lingüista que os está construindo, este pode implementar as teorias que desejar usando regras. Contudo, como tanto as teorias quanto as generalizações da linguagem devem ser explicitamente definidas, é aconselhável que se determine o domínio específico a ser trabalhado.

Além disto, as regras do sistema são manualmente construídas. Ou seja, para que se possa expressar um domínio relativamente grande, muitas regras são necessárias, e, conseqüentemente, muito trabalho manual precisa ser feito. Quando se quiser testar estas teorias, deve-se usar dados provenientes do domínio específico modelado.

Outro ponto a salientar é que este método normalmente utiliza programação simbólica.

3.2 Sistemas Rotuladores Baseados em Métodos Estatísticos

Os rotuladores estatísticos (ou estocásticos) seguem o pensamento empírico que esteve em alta nos anos 50 e que teve um renascimento nos últimos anos. Church [CHU93] descreve bem o motivo do crescente interesse nos métodos estatísticos no seguinte trecho:

“A ênfase atual nos métodos empíricos, na comunidade de reconhecimento de fala, é uma reação à falha das abordagens baseadas em conhecimento dos anos 70. Isto tem se tornado novamente popular com o objetivo de focalizar as restrições de alto nível da linguagem natural, de modo a reduzir o espaço de busca.”

O pensamento empírico tenta realizar a análise de textos irrestritos. Um sistema rotulador que utiliza estes métodos, por tratar de textos irrestritos, não apresenta limitações quanto ao domínio de aplicação. E, para conseguir tratar com tal

abrangência de domínio, usa técnicas de construção automática de regras. Estas regras são automaticamente inferidas a partir de um “corpus”. Este processo é totalmente automático e não necessita supervisão humana direta. A participação humana, neste caso, se dá somente pela construção de procedimentos de alto nível, para aumentar a precisão do sistema. A construção de um sistema deste tipo envolve inferência e manipulação de dados estatísticos.

As generalizações, que nos métodos baseados em regras devem ser especificamente descritas, nos métodos estatísticos, são automaticamente adquiridas a partir do “corpus”. Esta capacidade do sistema de generalizar a partir dos dados analisados é testada usando dados reais, ou seja, um texto de domínio irrestrito.

3.3 Sistemas Rotuladores Baseados em Regras X Sistemas Rotuladores Estatísticos

Os rotuladores baseados em regras são mais estruturados e podem ser melhor entendidos. Isto facilita muito quando se precisa fazer extensões ao sistema, uma vez que apresenta um modo mais natural de representar as regras.

Como já foi citado, teorias lingüísticas podem ser implementadas usando este método. Pode-se, então, pensar que, graças a isto, grande parte da língua pode ser modelada neste tipo de sistema. Entretanto, existem muitas exceções que, as teorias descritas no sistema, provavelmente, não conseguirão tratar. Devido à grandeza da língua, torna-se extremamente complexo construir manualmente regras para tratar todos os casos e todas as exceções nela presentes. O domínio em que este sistema pode ser aplicado é um domínio limitado pela abrangência das regras descritas. Além disto, a construção de uma teoria não é uma tarefa fácil, pois implica um projeto bem detalhado, e teorias nem sempre são fáceis de projetar [SCA92].

Geralmente, um sistema deste tipo define em torno de 1000 regras. Uma vez que as regras do sistema devem ser manualmente descritas, o custo e o tempo de desenvolvimento e de implementação deste tipo de sistema é extremamente alto. E, como estas regras são modeladas para um domínio muito específico, este sistema dificilmente poderá ser levado para outro domínio de aplicação.

Além destes aspectos, deve-se salientar que a precisão e a velocidade, apresentadas por tais rotuladores, são geralmente baixas. Além disto, o custo da construção de tal sistema é bastante alto.

Os métodos estatísticos, por outro lado, se destacam pela ênfase dada ao trabalho com textos “reais”, ou seja, textos de domínio irrestrito. Para tanto, oferece a facilidade da extração automática de padrões, com base em um “corpus”.

Além disto, possuem uma grande vantagem que é a possibilidade de repetir experiências já realizadas, graças à disponibilidade que se tem do “corpus” de treinamento. Isto significa que se pode repetir a experiência usando o mesmo “corpus” de treinamento e, que o resultado desta repetição será coerente com o resultado anteriormente obtido. Logo, pode-se perfeitamente duplicar os experimentos para confirmar os resultados obtidos.

Todavia, uma vez que o conhecimento do sistema é inferido a partir de dados reais, muitas vezes os resultados obtidos podem divergir das intuições lingüísticas. Isto se deve ao fato de o conhecimento do sistema refletir o que a ele foi apresentado: um “corpus” de textos reais. Este “corpus”, provavelmente, conterà erros difíceis de detectar quando se trabalha com milhões de palavras.

Além disto, o processo como o sistema infere e expressa o seu conhecimento, pode tornar difícil qualquer tentativa de inferência humana a partir disto. Neste caso, um sistema baseado em regras, se apresenta muito mais natural e fácil de ser compreendido.

Muitas vezes, para um sistema rotulador estatístico obter uma precisão maior, é necessário implementar alguns procedimentos de alto nível. Para isto, necessita-se de esforço humano. Se forem necessárias muitas intervenções humanas, estas podem se mostrar de alto custo. Além disto, esta participação humana pode, por muitas vezes, inserir conhecimentos no sistema um tanto quanto inconsistentes.

Contudo, a principal razão para o sucesso crescente dos métodos estatísticos é que o conhecimento pode ser automaticamente inferido e pouco deverá ser manualmente inserido no modelo. O uso destes métodos requer a manipulação de

um grande número de dados estatísticos, mas obtém-se uma precisão bastante alta: por volta de 95%.

Todos estes aspectos devem ser cuidadosamente pensados, quando se realiza a escolha do método que será utilizado para a construção do sistema rotulador. Atualmente, segundo Church [CHU92], os sistemas que usam métodos estatísticos conseguem superar em muito os sistemas baseados em regras, que têm somente um sucesso limitado.

3.3.1 Rotulador de Categorias Morfo-Sintáticas Misto

Eric Brill em sua recente tese [BRI93b], descreve um sistema rotulador baseado em regras, mas que também realiza uma certa análise estatística, obtendo, deste modo, uma performance invejável. Isto significa que, dado um conjunto de regras adquiridas através de treinamento, este sistema rotulador irá realizar o “tagging” de um “corpus”, de acordo com estas regras.

O sistema é composto por dois rotuladores: o rotulador inicial e o rotulador final. O rotulador inicial faz uma compilação do “corpus” marcado. Esta compilação consiste da criação de um dicionário onde, para cada uma das palavras, há somente o seu rótulo mais comum. Assim, quando for executar o “tagging” de um novo “corpus”, o rotulador inicial atribuirá o rótulo mais provável para a palavra, conforme descrito no dicionário, figura 3.2. Esta tarefa é executada sem levar em conta o contexto no qual as palavras estão inseridas.

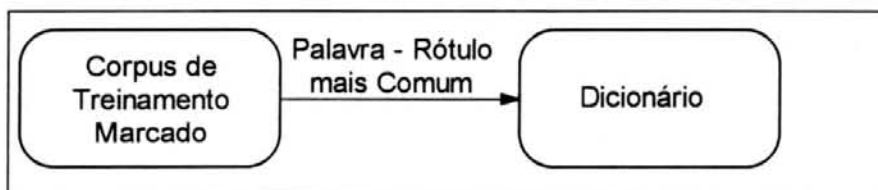


Figura 3.2 - Treinamento do Rotulador Inicial

Este sistema rotulador possui ainda dois procedimentos de alto nível que ajudam no tratamento de palavras desconhecidas (figura 3.3). O primeiro procedimento define que as palavras desconhecidas que iniciem por letra maiúscula tendem a ser substantivos próprios. O segundo procedimento especifica que as palavras desconhecidas receberão o rótulo mais comum para palavras que tenham aquela

terminação. Por exemplo, o rótulo mais comum para palavras terminando em “ar”, é verbo. Ambos os procedimentos são automaticamente inferidos a partir do “corpus” de treinamento marcado.

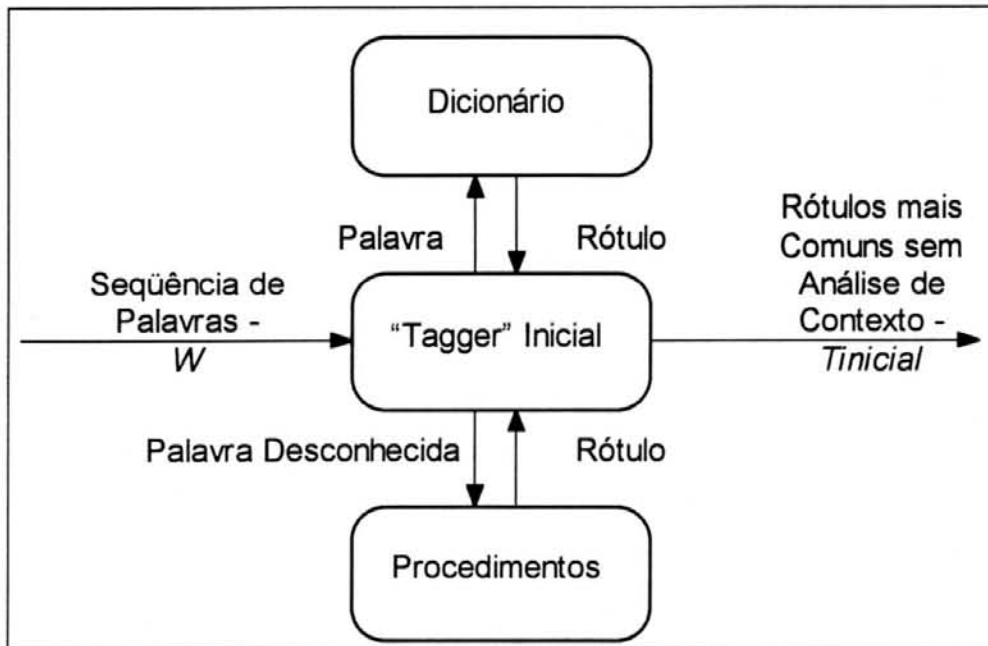


Figura 3.3 - Marcando com o Rotulador Inicial

Assim, “casa” nas seguintes sentenças receberia o rótulo N em ambos os casos:

Sentença 3.1: Aquela casa está a venda.

Sentença 3.2: Ele casa hoje.

O rotulador contextual inferirá automaticamente as regras de contexto, a partir do “corpus” de treinamento marcado. De que forma isto é feito? Faz-se o “tagging” do “corpus” de treinamento usando o rotulador inicial. A seguir, comparam-se automaticamente os resultados e gera-se uma lista de erros. A partir da aplicação desta lista de erros no “corpus”, serão geradas as regras de contexto. As aplicações que gerarem melhor resultado serão utilizadas como regras de contexto.

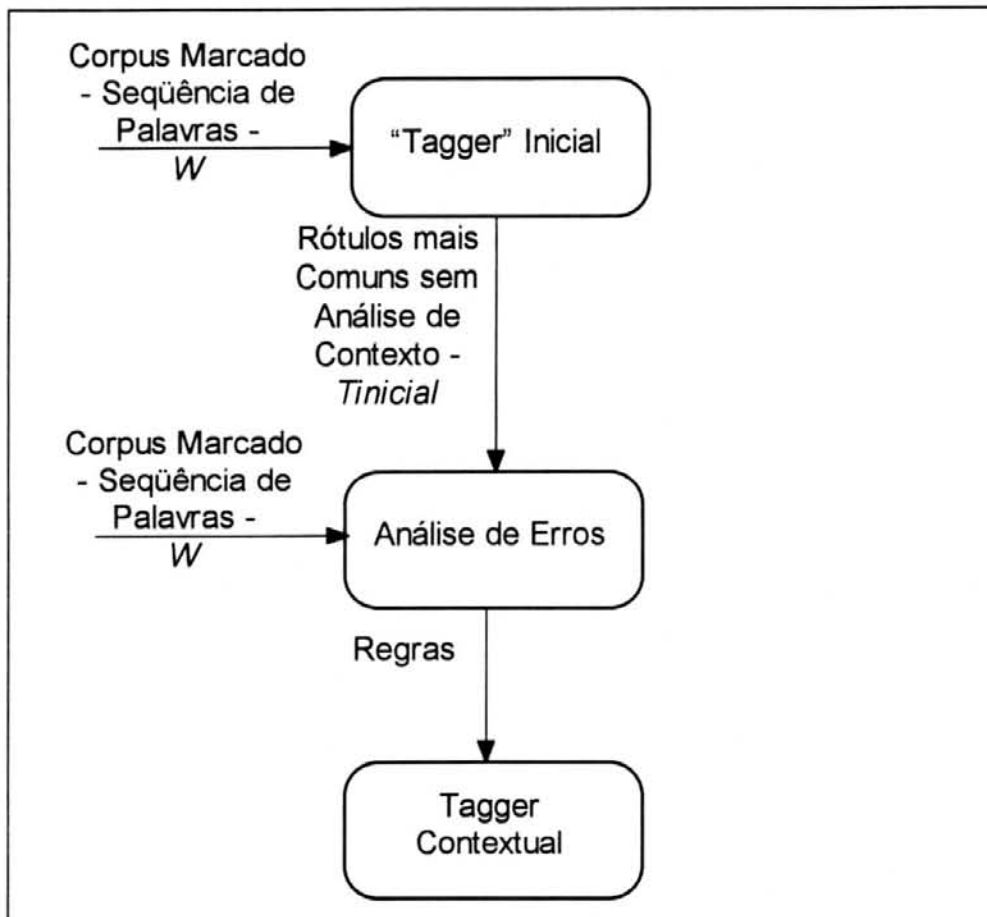


Figura 3.4 - Marcando com o Rotulador Inicial

Estas regras serão do tipo:

Tabela 3.1 - Padrões de Rótulos

| PADRÃO DE REGRA | SIGNIFICADO |
|----------------------|---|
| A B PREV TAG C | Troque de A para B se o rótulo anterior for C |
| A B PREV1OR2OR3TAG C | Troque de A para B se o primeiro rótulo anterior ou o segundo ou o terceiro for C |
| A B PREV1OR2TAG C | Troque de A para B se o primeiro rótulo anterior ou o segundo for C |
| A B NEXT1OR2TAG C | Troque de A para B se o primeiro ou o segundo rótulo posterior for C |
| A B NEXT TAG C | Troque de A para B se o rótulo posterior for C |
| A B SURROUND TAG C D | Troque de A para B se os rótulos vizinhos forem C e D |
| A B NEXTBIGRAM C D | Troque de A para B se o próximo bigrama for C D |
| A B PREVBIGRAM C D | Troque de A para B se o bigrama anterior for C D |

Assim, obtém-se um conjunto de regras que analisa a atribuição dos rótulos feita pelo rotulador inicial e os corrige conforme o contexto no qual as palavras aparecem.

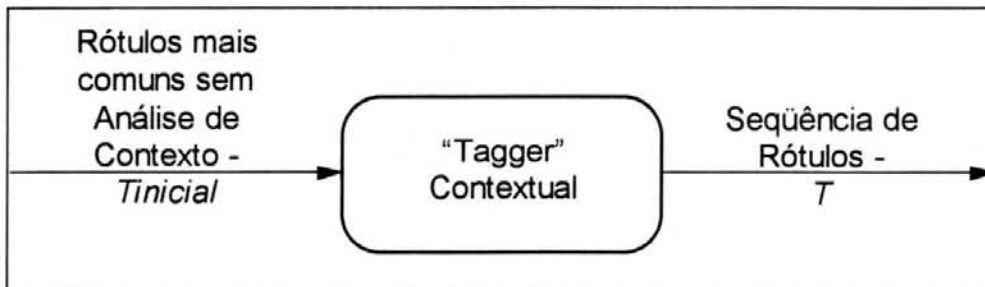


Figura 3.5 - Segunda Fase

Supondo-se que tal regra fosse gerada:

Regra1: N V prevtag PPR

na Sentença 3.2 (Ele casa hoje), o rótulo N para a palavra “casa” seria trocado pelo rótulo correto V, dado que “Ele” é um PPR (pronome pessoal do caso reto).

Este sistema rotulador consegue uma precisão de 94,9% para a Língua Inglesa, usando apenas 300 regras que foram automaticamente inferidas. Para atingir tal precisão, necessita de um “corpus” de treinamento bastante grande, com 1.000.000 de palavras [BRI92b]. Além disto, tem um processamento muito lento e o custo do treinamento é extremamente alto.

Como para a Língua Portuguesa ainda não se tem um “corpus” marcado de tal magnitude e, devido aos custos altos e ao tempo de processamento exigidos por este sistema rotulador, até o momento não foi possível usá-lo. Espera-se que futuramente se possa utilizá-lo. Devido a isto, neste trabalho serão abordados unicamente os sistemas rotuladores que utilizam métodos estatísticos.

3.4 Marcação Automática de Categorias Morfo-Sintáticas Usando a Teoria da Informação de Shannon

O processo de marcação automática de categorias morfo-sintáticas pode ser pensado em termos do modelo de Shannon. Mas como? De que forma um processo de “tagging” pode ser mapeado para o modelo de Shannon? Como foi visto na seção 2.4, no modelo de Shannon há uma seqüência de entrada no canal, que devido ao ruído ali existente, fica corrompida e é produzida como saída uma outra seqüência. Então, dada uma saída, tenta-se reconstituir a seqüência de entrada correspondente.

Pensando agora em termos de “tagging”: imagine-se que a entrada do canal de ruídos é uma seqüência de rótulos T , que fica corrompida, gerando como saída, uma seqüência de palavras W .



Figura 3.6 - Marcação através do modelo de Shannon

Então, tendo como saída do canal de ruídos uma seqüência de palavras e usando os recursos fornecidos pela Teoria da Informação, tenta-se reconstituir a seqüência de rótulos. Esta reconstituição é feita através do uso do sistema rotulador. Dada uma seqüência de palavras, o sistema rotulador tentará reconstituir a seqüência de rótulos correspondente.

Desta forma, a questão que se tenta responder é: “Que seqüência de rótulos T corresponde à seqüência de palavras W ?”. E, usando a fórmula 2.1, tem-se que:

$$\max P(T|W) = \max_T P(T)P(W|T)$$

Fórmula 3.1

onde $P(T|W)$ é a seqüência de rótulos mais provável para a seqüência de palavras W . $P(W|T)$ corresponde ao **modelo do canal** e é descrito em [CHU92] como sendo uma tabela que define para cada par (ou alinhamento) W, T da linguagem, um valor correspondente a sua probabilidade. Assim, para tentar descobrir qual é a seqüência de rótulos correta, para cada conjunto de palavras de uma língua e os seus rótulos, que tenham o mesmo tamanho da seqüência de entrada, será atribuída uma probabilidade. Esta probabilidade corresponde ao fato de que, esta seqüência de palavras, escolhida ao acaso, entre todas as possíveis da língua, e à qual está associado um conjunto de rótulos T , vir a ser a seqüência de palavras W . Uma vez que isto é um evento único, torna-se impossível de calcular tal probabilidade. Então, costuma-se usar uma aproximação mais simples para ela: o modelo de n-gramas. Apesar desta aproximação ser mais simples, costuma apresentar resultados surpreendentemente altos.

O conceito de bigramas, como foi explicado na seção 2.5, determina que a probabilidade de um rótulo depende apenas da probabilidade do rótulo anterior $P(t_i|t_{i-1})$. E a probabilidade de uma seqüência de rótulos $P(t_1, t_2, \dots, t_n)$ é dada pelo produtório da probabilidade de cada um dos rótulos $\prod_{i=1}^n P(t_i|t_{i-1})$. Usando-se o conceito de

bigramas para descrever o **modelo da linguagem** $P(T)$, também conhecido como **modelo original**, deriva-se:

$$P(T) = P(t_1, t_2, \dots, t_n) \approx \prod_{i=1}^n P(t_i | t_{i-1})$$

Fórmula 3.2

E aplicando-se o conceito de trigramas, se obtém:

$$P(T) = P(t_1, t_2, \dots, t_n) \approx \prod_{i=1}^n P(t_i | t_{i-2}, t_{i-1})$$

Fórmula 3.3

que determina que um dado rótulo é dependente somente dos dois rótulos anteriores $P(t_i | t_{i-2}, t_{i-1})$. E, da mesma forma que em bigramas, a probabilidade de uma seqüência de rótulos $P(t_1, t_2, \dots, t_n)$ é dada pelo produtório da probabilidade de cada um dos rótulos $\prod_{i=1}^n P(t_i | t_{i-2}, t_{i-1})$.

Já $P(W|T)$, que é o **modelo do canal**, é determinada por:

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

Fórmula 3.4

onde a probabilidade de uma palavra é dependente somente de seu rótulo $P(w_i | t_i)$. Conseqüentemente, a probabilidade de uma seqüência de palavras, dada uma seqüência de rótulos $P(W|T)$, é obtida pelo produtório da probabilidade de uma palavra, dado o seu rótulo $\prod_{i=1}^n P(w_i | t_i)$.

Como já foi dito, os modelos de n-gramas mais comuns são os bigramas e os trigramas. E estes modelos, apesar de extremamente simples, apresentam ótimos resultados. Pode-se, então, pensar: "Se analisando uma vizinhança próxima ($n=2$, a palavra anterior ou $n=3$ as duas palavras anteriores), obtém-se um resultado tão bom, por quê não usar um "n" maior e obter uma precisão melhor?". A resposta é simples: seqüências mais longas se constituem de eventos mais raros, possivelmente únicos. Conseqüentemente, a freqüência com que elas ocorrem é bastante baixa. Neste caso,

não se podem usar métodos empíricos para estimar probabilidades, a partir das frequências com que estas seqüências aparecem no “corpus”. Shabes [SCA92] apresenta dados extraídos do “Wall Street Journal Corpus”, que tem 800.000 palavras marcadas com rótulo, onde se pode ter uma idéia mais clara deste problema:

Tabela 3.2 : Bigramas X Trigramas

| WSJ Corpus | Bigramas | Trigramas |
|------------------------|----------|-----------|
| Rótulos | 48 | 48 |
| Combinações Possíveis | 2.304 | 110.592 |
| Combinações Observadas | 1.366 | 14.306 |

Como pode ser visto na tabela acima, dos 2.304 bigramas possíveis, apenas 1.366 foram encontrados no “corpus” de 800.000 palavras, ou seja, 59,29% dos bigramas possíveis. Isto significa que mais da metade dos bigramas apareceram no “corpus”. No mesmo “corpus”, dos 110.592 trigramas possíveis, apenas 14.306 ocorreram, ou seja, 12,94%.

Os bigramas e trigramas, apesar de serem extremamente simples, quando são bem construídos, costumam resultar em modelos com precisão bastante alta. São nestas aproximações, usando a Teoria de Shannon, que está baseado o sistema rotulador desenvolvido, descrito no capítulo 4.

3.5 HMM para Marcação Automática de Categorias Morfo-Sintáticas

A questão agora é como implementar o rotulador. Um formalismo que se adapta perfeitamente às necessidades que se apresentam é um caso particular dos Modelos de Markov: o “Hidden Markov Model” ou HMM.

A principal vantagem que o HMM apresenta é a possibilidade que se tem de fazer treinamento automático do modelo. Além disto, permite que se tenha flexibilidade na escolha de “corpus” de treinamento. Textos extraídos de qualquer

domínio podem ser usados para fazer o treinamento do modelo. Desta forma, pode-se escolher um determinado domínio de aplicação no qual se quer trabalhar com o sistema rotulador e usar textos deste domínio para treiná-lo. Caso se necessite ampliar o conjunto de rótulos, ele apresenta facilidade na introdução de novos rótulos ao modelo. Este formalismo, que foi explicado no capítulo 2, é utilizado em grande parte dos sistemas rotuladores estocásticos citados na literatura [CUT92, KEM94a].

Um HMM pode ser visto como tendo um modelo de linguagem markoviano, isto é, que tem estados finitos, e um modelo de canal, que depende somente do estado em que está o modelo da linguagem.

Mas como pensar em um HMM em termos de marcação automática ? Basta que se considere que os estados do HMM representem os rótulos e que as transições sejam as palavras da língua. Deste modo, cada estado do HMM corresponde à palavra que será produzida a seguir. Existem dois tipos de probabilidades:

- as probabilidades contextuais, ou de transição, que especificam que um determinado estado p seja seguido por um estado q ;
- as probabilidades lexicais, que especificam a probabilidade de um símbolo x ser emitido quando se está em um estado p .

A figura 3.7, a seguir, apresenta um HMM onde pode-se visualizar mais claramente as probabilidades lexicais e as probabilidades contextuais, que são apresentadas em separado:

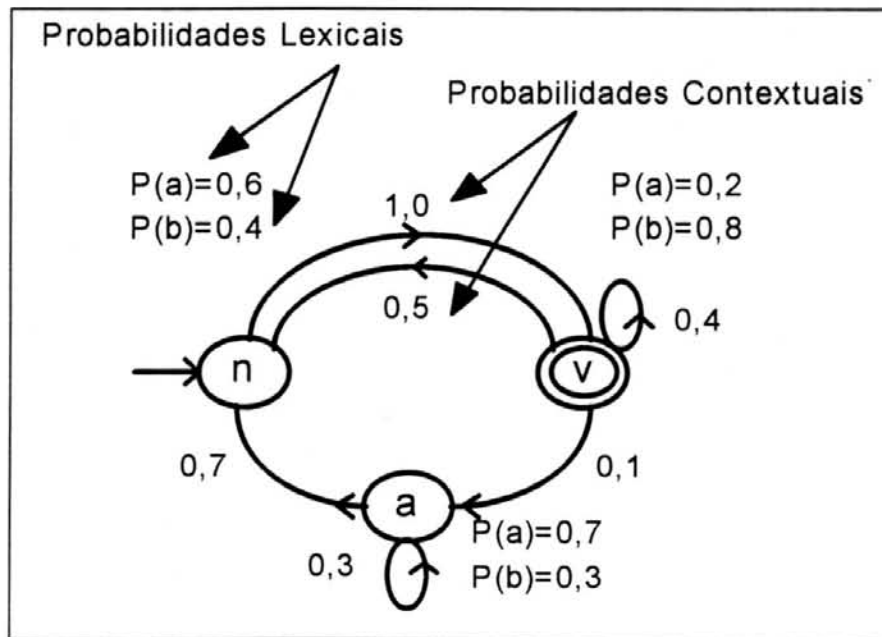


Figura 3.7 - HMM Com Três Estados

O HMM da figura 3.8⁴ é equivalente ao HMM da figura 3.7, só que apresenta as probabilidades lexicais e contextuais já integradas :

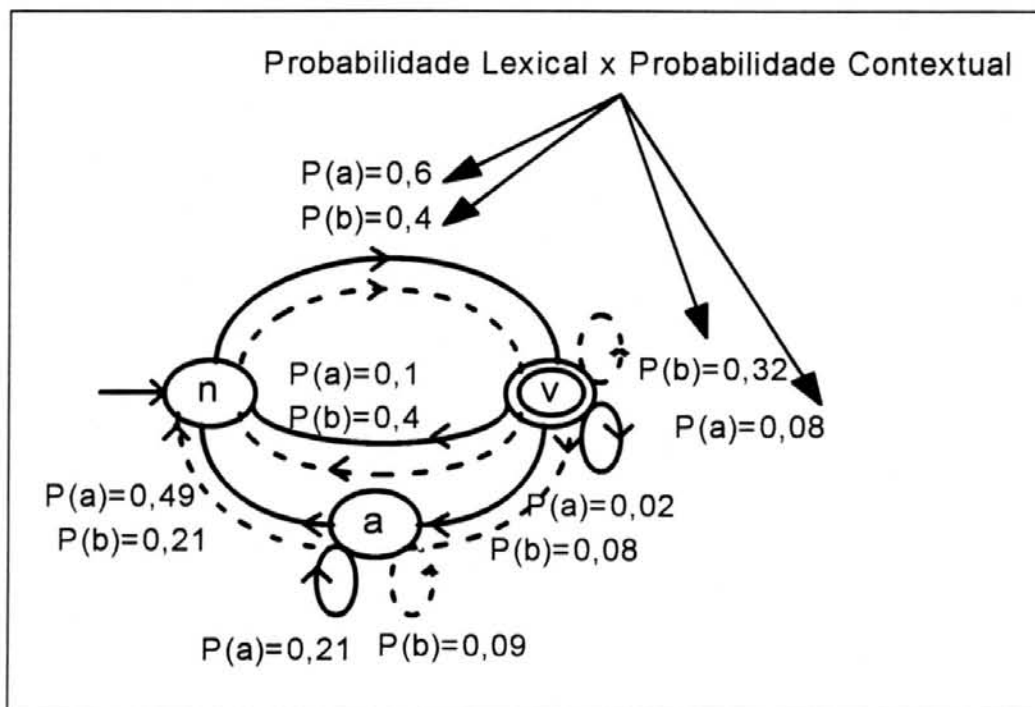


Figura 3.8 - HMM Com Três Estados (outra visualização)

⁴As transições com o símbolo B estão representadas pela linha pontilhada, para facilitar a visualização da Figura 3.4.

O conceito de n-gramas, mais especificamente o de bigramas, define que um rótulo só é dependente do rótulo imediatamente anterior a ele. Pode-se implementar o conceito de bigramas em HMMs quando se define que a transição para um estado depende apenas da transição anterior. Esta é a probabilidade contextual do modelo, que é definida como:

$$P(t_n|w_{1,n-1},t_{1,n-1})=P(t_n|t_{n-1})$$

Fórmula 3.5

que significa que, dada uma seqüência de palavras $w_{1,n-1}$ e uma seqüência de rótulos $t_{1,n-1}$, a probabilidade do próximo rótulo t_n depende somente do rótulo anterior t_{n-1} .

A probabilidade lexical de um bigrama especifica que a probabilidade de produção de uma palavra depende somente da probabilidade do rótulo dela. Em um HMM, isto significa que a emissão de um símbolo depende somente do estado onde se está no modelo. Desta forma, a probabilidade lexical é definida como:

$$P(w_n|w_{1,n-1},t_{1,n})=P(w_n|t_n)$$

Fórmula 3.6

onde se define que a probabilidade de uma palavra w_n , dada uma seqüência de palavras $w_{1,n-1}$ e uma seqüência de rótulos $t_{1,n}$, é dependente somente da probabilidade do seu rótulo t_n .

Conseqüentemente o modelo da linguagem é:

$$P(w_{1,n})= \sum_{t_{1,n+1}} \prod_{i=1}^n P(w_i|t_i)P(t_{i+1}|t_i)$$

Fórmula 3.7

Na figura 3.9 pode ser visto um HMM no qual foi implementado o conceito de bigrama, ou como é denominado, um HMM de primeira ordem ($n=1$). Este HMM tem dois estados, “n” e “v”. Considerando-se, por exemplo, a transição de estado “n” para o estado “v”, com o símbolo “a”, pode-se ver que a emissão de tal símbolo está condicionada a que o estado atual seja “n” $P(a|n)$. Já a probabilidade de o

próximo estado ser “v” está condicionado a ocorrência de “n” como o estado atual $P(v|n)$.

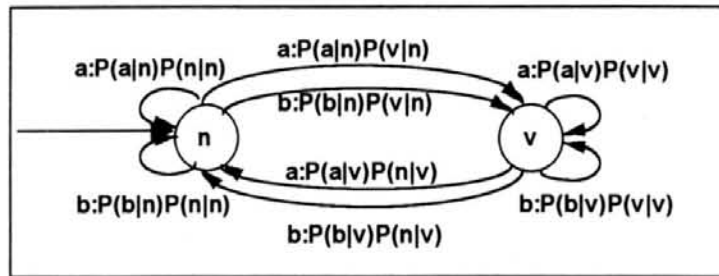


Figura 3.9 - HMM de Primeira Ordem Para Marcação Morfo-Sintática

E para o caso de se trabalhar com trigramas, no HMM, ter-se-á que um estado dependerá dos dois estados imediatamente anteriores a ele. Tem-se então, um HMM de segunda ordem, que implementa o trigrama ($n=2$). Esta é a probabilidade contextual, que é formalizada através da fórmula 3.8:

$$P(t_n | w_{1,n-1}, t_{1,n-1}) = P(t_n | t_{n-2} t_{n-1})$$

Fórmula 3.8

E o modelo da linguagem é definida por:

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

Fórmula 3.9

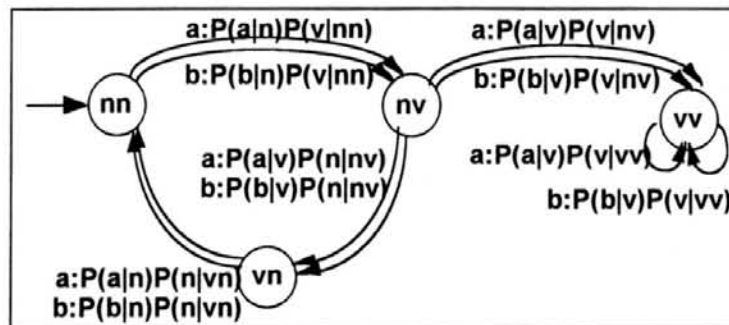


Figura 3.10 - HMM de Segunda Ordem para Marcação Morfo-Sintática

Na figura acima tem-se que, por exemplo, na transição de “nn” para “nv” com o símbolo “a”, a probabilidade lexical $P(a|n)$, da emissão de “a” depende do rótulo atual ser “n” e a probabilidade contextual $P(v|nn)$ é de o próximo rótulo vir a ser “v”, uma vez que o anterior é “n” e o atual é “n” também.

3.6 Algoritmos de Treinamento e de Marcação

Agora que se construiu o sistema rotulador, como um HMM, restam algumas questões:

- **o que é, exatamente, e como se faz o treinamento do HMM?**
- **como fazer “tagging”, usando o HMM?**

Após ter-se construído um “Hidden Markov Model”, tem-se definidos os estados e as transições entre eles. Contudo, não se sabe ainda quais são os valores dos parâmetros (probabilidades) deste HMM. Então, o próximo passo é fazer a estimativa dos valores destes parâmetros. Este processo também é conhecido como “**treinamento**”.

Este processo de treinamento pode ser visto como aquisição do conhecimento, porque é através dele que se vai incorporar no sistema rotulador o conhecimento desejado. Inicialmente ele não tem conhecimento nenhum, só tem estados e transições. Assim, para possibilitar este “aprendizado” é fornecido um “corpus” de treinamento e o sistema rotulador irá extrair dele o conhecimento necessário, figura 3.11.

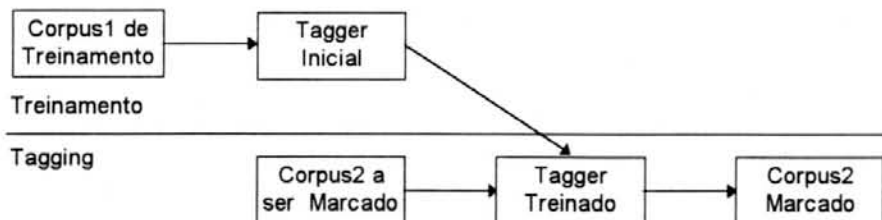


Figura 3.11 - Ciclo de Aquisição e Utilização do Conhecimento de um Rotulador

Esta extração do conhecimento se dará através da estimativa dos parâmetros do HMM. Desta forma, ele analisa e aprende os padrões linguísticos presentes no “corpus” de treinamento e irá modelar isto através das relações existentes entre os estados do HMM.

Para fazer o cálculo dos valores dos parâmetros do HMM, são usados algoritmos como o de Frequência Relativa - FR- e o de “Forward-Backward” - FB, explicados nas seções 3.6.1 e 3.6.2 respectivamente.

O treinamento é feito estimando os parâmetros do HMM. Após fazer esta estimativa, já se sabe exatamente os valores das probabilidades lexicais e das probabilidades contextuais; neste momento o sistema rotulador já assimilou o conhecimento. Assim, com base nestes dados, pode-se usar o sistema rotulador para fazer o “tagging”, ou marcação de textos. Um algoritmo muito utilizado para fazer o “tagging” é o algoritmo de Viterbi [CHK93, CUT92, MER94]. Ele é o mais utilizado nos trabalhos publicados na área, e realiza análise a nível de sentença.

Após ter implementado os algoritmos de treinamento e de “tagging”, usa-se o sistema rotulador para fazer a marcação de um texto qualquer e analisa-se o resultado obtido. Pode-se então, com base nos resultados obtidos, retreinar o “tagger, ajustando os seus parâmetros”, caso estes resultados não tenham sido satisfatórios. Esta é a fase de ajustes. Após realizados estes ajustes necessários, pode-se considerar que o rotulador está finalmente pronto para ser usado para fazer a marcação de textos. A seguir serão apresentados os algoritmos usados para fazer o treinamento e o “tagging” usando HMMs.

3.6.1 Frequência Relativa(FR)

Dado um “corpus” marcado, este algoritmo faz a estimativa dos parâmetros de um HMM, baseando-se nas frequências dos padrões que ocorrem neste “corpus”. Calcula-se a frequência $F(w_i, t_i)$ com que uma palavra w_i ocorre com o rótulo t_i (frequência lexical) e o número de vezes $F(t_{i-2}, t_{i-1}, t_i)$ em que um rótulo t_i é precedido dos tags t_{i-2} e t_{i-1} (frequência contextual).

O cálculo das probabilidades lexicais é um tanto intuitivo. Determinam-se todas as ocorrências de uma palavra com um determinado rótulo e se divide este valor pelo total de vezes em que este rótulo ocorre no “corpus”. Este procedimento é executado para cada um dos possíveis rótulos para a palavra. Por exemplo, dada a seguinte tabela:

Tabela 3.3 - Probabilidades Lexicais

| Palavra | Rótulo | Ocorrências da Palavra com o Rótulo | Ocorrências do Rótulo |
|----------|--------|-------------------------------------|-----------------------|
| o | ART | 835 | 1785 |
| caso | N | 57 | 2432 |
| terminou | VI | 95 | 1020 |
| . | PTO | 673 | 673 |

e a fórmula do cálculo das probabilidades lexicais:

$$P(w_i|t_i) = \frac{F(w_i, t_i)}{F(t_i)}$$

Fórmula 3.10

de onde se pode inferir que a probabilidade de um artigo (ART) ser a palavra “o” é de 835/1785 ou 0,47 ou ainda 47% das vezes em que um artigo ocorre. Um substantivo (N) terá uma probabilidade de 2,3% ($57/2432 = 0,023$) de ser a palavra “caso”. Já um verbo intransitivo (VI) será a palavra “terminou” em 9,3% ($95/1020 = 0,093$) das vezes. Enquanto que PTO será ‘.’ em 100% das vezes ($673/673$).

O cálculo das probabilidades contextuais se baseia na frequência de vezes que um dado n-grama $F(t_{i-2}, t_{i-1}, t_i)$ ocorre no “corpus”. Como será visto no capítulo 4, o sistema rotulador desenvolvido usa o cálculo feito para bigramas $F(t_{i-1}, t_i)$, onde se determina a probabilidade de um certo rótulo ser precedido por outro. Assim, tem-se a seguinte fórmula:

$$P(t_i|t_{i-1}) = \frac{F(t_{i-1}, t_i)}{F(t_i)}$$

Fórmula 3.11

E, no caso de se desejar descobrir a probabilidade de um artigo ser seguido por um substantivo, se aplica a fórmula 3.11:

$$P(N|ART) = \frac{F(ART, N)}{(N)}$$

que, segundo os dados coletados do “corpus”, resulta em:

$$P(N|ART) = 1576/3798 = 0,41.$$

Este método tem se apresentado bastante satisfatório, sendo fácil de implementar e produzindo bons resultados para modelos simples. Uma desvantagem deste algoritmo é que ele atribui uma probabilidade zero para seqüências de rótulos que não ocorreram no “corpus” de treinamento. Isto gerará problemas para alinhamentos que tenham tais seqüências, aos quais será atribuída uma probabilidade de zero. Logo, para este tipo de sentenças, o algoritmo é ineficiente. Este problema, da esparcidade dos dados, será discutido mais detalhadamente na seção 3.6.4.

3.6.2 Algoritmo de Forward-Backward

O algoritmo de “Forward-Backward” é um outro algoritmo usado para estimar os parâmetros do HMM. Para tanto, é calculada a probabilidade que um HMM atribui a um dado “corpus” de treinamento. A partir deste resultado, tenta-se ajustar as probabilidades das transições do HMM, de forma a maximizar a probabilidade atribuída ao “corpus” de treinamento.

Este algoritmo calcula, recursivamente, dois conjuntos de probabilidades: a “*forward probability*” e a “*backward probability*”.

3.6.2.1 Forward e Backward Probabilities

A “*forward probability*”, $\alpha_i(t)$ é a probabilidade conjunta de no tempo t o HMM estar no estado s^i , dada uma seqüência de símbolos $\{w_1, w_2, \dots, w_{t-1}\}$. Faz isto de maneira incremental, calculando uma palavra por ciclo de execução, começando pela primeira. Pode-se dizer que esta probabilidade é calculada através da movimentação progressiva na seqüência. Assim:

$$\alpha_i(t) \stackrel{\text{def}}{=} P(w_{1,t-1}, S_t = s^i), t > 1$$

Fórmula 3.12

Em $t=1$ tem-se a entrada no HMM que pode ser descrita por:

$$\alpha_i(1) \stackrel{\text{def}}{=} \begin{cases} 1, & \text{se } i = 1 \\ \text{senão} & 0 \end{cases}$$

Fórmula 3.13

Através do cálculo de todos os $\alpha_i(n+1)$ para a seqüência de símbolos $w_{1,n}$, obtém-se facilmente a probabilidade da seqüência $P(w_{1,n})$. Isto pode ser expresso em:

$$P(w_{1,n}) = \sum_{i=1}^{\sigma} P(w_{1,n}, S_{n+1} = s^i) = \sum_{i=1}^{\sigma} \alpha_i(n+1)$$

Fórmula 3.14

Pela fórmula seguinte, observa-se que, tendo-se calculado todos os $\alpha_j(t)$, pode-se facilmente calcular o valor de $\alpha_j(t+1)$:

$$\alpha_j(t+1) = \sum_{i=1}^{\sigma} \alpha_i(t) P(s^i \xrightarrow{w_t} s^j)$$

Fórmula 3.15

A execução do cálculo da *forward probability* pode ser melhor entendida através do exemplo a seguir. Para este exemplo, vai ser usado o HMM da figura 2.3, apresentado novamente na figura 3.12, para uma visualização melhor:

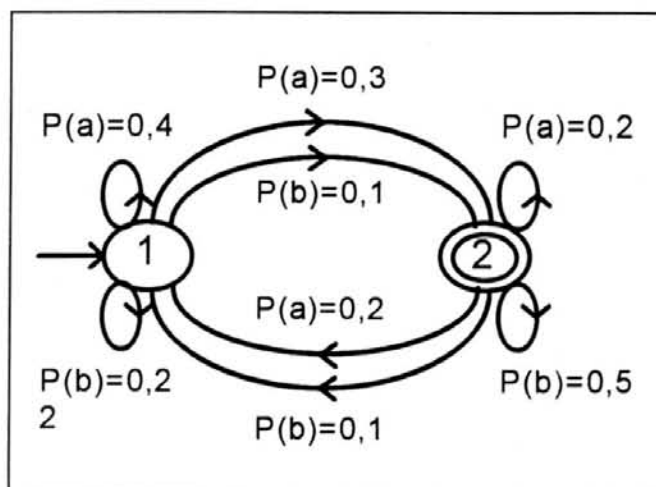


Figura 3.12 - HMM Com Dois Estados

Os passos do cálculo da “forward probability” da seqüência “aab”, dado o HMM da figura 3.12, podem ser vistos em:

$$\alpha_1(1)=1,0$$

$$\alpha_2(1)=0,0$$

$$\alpha_1(2)=1,0 * 0,4 + 0,0 * 0,2 = 0,4$$

$$\alpha_2(2)=1,0 * 0,3 + 0,0 * 0,2 = 0,3$$

$$\alpha_1(3)=0,4 * 0,4 + 0,3 * 0,2 = 0,22$$

$$\alpha_2(3)=0,4 * 0,3 + 0,3 * 0,2 = 0,18$$

$$\alpha_1(4)=0,22 * 0,2 + 0,18 * 0,1 = 0,062$$

$$\alpha_2(4)=0,22 * 0,1 + 0,18 * 0,5 = 0,112$$

Estes valores estão descritos na seguinte tabela:

Tabela 3.4 - Forward Probabilities

| Linha do Tempo | 1 | 2 | 3 | 4 |
|----------------|------------|-----|------|-------|
| Símbolos | ϵ | a | aa | aab |
| $\alpha_1(t)$ | 1,0 | 0,4 | 0,22 | 0,062 |
| $\alpha_2(t)$ | 0,0 | 0,3 | 0,18 | 0,112 |
| $P(w_{1,t})$ | 1,0 | 0,7 | 0,40 | 0,174 |

Por outro lado, a “backward probability” $\beta_i(t)$ pode ser definida como a probabilidade de, estando-se no estado s^i , no tempo t , se ter a seqüência $\{w_t, \dots, w_n\}$. A “backward probability” é determinada exatamente do mesmo modo que a “forward

probability”, porém é iniciada pela última palavra e vai sendo executada em direção ao começo da seqüência.

$$\beta_i(t) \stackrel{\text{def}}{=} P(w_{t,n} | S_t = s^i)$$

Fórmula 3.16

Observa-se aqui que a “backward probability” pode ser usada para o cálculo da probabilidade da seqüência $P(w_{1,n})$, de maneira similar à “forward probability”:

$$\beta_1(1) = P(w_{1,n} | S_1 = s^1) = P(w_{1,n})$$

Fórmula 3.17

O cálculo é feito de maneira recursiva:

$$\beta_i(t-1) = \sum_{j=1}^{\sigma} P(s^i \xrightarrow{w_{t-1}} s^j) \beta_j(t)$$

Fórmula 3.18

Começando pelo final da seqüência em direção ao começo, tem-se que o ciclo base (o primeiro) deve ser igual a 1:

$$\beta_i(n+1) = P(\epsilon | S_{n+1} = s^i) =$$

Fórmula 3.19

Assim, para a seqüência “aab”, tem-se os seguintes valores:

$$\beta_1(4) = 1,0$$

$$\beta_2(4) = 1,0$$

$$\beta_1(3) = 1,0 * 0,2 + 1,0 * 0,1 = 0,3$$

$$\beta_2(3) = 1,0 * 0,5 + 1,0 * 0,1 = 0,6$$

$$\beta_1(2) = 0,3 * 0,4 + 0,6 * 0,3 = 0,3$$

$$\beta_2(2) = 0,6 * 0,2 + 0,3 * 0,2 = 0,18$$

$$\beta_1(2)=0,3 * 0,4 + 0,18 * 0,3 = 0,174$$

$$\beta_2(1)=0,18 * 0,2 + 0,3 * 0,2 = 0,096$$

e a seguinte tabela:

Tabela 3.5 - Backward Probabilities

| | | | | |
|----------------|------------|-----|------|-------|
| Linha do Tempo | 4 | 3 | 2 | 1 |
| Símbolos | ϵ | b | ab | aab |
| $\beta_1(t)$ | 1,0 | 0,3 | 0,3 | 0,174 |
| $\beta_2(t)$ | 1,0 | 0,6 | 0,18 | 0,096 |

Como pode ser visto na tabela 3.5, o $\beta_1(1)=0,174$ é igual ao $P(w_{1,4})$ da tabela 3.4, obedecendo ao que a Fórmula 3.15 estabelece. O valor de $\beta_2(1)$ pode ser desconsiderado, uma vez que não tem nenhuma importância para o cálculo da probabilidade da seqüência.

3.6.2.2 Funcionamento do Algoritmo

O algoritmo de “Forward-Backward” é usado para fazer o treinamento de HMMs. Ele utiliza a “forward probability” e a “backward probability”, para ajustar os parâmetros probabilísticos do HMM, de modo que seja atribuída a maior probabilidade possível à seqüência de treinamento analisada.

Isto auxiliará muito na análise de outras seqüências de símbolos, pois com os parâmetros ajustados, o HMM saberá qual o melhor caminho a seguir para reconhecer uma seqüência similar.

A idéia básica é contar quantas vezes os caminhos são utilizados no reconhecimento de seqüências de treinamento, e dar um peso maior aos que são utilizados mais vezes. Como não se pode ter certeza de quais os caminhos que foram percorridos para reconhecer a dada seqüência, visto que o HMM é não determinístico, usa-se um artifício: supõe-se que todas as transições possíveis para aquela seqüência

foram seguidas, fórmula 3.20. A seguir estes valores são ajustados de acordo com a probabilidade do caminho no qual a transição se encontra, e todos os valores são somados, fórmula 3.21.

$$P_c(s^i \xrightarrow{w^k} s^j) = \frac{C(s^i \xrightarrow{w^k} s^j)}{\sum_{l=1, m=1}^{\sigma, \omega} C(s^i \xrightarrow{w^m} s^l)}$$

Fórmula 3.20

$$C(s^i \xrightarrow{w^k} s^j) = \sum_{s_{1,n+1}} P(s_{1,n+1} | w_{1,n}) \eta(s^i \xrightarrow{w^k} s^j, s_{1,n}, w_{1,n})$$

Fórmula 3.21

Onde $\eta(s^i \xrightarrow{w^k} s^j, s_{1,n}, w_{1,n})$ se refere à contagem das vezes que a transição $s^i \xrightarrow{w^k} s^j$ aparece na seqüência de estados $s_{1,n}$ utilizados para reconhecer a seqüência de símbolos $w_{1,n}$.

Então, a cada ciclo de execução do algoritmo, os valores dos parâmetros são ajustados, gerando assim novos valores. Estes novos valores, por sua vez, serão utilizados como entrada para o próximo ciclo do algoritmo, ou seja, serão ajustados e produzirão outros novos valores. Este processo se repete até que as probabilidades geradas para as transições se estabilizem, ou até que estabilize a probabilidade atribuída ao “corpus”. A segunda opção é considerada a mais acessível, uma vez que é mais fácil fazer a comparação de dois números (a probabilidade antiga e a nova do “corpus”) do que comparar dois conjuntos de números (probabilidades antigas e novas das transições).

A seguir é apresentada a execução do algoritmo de forward-backward para o HMM da figura 3.13. São mostradas 3 iterações do algoritmo, para a seqüência “aabab”. Os caminhos que podem gerar esta seqüência são “111111”, “111112”, “111211”, “111212”.

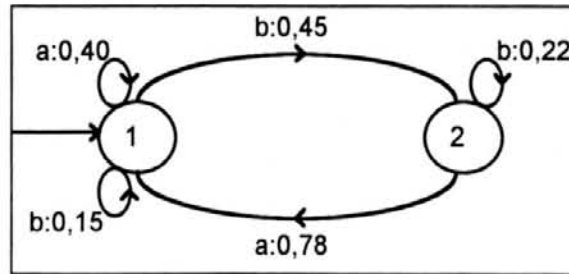


Figura 3.13 - HMM Inicial

As probabilidades dos caminhos são:

$$111111 = 0,40 \times 0,40 \times 0,15 \times 0,40 \times 0,15 = 0,00144$$

$$111112 = 0,40 \times 0,40 \times 0,15 \times 0,40 \times 0,45 = 0,00432$$

$$111211 = 0,40 \times 0,40 \times 0,45 \times 0,78 \times 0,15 = 0,00842$$

$$111212 = 0,40 \times 0,40 \times 0,45 \times 0,78 \times 0,45 = 0,02527$$

Aplicando-se a fórmula 3.20:

$$P(1 \xrightarrow{a} 1) = (0,00144 \times 3) + (0,00432 \times 3) + (0,00842 \times 2) + (0,02527 \times 2) = 0,00432 + 0,01296 + 0,01684 + 0,05054 = 0,08466 \cong 0,085$$

obtem-se os valores relativos ao número de vezes que a transição foi percorrida para cada um dos possíveis caminhos (valores da terceira coluna da tabela 3.6). E o valor final, 0,08466, é a soma de todos estes valores (6ª linha 3ª coluna da tabela 3.6). Este valor é arredondado e gera o valor a ser utilizado nos cálculos, 0,85 (7ª linha 3ª coluna da tabela 3.6).

$$P(1 \xrightarrow{a} 1) = \frac{0,085}{0,085 + 0,016 + 0,063} = \frac{0,085}{0,164} = 0,52$$

a seguir este valor é dividido pela soma de todos os valores das transições que partem deste mesmo estado, gerando o novo valor da transição, 0,52 (8ª linha 3ª coluna da tabela 3.6).

para a transição de 1 para 1 com o símbolo 'a'. Os valores para as demais transições podem ser vistos na tabela 3.6 e figura 3.14.

Tabela 3.6 - Primeira Iteração

| Caminho | P(Cam) | $1 \xrightarrow{a} 1$ | $1 \xrightarrow{b} 1$ | $1 \xrightarrow{b} 2$ | $2 \xrightarrow{b} 2$ | $2 \xrightarrow{a} 1$ | $\Sigma 1 \xrightarrow{w} s$ | $\Sigma 2 \xrightarrow{w} s$ |
|-------------------------|---------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|------------------------------|------------------------------|
| 111111 | 0,00144 | 0,00432 | 0,00288 | 0,0 | 0,0 | 0,0 | | |
| 111112 | 0,00432 | 0,01296 | 0,00432 | 0,00432 | 0,0 | 0,0 | | |
| 111211 | 0,00842 | 0,01684 | 0,00842 | 0,00842 | 0,0 | 0,00842 | | |
| 111212 | 0,02527 | 0,05054 | 0,0 | 0,05054 | 0,0 | 0,02527 | | |
| Total | 0,01333 | 0,08466 | 0,01562 | 0,06328 | 0,0 | 0,03369 | | |
| Valor ser usado | 0,01 | 0,085 | 0,016 | 0,063 | 0,00 | 0,034 | 0,164 | 0,034 |
| Novo P(\rightarrow) | | 0,52 | 0,1 | 0,38 | 0,0 | 1,0 | | |

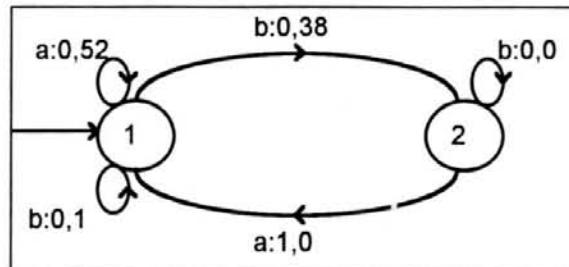


Figura 3.14 - HMM Após a Primeira Iteração

Pode-se observar que o valor das probabilidades das transições mais utilizadas aumentou, e as transições menos utilizadas tiveram suas probabilidades reduzidas. Este ajuste se dará, novamente, a cada iteração do algoritmo. Assim, espera-se que, após a última iteração, as probabilidades estejam ajustadas conforme a maior ou menor utilização das transições.

Tabela 3.7 - Segunda Iteração

| Caminho | P(Cam) | 1 \xrightarrow{a} 1 | 1 \xrightarrow{b} 1 | 1 \xrightarrow{b} 2 | 2 \xrightarrow{b} 2 | 2 \xrightarrow{a} 1 | $\Sigma 1 \xrightarrow{w} s$ | $\Sigma 2 \xrightarrow{w} s$ |
|-------------------------|---------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|------------------------------|------------------------------|
| 111111 | 0,00141 | 0,00423 | 0,00282 | 0,0 | 0,0 | 0,0 | | |
| 111112 | 0,00534 | 0,01602 | 0,00534 | 0,00534 | 0,0 | 0,0 | | |
| 111211 | 0,01028 | 0,02056 | 0,01028 | 0,01028 | 0,0 | 0,01028 | | |
| 111212 | 0,03905 | 0,07810 | 0,0 | 0,07810 | 0,0 | 0,03905 | | |
| Total | 0,05608 | 0,11891 | 0,01844 | 0,09372 | 0,0 | 0,04933 | | |
| Valor ser usado | 0,06 | 0,12 | 0,02 | 0,09 | 0,0 | 0,05 | 0,230 | 0,05 |
| Nova P(\rightarrow) | | 0,52 | 0,09 | 0,39 | 0,0 | 1,0 | | |

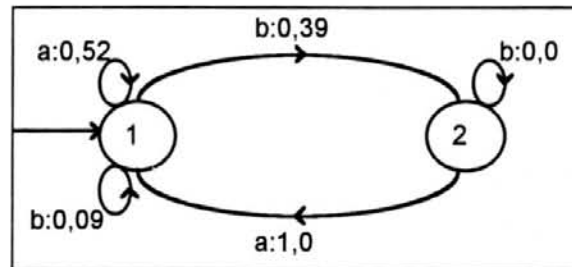


Figura 3.15 - HMM Após a Segunda Iteração

Tabela 3.8 - Terceira Iteração

| Caminho | P(Cam) | $1 \xrightarrow{a} 1$ | $1 \xrightarrow{b} 1$ | $1 \xrightarrow{b} 2$ | $2 \xrightarrow{b} 2$ | $2 \xrightarrow{a}$ | $\Sigma 1 \xrightarrow{w} s$ | $\Sigma 2 \xrightarrow{w}$ |
|-------------------------|---------|-----------------------|-----------------------|-----------------------|-----------------------|---------------------|------------------------------|----------------------------|
| 111111 | 0,00114 | 0,00342 | 0,00228 | 0,0 | 0,0 | 0,0 | | |
| 111112 | 0,00494 | 0,01482 | 0,00494 | 0,00494 | 0,0 | 0,0 | | |
| 111211 | 0,00949 | 0,01898 | 0,00949 | 0,00949 | 0,0 | 0,00949 | | |
| 111212 | 0,04113 | 0,08226 | 0,0 | 0,08226 | 0,0 | 0,04113 | | |
| Total | 0,0567 | 0,11948 | 0,01671 | 0,09669 | 0,0 | 0,05062 | | |
| Valor ser usado | 0,06 | 0,12 | 0,02 | 0,1 | 0,0 | 0,05 | 0,240 | 0,05 |
| Nova P(\rightarrow) | | 0,5 | 0,08 | 0,42 | 0,0 | 1,0 | | |

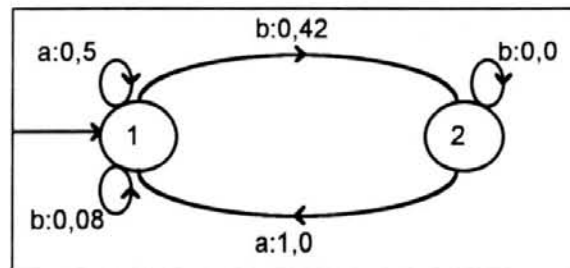


Figura 3.16 - HMM Após a Terceira Iteração

O algoritmo é executado desta forma até que se tenha chegado a valores ideais, ou até que o número total de iterações, que já foi definido, tenha sido executado.

3.6.2.3 Problemas

Na seção acima, foi apresentado o algoritmo de Forward-Backward para realizar o ajuste dos parâmetros probabilísticos de um HMM. Quando se utiliza este algoritmo, costuma-se lidar com alguns problemas.

O primeiro dos problemas diz respeito aos valores iniciais dos parâmetros do HMM. Estes valores devem ser atribuídos, antes do início da execução do algoritmo. Como não há maneira de determinar estes valores, os mesmos são atribuídos ao acaso e o algoritmo se encarrega de ajustá-los.

O caso das probabilidades estarem em um “ponto crítico” é o segundo problema apresentado: pode haver situações em que o algoritmo se defronte com duas possibilidades igualmente boas para seguir. Neste caso, ele não sabe qual das possibilidades deve escolher para obter o melhor resultado. Ficando neste impasse, não consegue aprimorar as probabilidades do modelo. Esta situação pode ser vista no exemplo abaixo, que usa o mesmo HMM da figura 3.13, para reconhecer a seqüência “aabbab”, mas com diferentes probabilidades iniciais:

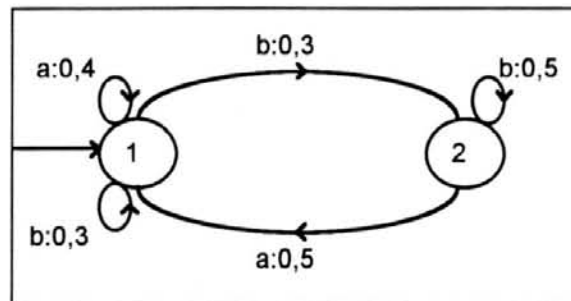


Figura 3.17 - Segundo HMM Inicial

Tabela 3.9 - Primeira Iteração

| Caminho | P(Cam) | $1 \xrightarrow{a} 1$ | $1 \xrightarrow{b} 1$ | $1 \xrightarrow{b} 2$ | $2 \xrightarrow{b} 2$ | $2 \xrightarrow{a} 1$ | $\Sigma 1 \xrightarrow{w} s$ | $\Sigma 2 \xrightarrow{w} s$ |
|-------------------------|---------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|------------------------------|------------------------------|
| 1111111 | 0,00173 | 0,00519 | 0,00519 | 0,0 | 0,0 | 0,0 | | |
| 1111112 | 0,00173 | 0,00519 | 0,00346 | 0,00173 | 0,0 | 0,0 | | |
| 1111211 | 0,00216 | 0,00432 | 0,00432 | 0,00216 | 0,0 | 0,00216 | | |
| 1111212 | 0,00216 | 0,00432 | 0,00216 | 0,00432 | 0,0 | 0,00216 | | |
| 1112211 | 0,00360 | 0,00720 | 0,00360 | 0,00360 | 0,00360 | 0,00360 | | |
| 1112212 | 0,00360 | 0,00720 | 0,0 | 0,00720 | 0,00360 | 0,00360 | | |
| Total | 0,01498 | 0,03342 | 0,01873 | 0,01901 | 0,00720 | 0,01152 | | |
| Valor ser usado | 0,02 | 0,03 | 0,02 | 0,02 | 0,01 | 0,01 | 0,07 | 0,02 |
| Nova P(\rightarrow) | | 0,4 | 0,3 | 0,3 | 0,5 | 0,5 | | |

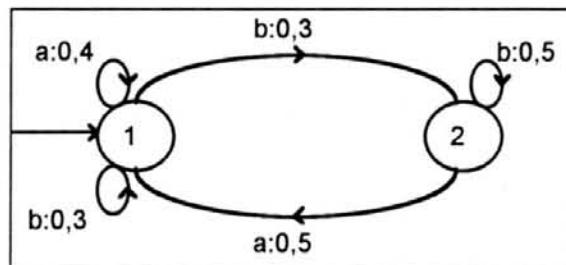


Figura 3.18 - HMM Após a Primeira Iteração

Tabela 3.10 - Segunda Iteração

| Caminho | P(Cam) | 1 \xrightarrow{a} 1 | 1 \xrightarrow{b} 1 | 1 \xrightarrow{b} 2 | 2 \xrightarrow{b} 2 | 2 \xrightarrow{a} 1 | $\Sigma 1 \xrightarrow{w} s$ | $\Sigma 2 \xrightarrow{w} s$ |
|-------------------------|---------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|------------------------------|------------------------------|
| 1111111 | 0,00173 | 0,00519 | 0,00519 | 0,0 | 0,0 | 0,0 | | |
| 1111112 | 0,00173 | 0,00519 | 0,00346 | 0,00173 | 0,0 | 0,0 | | |
| 1111211 | 0,00216 | 0,00432 | 0,00432 | 0,00216 | 0,0 | 0,00216 | | |
| 1111212 | 0,00216 | 0,00432 | 0,00216 | 0,00432 | 0,0 | 0,00216 | | |
| 1112211 | 0,00360 | 0,00720 | 0,00360 | 0,00360 | 0,00360 | 0,00360 | | |
| 1112212 | 0,00360 | 0,00720 | 0,0 | 0,00720 | 0,00360 | 0,00360 | | |
| Total | 0,01498 | 0,03342 | 0,01873 | 0,01901 | 0,00720 | 0,01152 | | |
| Valor ser usado | 0,02 | 0,03 | 0,02 | 0,02 | 0,01 | 0,01 | 0,07 | 0,02 |
| Nova P(\rightarrow) | | 0,4 | 0,3 | 0,3 | 0,5 | 0,5 | | |

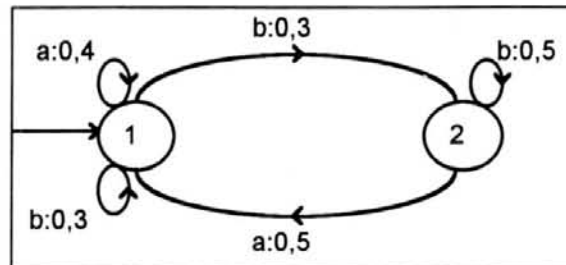


Figura 3.19 - HMM Após a Segunda Iteração

Se forem comparados os valores resultantes da primeira e da última iteração, pode-se observar que os mesmos não se modificaram. Isto pode ser dito tanto em relação à probabilidade da seqüência (0,2), quanto em relação às probabilidades das transições (0,4; 0,3; 0,3; 0,5; 0,5). O algoritmo não consegue mais aprimorar os parâmetros probabilísticos do modelo. Para resolver este problema uma solução muito simples é a de adicionar um certo "ruído" (um valor qualquer) aos valores iniciais das transições, de tal forma que permita que o algoritmo saia do impasse e faça sua escolha.

Outra questão que surge diz respeito aos valores globais. O algoritmo de Forward-Backward é executado na tentativa de encontrar os melhores valores para as probabilidades do modelo. No entanto, apesar deste algoritmo ter garantias de

achar os melhores valores locais, não se pode garantir que ache os melhores valores globais.

Para entender melhor este problema, considere, agora, o seu método de funcionamento. Este algoritmo realiza a sua execução até que ache o melhor valor possível, ou seja, até que chegue em um ponto onde qualquer mudança reduza a probabilidade dos parâmetros. Quando ele chega neste ponto, encerra a sua execução. Contudo, este pode não ser o melhor valor possível, quando se considera todo o espaço de busca. Isto pode ser visto no seguinte gráfico:

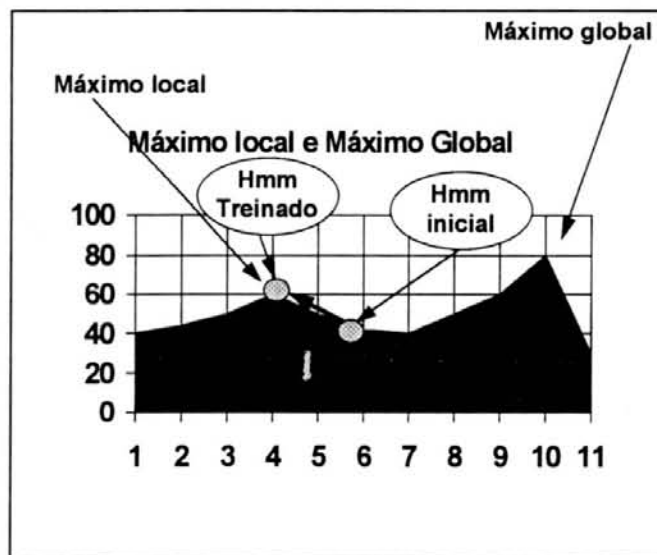


Figura 3.20 - Gráfico do Melhor Local x Melhor Global

Da execução do algoritmo irá resultar que o HMM irá se aprimorar após o treinamento, atingindo um ponto máximo, a partir de onde se encontra. Entretanto, este ponto que é atingido não é o máximo global, mas apenas o máximo local. O máximo local, que será atingido é o ponto que contém os melhores valores para as probabilidades naquela região. Contudo, na verdade, os melhores valores se encontram em outro ponto: no máximo global.

Os fatores que determinam o ponto máximo encontrado pelo algoritmo, são os valores iniciais aleatoriamente atribuídos. Devido a isto, encontrar o “máximo global”, é basicamente uma questão de sorte na escolha dos parâmetros iniciais.

3.6.3 Algoritmo de Viterbi

Este algoritmo faz o “tagging” a nível de sentença. Quando se faz o “tagging” de uma sentença, tenta-se descobrir qual a seqüência mais provável de rótulos para uma determinada sentença. Em um HMM, isto equivale a descobrir a seqüência de estados percorridos para gerar aquela sentença, ou seja, o caminho mais provável.

O algoritmo de Viterbi é uma maneira muito simples e otimizada de fazer o cálculo do caminho mais provável. Tem-se uma seqüência de símbolos $W_{1,t-1}$ e um tempo t para realizar o reconhecimento. O algoritmo é executado analisando um símbolo da seqüência a cada espaço de tempo, começando pelo símbolo inicial. Para o símbolo que está sendo analisado, o algoritmo calcula o caminho mais provável para chegar a cada um dos estados existentes. Assim, somente se fará o cálculo do **melhor caminho resultante em cada um dos estados**.

Para o HMM da figura 3.21, quer se descobrir qual a seqüência de estados que tem maior probabilidade de ter gerado a seqüência “aab”. Como resultado da aplicação do algoritmo de Viterbi, se obteve os caminhos que aparecem na figura 3.22.

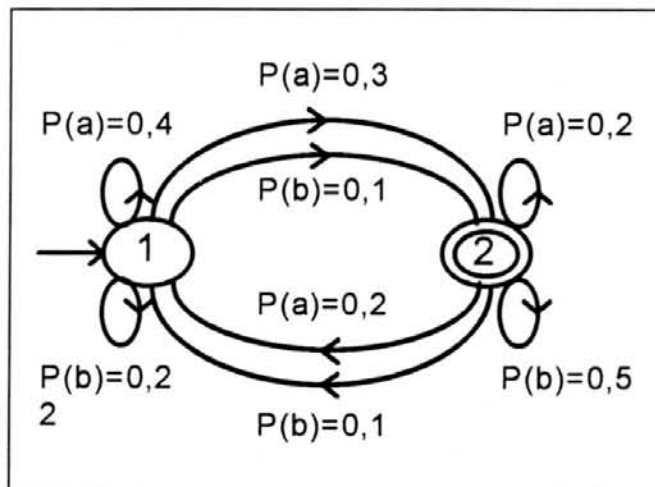


Figura 3.21 - HMM Com Dois Estados

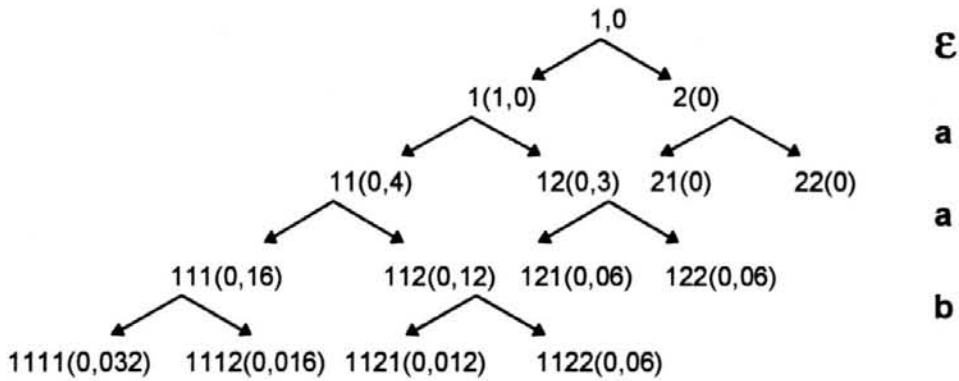


Figura 3.22 - Árvore da Sequência "aab" usando o Algoritmo de Viterbi

No primeiro espaço de tempo, é realizada a entrada no HMM com o símbolo ϵ , que tem probabilidade 1 para o estado 1 e 0 para o estado 2. A seguir, em $t=2$, vem o símbolo **a** com a maior probabilidade de chegar ao estado 1 de 0,4 (11) e ao estado 2 de 0,3 (12). As opções restantes são descartadas. Tem-se, novamente, o símbolo **a**, e o melhor caminho para se chegar a 1 é através da transição para 1 mesmo (111), com probabilidade de 0,16 e para 2 (112) com 0,12. Somente estas opções são consideradas. O último símbolo é **b**, e ao estado 1 se chega com 0,032 (1111) e a 2 com 0,06 (1122). Destes dois, o caminho que apresenta uma maior certeza é 1122 com 0,06.

Observa-se aqui que, em cada uma das unidades de tempo, dado um símbolo, não basta armazenar somente o MELHOR dentre todos os caminhos, mas sim o melhor caminho que chegue a cada um dos estados com aquele símbolo. Por exemplo, na figura 3.22, no terceiro nível, o melhor caminho que acaba em 1 (111) tem uma probabilidade de 0,16, enquanto que aquele que acaba em 2 (112) tem 0,12. Se somente o caminho 111 fosse armazenado (e o 112 descartado), não se chegaria ao melhor caminho para toda a sequência (1122), que é uma ramificação de 112.

Este mesmo problema, da procura da sequência de rótulos mais provável para a sequência "aab", solucionado sem o uso do algoritmo de Viterbi resultou nos caminhos da figura 3.23.

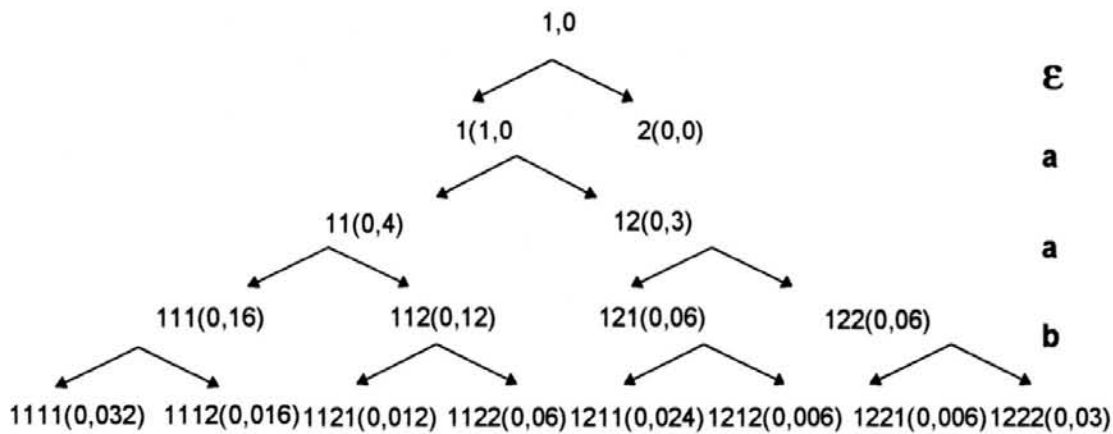


Figura 3.23 - Árvore da Sequência "aab"

Comparando-se com o resultado das figura 3.22 e 3.23, vê-se que na figura 3.23 foram gerados 8 possíveis caminhos, enquanto que na 3.22, com o uso do algoritmo de Viterbi, foram gerados apenas 4. O número de ramificações geradas diminuiu significativamente. Logo, foram feitos muito menos cálculos. E chega-se ao mesmo resultado: o melhor caminho apresenta uma certeza de 6% (1122).

A árvore apresentada tem somente duas sub-árvores em cada nível, devido à existência de apenas dois estados neste HMM (que pode ser visto na figura 3.2). **Isto significa que a procura do melhor caminho, que era um problema exponencial, se tornou linear com o uso do algoritmo de Viterbi.**

Em uma definição mais formal, o algoritmo de Viterbi pode ser descrito por:

$$i(1) = s^i$$

Fórmula 3.22

$$i(t+1) = \sigma_j(t)os^i, \quad j = \arg \max_{k=1}^{\sigma} P(\sigma_k(t))P(s^k \xrightarrow{w_t} s^i)$$

Fórmula 3.23

onde $\sigma_i(t)$ é uma seqüência de estados com a máxima probabilidade de gerar a seqüência de símbolos $w_{1,t-1}$ (de tamanho t-1) e que acaba no estado s^i .

A primeira equação, fórmula 3.22, pode ser lida como: o melhor caminho para se gerar uma seqüência de símbolos de tamanho 1 e se chegar no estado s^i é o próprio estado s^i . Já a segunda equação, fórmula 3.23, define que o melhor

caminho a que se pode chegar no tempo $t+1$ é simplesmente uma continuação do melhor caminho obtido no tempo t , para cada um dos estados.

3.6.4 Esparcidade dos Dados

Este problema, que ocorre com ambos os algoritmos, o de Frequência Relativa (FR) e o de “Forward-Backward” (FB), diz respeito à escolha de um “corpus” de treinamento que represente bem a realidade. Isto é necessário, uma vez que o HMM se ajustará a este “corpus” refletindo a ocorrência das estruturas gramaticais do mesmo. Se este “corpus” de treinamento for bem representativo, o HMM poderá reconhecer estruturas similares em outros “corpora”.

No entanto, quando se trabalha com volumes de dados muito grandes, certas construções gramaticais mais raras costumam aparecer. Caso estas estruturas raras não tenham sido encontradas no “corpus” de treinamento, a probabilidade que o modelo atribuirá a elas será igual a zero. Este fato gera uma série de problemas, pois, dada uma seqüência de palavras que contenha esta estrutura, o modelo atribuirá às mesmas uma probabilidade de zero. Apesar da frequência com que uma única destas estruturas raras ocorre não ser muito significativa, quando se leva em conta a ocorrência de **todas** as estruturas raras em um “corpus”, se terá um valor bastante alto. É o que nos afirma a lei de Zipf [SCA92]: *“Muitos poucos fenômenos ocorrem frequentemente. A maioria dos fenômenos ocorrem muito infrequentemente. Deve sempre ser avaliado se uma dada quantificação estatística faz sentido”*. Neste caso, o resultado será um modelo inútil que não consegue reconhecer certas sentenças.

Tome-se como exemplo o HMM da figura 3.13, a seqüência “de treinamento “aabab” e a Tabela 3.6. Após a primeira iteração, a transição $2 \xrightarrow{b} 2$ fica com valor zero, porque não ocorre nenhuma vez naquela seqüência de treinamento. Isto leva o HMM a “supor” que tal transição não deva ser utilizada e desta forma a desabilita. Considere agora, que o HMM deva reconhecer a seqüência “aabab”. Tem-se os seguintes caminhos: 1111111, 1111112, 1111211, 1111212, 1112211, 1112212, onde alguns deles utilizam aquela transição. Contudo, como esta transição tem valor

zero, e o HMM a desabilitou, ele nem mesmo leva em consideração a possibilidade dos caminhos 1112211 e 1112212 terem sido utilizados.

3.6.4.1 Refinamento

Para resolver este problema, são usadas técnicas para fazer o refinamento dos valores. Uma técnica bastante simples para resolver este problema é definir um limite inferior para os valores dos parâmetros. Se algum parâmetro tiver valor inferior ao limite definido, o valor do parâmetro será trocado por este limite a cada iteração. Após, os parâmetros são recalculados, para que incorporem estes ajustes. Isto evita que se tenha seqüências consideradas impossíveis [BRIS94].

Outra possibilidade é definir que os parâmetros que tenham valor inferior ao limite definido, recebam o valor zero. Esta última abordagem se constitui numa forma de aprendizado gramatical, porque o número de transições é reduzido a cada iteração, resultando num modelo mais enxuto, mais restrito [BRIS94].

Uma boa descrição dos métodos de refinamento como o “Método de Good-Turing” e o “Método de Deleted Interpolation”, usados por [CUT92], [CHU88] entre outros, pode ser encontrada em [CHU91].

3.7 Treinamento com Corpus Marcado X Treinamento com Corpus não Marcado

Os rotuladores estocásticos podem ser classificados em dois tipos, dependendo do modo com que são treinados. Todavia, independentemente do tipo de treinamento, o modelo básico utilizado para construir o sistema rotulador é o mesmo.

O treinamento de um HMM difere quanto ao tipo de “corpus” utilizado: pode se usar tanto um “corpus” marcado quanto um “corpus” não marcado. A escolha do tipo de treinamento dependerá da disponibilidade ou não de um “corpus” marcado para a língua em questão.

No primeiro tipo de treinamento, quando se dispõe de um “corpus” marcado, o treinamento é relativamente fácil. Se pode facilmente obter as freqüências com que os padrões da linguagem ocorrem, usando o algoritmo de FR. Como já foi

visto na seção 3.6.1, a partir do “corpus” marcado, se pode facilmente estimar as frequências das palavras (probabilidades lexicais) e dos n-gramas (probabilidades contextuais).

Relembrando, probabilidades lexicais são definidas como $P(w|t)$: a probabilidade de uma palavra dado o seu rótulo.

Probabilidades contextuais são calculadas como $P(t_i|t_{i-1}, \dots, t_{i-n})$ com o valor do “ n ” dependendo do tamanho da janela contextual que está sendo usada ($P(t_i|t_{i-1})$ no caso de bigramas ou $P(t_i|t_{i-1}, t_{i-2})$ no caso de trigramas).

Este tipo de treinamento se procede da seguinte maneira: primeiramente um “corpus” marcado é usado para treinar (estimar os parâmetros) o modelo. A seguir, o modelo treinado é usado para marcar um outro “corpus” qualquer, que então é manualmente corrigido e, após, usado para retreinar o modelo. Feito isto, o sistema rotulador pode ser usado para rotular outros “corpora”.

Quando se usa este tipo de treinamento, é aconselhável que se tenha um “corpus” marcado bastante grande, para que as estimativas dos parâmetros sejam confiáveis e assim se evite o problema da esparcidade dos dados, seção 3.6.4. Church [CHU88] usa o “Brown Corpus” marcado, com 1.100.000 palavras e consegue uma precisão que varia de 95-99%. Já em [CHA93] são usadas apenas 12.284 palavras para o treinamento, com a precisão variando entre 76,4% e 96,83%.

Merialdo [MER94] realizou algumas experiências comparando a performance dos algoritmos de Frequência Relativa e o de “Forward-Backward” para fazer o treinamento de um tagger, chegando à seguinte constatação:

- a estimativa de parâmetros feita mediante a contagem das frequências (usando o FR), a partir de um “corpus” marcado, resulta em uma maior precisão, posteriormente, quando se faz o “tagging”;
- quanto mais texto marcado for usado para o treinamento, melhores serão os resultados obtidos.

No segundo tipo de treinamento de rotuladores estocásticos, não se necessita de texto marcado. Isto se constitui numa possibilidade para a maioria das línguas, uma vez que são poucas que possuem um “corpus” marcado. Entretanto, apesar disto, necessita-se de um grande dicionário para determinar os possíveis rótulos para as palavras e de um grande “corpus” de treinamento (não marcado), a partir do qual os padrões lingüísticos serão automaticamente inferidos.

Se um dicionário “on-line” não está disponível para a língua do “corpus” sendo marcado ou se os rótulos existentes no dicionário não podem ser mapeados nos rótulos desejados, então muito trabalho manual é necessário para prover este material de treinamento, que é indispensável.

O sistema rotulador descrito por [CUT92] é deste tipo. Ele utiliza o algoritmo de “Forward-Backward” para fazer o treinamento do modelo. Este treinamento se realiza com a estimativa dos parâmetros de um HMM, a partir de um “corpus” não marcado. Após, este modelo treinado é usado para fazer o “tagging” de outros “corpora” quaisquer. Para o treinamento, foram usadas 500.000 palavras do “Brown Corpus” não marcadas, sendo que as 500.000 restantes foram usadas para verificar a performance obtida, que é superior a 96%.

A vantagem do uso desta abordagem é que não é necessário que se disponha de um “corpus” marcado, o que é o caso da maioria das línguas. Além disto, se assume que o modelo correto é aquele no qual os rótulos são usados para melhor refletir a seqüência de palavras. Contudo, apesar desta performance bastante elevada, ainda não está claro se o treinamento com “corpora” não marcados provê um método efetivo e portátil de “tagging”. Por exemplo, em [CUT92], para se obter tão alta precisão, foi necessário um grande dicionário com informações sobre rótulos e as inflexões das palavras. Além disto, um grande número de procedimentos de alto nível foi manualmente construído, baseado na análise manual de erros.

Para este tipo de treinamento, Merialdo [MER94] sugere que, se possível, se utilize um pequeno “corpus” manualmente marcado, apenas para construir e inicializar os parâmetros do modelo, antes de iniciar o treinamento. Para tanto, a partir deste “corpus” marcado, são extraídas as freqüências usando o algoritmo de FR.

Com isto, se pode obter um melhor modelo inicial e um aumento resultante de 3% no tratamento de palavras ambíguas [CHA93].

Em ambos os tipos de treinamento, além dos programas não apresentarem quase nenhuma restrição aos textos a serem analisados, eles conseguem ter uma precisão bastante alta, em torno de 96%, usando apenas modestos recursos de tempo e espaço [CHU93]. No entanto, a maioria dos sistemas rotuladores encontrados na literatura segue a primeira estratégia, a do uso de “corpus” marcado [CHU88, KEM94a, MER94].

4. IMPLEMENTAÇÃO DO SISTEMA ROTULADOR

O sistema rotulador, descrito neste trabalho, foi projetado, implementado e testado em conjunto com o grupo de Processamento de Linguagem Natural da Universidade Nova de Lisboa, em Portugal. Este trabalho faz parte do projeto de cooperação internacional "Processamento de Língua Natural" patrocinado pela JNICT e pelo CNPq.

Este sistema realiza o tratamento de textos irrestritos, para a Língua Portuguesa. O rotulador, tem como objetivo processar um texto de entrada e obter como resultado os rótulos de categorias morfo-sintáticas correspondentes a cada uma das palavras deste texto. Para isto, ele deve resolver a ambigüidade lexical destas palavras, sendo capaz de lidar com palavras desconhecidas, que não estejam definidas no seu dicionário.

Os métodos estatísticos se mostram perfeitamente adequados ao problema proposto, uma vez que se necessita de uma ferramenta que possa suportar a análise de textos irrestritos, sem apresentar maiores problemas, e com um custo relativamente baixo. Outro motivo para a escolha de métodos estatísticos, é o fato de o conhecimento do sistema ser inferido automaticamente, sem a necessidade de supervisão humana. O tempo de implementação de tal sistema é linear, uma vez que é totalmente independente do tempo de modelagem do conhecimento que ele utiliza, que é realizado de modo automático.

Um sistema como o rotulador, que realiza uma "pré-análise" dos dados de entrada, facilita enormemente o trabalho de um outro sistema cuja proposta seja realizar uma análise de mais alto nível nestes dados, tal como um analisador sintático. Assim, este sistema, de mais alto nível, já não precisa se preocupar em resolver questões básicas, como ambigüidades à nível lexical e tratamento de palavras desconhecidas. E, desta forma, pode tratar de textos irrestritos.

Este sistema rotulador foi totalmente projetado para trabalhar com a Língua Portuguesa, para a qual, até o momento, não havia sido feito nenhum deste

tipo. Basicamente, este sistema é capaz de receber, como entrada, uma sentença na Língua Portuguesa, processá-la e gerar, como saída, os rótulos de categorias morfo-sintáticas correspondentes às palavras que formam esta sentença.

Para ser capaz de executar tal tarefa, em uma etapa anterior é feito o treinamento do sistema. Durante o treinamento, é apresentado, ao sistema, um “corpus” de treinamento manualmente marcado com os rótulos de categorias morfo-sintáticas de cada palavra deste “corpus”. Este sistema analisa os padrões lingüísticos presentes no “corpus” e incorpora este conhecimento através do uso de HMMs. Pode, então, usar o conhecimento adquirido para fazer a marcação de outros “corpora” quaisquer que lhe forem apresentados. Este sistema utiliza métodos estatísticos para realizar a marcação dos “corpora”.

Apesar de ter sido projetado para a Língua Portuguesa, nada impede que o sistema rotulador seja treinado e usado para uma outra língua qualquer. Isto poderá facilmente ser feito, se houver um “corpus” de treinamento marcado a disposição.

O processo de construção do sistema rotulador será explicado nas próximas seções.

4.1 Conjunto de Rótulos

A primeira tarefa a ser realizada é a definição de um conjunto de rótulos a ser usado na marcação do Radiobras Corpus. Este conjunto deve ser capaz de conter os rótulos que formam as estruturas lingüísticas que se deseja modelar.

Para definir o conjunto de rótulos, primeiramente se fez um breve estudo, baseado na análise de um pequeno “corpus” marcado. Neste estudo, procurou-se verificar quais as categorias que deveriam ser incluídas no conjunto de rótulos.

Um critério utilizado, para definir a inclusão ou não de um rótulo, foi o das limitações apresentadas pelo rotulador, um HMM de primeira ordem. Como já foi dito, por ser um HMM de primeira ordem (bigrama), o rotulador consegue usar como

contexto, no máximo um rótulo anterior ao que está sendo analisado. Sendo assim, não haveria nenhum sentido em modelar dependências de longa distância, que não pudessem ser modeladas por um bigrama. Logo, os rótulos escolhidos, foram aqueles que apresentaram relações que podiam ser resolvidas dentro deste limite imposto por um HMM de primeira ordem.

Outro critério utilizado foi o do tamanho do conjunto de rótulos. Este é um critério muito importante, uma vez que o tamanho do conjunto de rótulos tem influência direta no número de parâmetros do HMM. Quanto mais parâmetros houver, maior deverá ser o “corpus” de treinamento, para que se possa obter estimativas confiáveis. No entanto, mesmo com esta limitação, o tamanho dos conjuntos de rótulos citados na literatura varia muito. Para a Língua Inglesa, este número varia de 48 rótulos usados no PennTrebek à 425 rótulos usados no SUSANNE. Para a Língua Francesa, se tem 88 rótulos e para a Chinesa, 46 rótulos.

Assim sendo, definiu-se, para a Língua Portuguesa (do Brasil), um conjunto com 33 rótulos e para a Língua Portuguesa (de Portugal), um conjunto com 45. Estes números diferem, devido às diferentes estruturas lingüísticas usadas em cada um destes países. Mas, apesar deste trabalho estar sendo feito tanto para o Português do Brasil quanto para o Português de Portugal, somente o primeiro será descrito aqui.

Nesta primeira etapa do projeto, decidiu-se que os rótulos teriam somente informações sobre a categoria morfo-sintática das palavras. Não foram acrescentadas características específicas como gênero, número, modo, tempo verbal, etc.

O conjunto de rótulos definido está listado na tabela 4.1:

Tabela 4.1 - Conjunto de Rótulos Usado

| Rótulos | Significado | Exemplos |
|---------|--|------------------|
| AF | Afixo | ex, sub |
| ART | Artigo | o, a |
| ADJ | Adjetivo | maravilhoso |
| ADV | Advérbio | agora, ontem |
| CH | Caractere | “ ” ’ , , ’ |
| CONJ | Conjunção | e, também |
| CONT | Contração | do, na |
| DPTO | Dois Pontos | : |
| INT | Ponto de Interrogação, Ponto de Exclamação | ?, ! |
| N | Substantivo | revista, carro |
| NC | Numeral Cardinal | dois, mil |
| NCOL | Numeral Coletivo | centenas |
| NO | Numeral Ordinal | primeiro |
| NP | Substantivo Próprio | Portugal, Mário |
| PAR | Parênteses | (,) |
| PD | Pronome Demonstrativo | aquele, este |
| PIND | Pronome Indefinido | alguma, ninguém |
| PPOA | Pronome Pessoal do Caso Oblíquo | lhe, me, a, o |
| PPS | Pronome Possessivo | meu, vosso |
| PR | Pronome Relativo | que, qual |
| PPR | Pronome Pessoal do caso Reto | ele, nós |
| PREP | Preposição | para, de |
| PTO | Ponto Final | . |
| TRAV | Hífen, Travessão | -, _ |
| UNKNOWN | Palavra Desconhecida | |
| VAUX | Verbo Auxiliar | ficar, continuar |
| VI | Verbo Intransitivo | correr |
| VIRG | Virgula | , |
| VPP | Verbo no Particípio | sonhado, parado |
| VSER | Verbo Ser | sou, sois |
| VTD | Verbo Transitivo Direto | comprar |
| VTER | Verbo Ter | tenho, terão |
| VTI | Verbo Transitivo Indireto | sonhar |

Como se pode observar, na tabela 4.1, as principais categorias morfológicas estão descritas no conjunto de rótulos: os artigos (ART), os adjetivos (ADJ), os advérbios (ADV), as conjunções (CONJ), os substantivos (N, NP), os

numerais (NO, NC, NCOL), as preposições (PREP, CONT), os pronomes (PD, PIND, PPOA, PPS, PR, PPR), os verbos (VI, VTD, VTI, VAUX, VPP, VSER, VTER).

Há categorias gramaticais inteiras que são representadas por um único rótulo, como é o caso dos artigos, enquanto que outras podem ser descritas por mais de um rótulo, como os verbos. Estas subdivisões, em uma mesma categoria gramatical, foram criadas quando se tornou necessário distinguir determinadas características relevantes para a captura dos padrões lingüísticos. No caso das preposições, definiu-se que estas seriam descritas por PREP. Já, no caso das contrações formadas por preposição seguida de artigo, como por exemplo “do”, ou ainda por preposição seguida de pronome, como “daquela”, decidiu-se definir um rótulo especial (CONT). Isto ocorre por englobarem mais de uma categoria sintática em uma única palavra. A seguir, serão brevemente descritos os rótulos e o domínio atingido por eles. Os verbos serão explicados com maiores detalhes devido ao nível de complexidade que apresentam.

Cada uma das seguintes categorias é descrita por um único rótulo: os **artigos** são descritos por **ART**, os **adjetivos** por **ADJ**, os **advérbios** por **ADV**, as **conjunções** por **CONJ**.

Além destas, existem as categorias gramaticais que são subdivididas, como é o caso de **N**, que descreve todos os **substantivos**, com exceção dos **substantivos próprios** que são descritos por **NP**.

Há também o caso dos numerais, onde o rótulo **NO** engloba os **numerais ordinais**, **NC**, os **numerais cardinais** e **NCOL** os **numerais coletivos** e demais numerais.

Já as **preposições** são descritas por **PREP**. Contudo, como já foi explicado acima, há também o rótulo **CONT**, que engloba as **contrações** de preposições e artigos e de preposições e pronomes.

Os pronomes são divididos em seis classes, cada uma descrita por um rótulo: **pronomes demonstrativos (PD)**, **pronomes indefinidos (PIND)**, **pronomes**

pessoais oblíquos (PPOA), pronomes pessoais do caso reto (PPR), pronomes possessivos (PPS), e pronomes relativos e interrogativos (PR).

Os verbos foram divididos em 7 classes. Estas classes foram delimitadas pelas observações dos principais padrões encontrados. O verbo *ser* e o verbo *ter*, devido à sua importância, têm rótulos exclusivos (VSER e VTER). Contudo, com exceção destes dois verbos, os demais são englobados pelos outros 5 rótulos (VAUX, VPP, VI, VTD e VTI). Construções verbais simples são classificadas de acordo com a transitividade do verbo: verbos intransitivos (VI), transitivos diretos (VTD) e transitivos indiretos (VTI).

Uma outra classe verbal definida é a de VAUX. Este rótulo foi definido tendo em vista construções simples, com um verbo ou ainda construções mais complexas como as locuções verbais, que necessitem de verbos auxiliares. Há também a classe descrita pelo rótulo VPP, que é encontrado nas locuções verbais.

As sete classificações serão descritas abaixo:

- 1) **verbo ser (VSER)** - este verbo tem um rótulo exclusivo para ele, por ser um dos verbos mais frequentemente utilizados na Língua Portuguesa. Ele está presente em várias das estruturas verbais existentes. Pode ser usado tanto em construções verbais simples quanto em estruturas compostas. Por exemplo: **era** uma moça, havia **sido**, **és** chamado, **fostes**, etc.
- 2) **verbo ter (VTER)** - este verbo também tem um rótulo exclusivo. A exemplo do verbo *ser*, também aparece com frequência nas construções verbais simples e compostas. Exemplo: **tenho** andado, havíamos **tido**, **teríeis**, etc.
- 3) **verbo intransitivo (VI)** - este rótulo é atribuído aos verbos intransitivos, como por exemplo: nadar, correr, etc.
- 4) **verbo transitivo direto (VTD)** - é o rótulo que está relacionado com os verbos transitivos diretos. Exemplos deste tipo de verbo são: **amar** (alguém), **saber** (alguma coisa), etc.

- 5) **verbo transitivo indireto (VTI)** - nesta categoria se incluem os verbos transitivos indiretos como: **comprar** (de alguém), **vender** (a alguém), etc.
- 6) **verbos auxiliares (VAUX)** - este rótulo representa os verbos auxiliares e os verbos de ligação (exceto os verbos ser e ter, que tem rótulos exclusivos): estar, andar, achar-se, ficar, fazer-se, haver, continuar, permanecer, parecer, etc. Podem ser utilizados tanto em construções simples quanto em compostas. As construções compostas, conhecidas como locuções verbais, são formadas por um verbo auxiliar (VAUX, VSER ou VTER) seguido por um verbo principal (VPP, VSER ou VTER). Exemplo: Ele **parecia** doente (construção simples), **fiquei fazendo** a lição de casa (construção composta).
- 7) **verbos no particípio (VPP)** - este rótulo é atribuído a verbos que façam parte de locuções verbais. Este rótulo foi criado para representar construções lingüísticas formadas por um verbo auxiliar seguido de um verbo indicativo da ação. Este segundo verbo pode estar no infinitivo, no particípio ou no gerúndio, tal como: tinha **corrido**, haviam **voltado**, estávamos **comprando**, saberiam **comprar**, etc.

Como se pode observar, os verbos transitivos, tanto diretos como indiretos, estão definidos no conjunto de rótulos. Resolveu-se descrever estes rótulos porque o complemento verbal (objeto direto ou objeto indireto) geralmente ocorre em uma distância que pode ser abrangida pelo modelo de bigramas.

*Sentença 4.1: O pintor **expõe** os seus quadros.*

*Sentença 4.2: Maria já **agradeceu** ao Paulo.*

Na sentença 4.1 acima, *os seus quadros* é o objeto direto do verbo **expor**, enquanto que na sentença 4.2, *ao Paulo* é o objeto indireto do verbo **agradecer**.

Entretanto, apesar dos verbos transitivos diretos e dos transitivos indiretos estarem sendo tratados, o mesmo não ocorre com os verbos que apresentam ambas as transitividades: os **verbos transitivos diretos e indiretos**. Isto ocorre porque a distância necessária para que se possa capturar ambos os complementos (o objeto

direto e o indireto) supera em muito a distância permitida pelos bigramas, que é de apenas um rótulos além do que está sendo analisado.

Sentença 4.3: Eu ganhei um fogão dos meus tios no mês passado.

Na sentença 4.3, o verbo **ganhar** tem, como objeto direto, *um fogão* e, como objeto indireto, *dos meus tios*.

Para solucionar este problema, decidiu-se tratar apenas o complemento que seja imediatamente vizinho ao verbo. Assim, o verbo terá a transitividade definida de acordo com o complemento mais próximo a ele. No caso da sentença 4.3, o verbo **ganhar** é definido pelo rotulador como um verbo transitivo direto, pois o complemento *um fogão* é um objeto direto. Já no caso de:

Sentença 4.4: Eu ganhei dos meus tios um fogão.

ganhar é definido como verbo transitivo indireto, pois o complemento *dos meus tios* é um objeto indireto. É uma solução bastante simples mas eficiente que foi encontrada para solucionar tal problema.

Os **sinais gráficos** possuem também um papel muito importante a cumprir: são eles que delimitam as fronteiras de uma sentença. Por este motivo, para eles também foram definidos os respectivos rótulos (**CH, DPTO, INT, PAR, PTO, TRAV, VIRG**). Porém, pode-se pensar que, uma vez que além dos rótulos das palavras, se incluem os rótulos dos sinais gráficos, o trabalho do sistema rotulador irá aumentar. Felizmente, na realidade, isto não acontece, porque os sinais gráficos têm uma grande vantagem: à nível sintático, eles não apresentam nenhuma ambigüidade. Uma vírgula é sempre uma vírgula e o rotulador a rotulará como tal.

Para as palavras que não estão definidas no dicionário, há um rótulo especial: **UNKNOWN**.

4.2 Classes de Ambigüidade

O “corpus” que está sendo usado neste trabalho é o Radiobras Corpus, marcado com os rótulos de categorias morfo-sintáticas, descritos na tabela 4.1. Devido ao seu tamanho relativamente pequeno, 21.000 palavras manualmente marcadas, resolveu-se utilizar o conceito de classes de ambigüidade. Como já foi explicado na seção 2.3.1, o uso de classes de ambigüidade diminui o problema da esparcidade dos dados. Desta forma, definiu-se que todas as palavras do “corpus” pertencem a alguma classe de ambigüidade. Assim, usando-se as classes de ambigüidade, espera-se que as estimativas feitas a partir deste pequeno “corpus” se tornem mais confiáveis.

4.3 O Radiobras Corpus

A Editoria de Ciência e Tecnologia da Agência Brasil transmite o boletim “C&T Radiobras”, com notícias com tópicos de interesse geral, em formato eletrônico, através da Internet. Estes boletins serviram para construir o Radiobras Corpus. Para tanto foram coletados boletins, durante cerca de 18 meses, totalizando 141.043 palavras. Uma parte de um boletim da Radiobras será apresentada, para que se possa visualizar o formato da mesma, figura 4.1:

C & T RADIOBRAS Numero: 181
 29 de julho de 1994 Editora: Marta Crisostomo

INDICE:
 COLUNA DE CIENCIA E TECNOLOGIA E MEIO AMBIENTE
 SOJA/INVIABILIDADE
 PESQUISA/PARQUES

SOJA/INVIABILIDADE
 Brasilia, 29 (Agencia Brasil - ABR) - Pesquisadores do Mato Grosso do Sul estao preocupados com o cultivo da soja feito de forma indiscriminada. Eles estiveram reunidos na unidade de pesquisa da Empresa Brasileira de Pesquisa Agropecuaria (EMBRAPA), em Dourados, este mes, e constataram que o nao cumprimento da rotacao de culturas que pode facilitar o aumento de doencas de dificil controle. Outra preocupacao eh com a monocultura da soja, que tambem contribui para a proliferacao de pragas.
 MC/AM

PESQUISA/PARQUES
 Rio de Janeiro, 29 (Agencia Brasil - ABR) - Para provar que os parques ecologicos sao um excelente local para o desenvolvimento de pesquisas, o Instituto Estadual de Florestas (IEF) de Minas Gerais decidiu equipar o Parque Florestal do Rio Doce, no leste de Minas, com facilidades para atrair cientistas de todo o pais. "Parque ecologico soh para turista ver eh coisa do passado. Temos que estimular a presenca de cientistas nos parques porque isso ajuda a conservacao e a troca de conhecimentos entre o pesquisador e os funcionarios", disse o zoologo Celio Valle, diretor do IEF, no 20' Congresso Brasileiro de Zoologia, realizado no Rio. No parque do Rio Doce, onde se encontra o maior conjunto lacustre do Brasil, com 40 lagoas, foram montados alojamentos confortaveis para 80 cientistas.
 MC/TC
 NNNN

Assinatura: envie mensagem para listserv@cr-df.rnp.br com o seguinte conteudo:
 subscribe ct-radiobras seu nome-sua instituicao(sigla)

O C&T RADIOBRAS e um servico da Editoria de Ciencia e Tecnologia da Agencia Brasil (Radiobras), especializado em divulgar, para jornais de todo o pais, a ciencia Brasileira. Precisamos que pesquisadores de todas as areas entrem em contato conosco, apresentando seus trabalhos, que serao pautados e transformados em material jornalistico para distribuicao via jornais e pela Internet. Eventos cientificos tambem sao divulgados , em forma de notas.

Contribuicoes: enviar para: radiobr@cnpq.br ou
 Marta Crisostomo
 Editoria de C&T - Agencia Brasil
 SRTVS quadra 701 conj, B lote 3 - Ed. Radio Nacional
 70.332-900 Brasilia - DF
 Fone: (061) 223-9878 e 322-1695 - Fax: (061) 226-1377

Figura 4.1 - Boletim da Radiobras

O boletim começa com seus dados indicativos, sendo seguido por um índice com os assuntos tratados e, finalmente, começam as notícias, cada uma com o seu título. Após o título, há dados sobre local, data e agência responsável e a notícia propriamente dita. Para sinalizar o final da notícia, é colocado MC/ seguido de outros dois símbolos (TC, AM, MW, etc), figura 4.2.

```

*****
C & T RADIOBRAS          Numero: 181
29 de julho de 1994      Editora: Marta Crisostomo
*****
INDICE:
COLUNA DE CIENCIA E TECNOLOGIA E MEIO AMBIENTE
SOJA/INVIABILIDADE
PESQUISA/PARQUES

SOJA/INVIABILIDADE
Brasilia, 29 (Agencia Brasil - ABR) - Pesquisadores do Mato
Grosso do Sul estao preocupados com o cultivo da soja feito de
forma indiscriminada. Eles estiveram reunidos na unidade de
pesquisa da Empresa Brasileira de Pesquisa Agropecuaria
(EMBRAPA), em Dourados, este mes, e constataram que o nao
cumprimento da rotacao de culturas que pode facilitar o aumento
de doencas de dificil controle. Outra preocupacao eh com a
monocultura da soja, que tambem contribui para a proliferacao de
pragas.
MC/AM

-----
PESQUISA/PARQUES
Rio de Janeiro, 29 (Agencia Brasil - ABR) - Para provar que
os parques ecologicos sao um excelente local para o
desenvolvimento de pesquisas, o Instituto Estadual de Florestas
(IEF) de Minas Gerais decidiu equipar o Parque Florestal do Rio
Doce, no leste de Minas, com facilidades para atrair cientistas
de todo o pais. "Parque ecologico soh para turista ver eh coisa
do passado. Temos que estimular a presenca de cientistas nos
parques porque isso ajuda a conservacao e a troca de
conhecimentos entre o pesquisador e os funcionarios", disse o
zooologo Celio Valle, diretor do IEF, no 20' Congresso Brasileiro
de Zoologia, realizado no Rio. No parque do Rio Doce, onde se
encontra o maior conjunto lacustre do Brasil, com 40 lagoas,
foram montados alojamentos confortaveis para 80 cientistas.
MC/TC
NNNN

```

Figura 4.2 - Notícia da Radiobras

O final do boletim é sinalizado por NNNN. Seguem, após, informações gerais sobre a Radiobras, figura 4.3.

```

*****
Assinatura: envie mensagem para listserver@cr-df.rnp.br com o seguinte conteudo:
subscribe ct-radiobras seu nome-sua instituicao(sigla)
-----
O C&T RADIOBRAS e um servico da Editoria de Ciencia e Tecnologia da Agencia
Brasil (Radiobras), especializado em divulgar, para jornais de todo o pais, a
ciencia Brasileira. Precisamos que pesquisadores de todas as areas entrem em
contato conosco, apresentando seus trabalhos, que serao pautados e transforma-
dos em material jornalistico para distribuicao via jornais e pela Internet.
Eventos cientificos tambem sao divulgados , em forma de notas.
-----
Contribuicoes: enviar para: radiobr@cnpq.br ou
Marta Crisostomo
Editoria de C&T - Agencia Brasil
SRTVS quadra 701 conj, B lote 3 - Ed. Radio Nacional
70.332-900 Brasilia - DF
Fone: (061) 223-9878 e 322-1695 - Fax: (061) 226-1377
*****

```

Figura 4.3- Final do Boletim da Radiobras

Antes de iniciar qualquer tentativa de processamento, um “corpus” deve ser pré-processado, como será explicado a seguir.

4.3.1 Preparação do corpus

Para preparar o “corpus” para a análise, de acordo com esta ferramenta, há três tarefas que devem ser executadas. A primeira diz respeito à retirada de frases que não apresentem estrutura sintática, como títulos e figuras. A segunda, trata da transformação de todos os caracteres maiúsculos em minúsculos. E a terceira, da utilização exclusiva do *ponto final* para delimitação da extensão da sentença.

Como se quer modelar, no rotulador, os padrões lingüísticos de sentenças escritas na Língua Portuguesa, não há sentido em analisar títulos, figuras, tabelas, gráficos, cabeçalhos, caracteres de controle, etc. Para tanto, resolveu-se fazer um pré-processamento do “corpus” e retirar todas estas informações que não se deseja utilizar na análise.

Este mesmo procedimento deve ser repetido para a marcação de outros “corpora”. Desta forma, antes do sistema rotulador fazer a marcação de um outro “corpus” qualquer, este deverá ser pré-processado.

Do boletim mostrado acima, somente o seguinte trecho será analisado:

Pesquisadores do Mato

Grosso do Sul estão preocupados com o cultivo da soja feito de forma indiscriminada. Eles estiveram reunidos na unidade de pesquisa da Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), em Dourados, este mês, e constataram que o não cumprimento da rotação de culturas que pode facilitar o aumento de doenças de difícil controle. Outra preocupação é com a monocultura da soja, que também contribui para a proliferação de pragas.

Para provar que

os parques ecológicos são um excelente local para o desenvolvimento de pesquisas, o Instituto Estadual de Florestas (IEF) de Minas Gerais decidiu equipar o Parque Florestal do Rio Doce, no leste de Minas, com facilidades para atrair cientistas de todo o país. "Parque ecológico só para turista ver é coisa do passado. Temos que estimular a presença de cientistas nos parques porque isso ajuda a conservação e a troca de conhecimentos entre o pesquisador e os funcionários", disse o zoológico Celio Valle, diretor do IEF, no 20º Congresso Brasileiro de Zoologia, realizado no Rio. No parque do Rio Doce, onde se encontra o maior conjunto lacustre do Brasil, com 40 lagoas, foram montados alojamentos confortáveis para 80 cientistas.

Figura 4.4 - Boletim Após o Processo de Retirada de Frases que não Apresentem Estrutura Sintática

Selecionado o texto a ser analisado, o próximo passo é transformá-lo em um texto só com caracteres minúsculos. Isto deve ser feito porque os caracteres maiúsculos estão sendo usados exclusivamente pelos rótulos. Assim, após o processo de troca, o texto da figura 4.4 fica conforme indicado na figura 4.5.

pesquisadores do mato grosso do sul estão preocupados com o cultivo da soja feito de forma indiscriminada. eles estiveram reunidos na unidade de pesquisa da empresa brasileira de pesquisa agropecuária (embrapa), em dourados, este mês, e constataram que o não cumprimento da rotação de culturas que pode facilitar o aumento de doenças de difícil controle. outra preocupação é com a monocultura da soja, que também contribui para a proliferação de pragas.

para provar que os parques ecológicos são um excelente local para o desenvolvimento de pesquisas, o instituto estadual de florestas (ief) de minas gerais decidiu equipar o parque florestal do rio doce, no leste de minas, com facilidades para atrair cientistas de todo o país. "parque ecológico só para turista ver é coisa do passado. temos que estimular a presença de cientistas nos parques porque isso ajuda a conservação e a troca de conhecimentos entre o pesquisador e os funcionários", disse o zoólogo celió valle, diretor do ief, no 20º congresso brasileiro de zoologia, realizado no rio. no parque do rio doce, onde se encontra o maior conjunto lacustre do brasil, com 40 lagoas, foram montados alojamentos confortáveis para 80 cientistas.

Figura 4.5 - Boletim Após o Processo de Troca

A última tarefa é garantir a utilização do ponto final (".") apenas para sinalizar o final das sentenças. Uma situação que deve ser tratada é a aplicação do ponto final para a separação de dígitos dos numerais cardinais. Costuma-se usar tanto o ponto final quanto a vírgula. Para solucionar esta situação, o número "12.000", por exemplo, será transformado em "12000".

Outro caso a ser tratado é o dos pronomes de tratamento, que costumam ser abreviados usando o ponto final. Desta forma, "Sr." será transformado em "Senhor", "V. Ema" em "Vossa Eminência" e assim por diante.

Este procedimento se torna necessário, uma vez que o sistema rotulador considera que, a extensão de cada sentença do "corpus", é delimitada pelo ponto final. Então, todas as palavras que estiverem situadas entre um ponto final e outro, serão consideradas componentes de uma mesma sentença.

4.3.2 Marcando o Radiobras Corpus

Após este pré-processamento, foram separadas 20.982 palavras, que foram manualmente marcadas. Para realizar esta marcação, foram usados os rótulos descritos em 4.1.

Para facilitar o trabalho de marcação manual, o “corpus” foi processado pelo módulo classificador do sistema rotulador, que será explicado na seção 4.5.1. Deste processamento, resultaram as classes de ambigüidade de cada uma das palavras do “corpus”.

em PREP cada PIND cruzamento N serao N_VSER instalados
VPP lacos N sensores N (PAR fios N detectores N para
CONJ_PREP_VI veiculos N colocados ADJ nas CONT vias N
) PAR , VIRG oito NC por PREP cruzamento N , VIRG ateh
ADV_PREP o ART_PD_PPOA final N de N_PREP convenio
N de N PREP pesquisa N . PTO

Figura 4.6 - Classes de Ambigüidade das Palavras

A partir do resultado gerado pelo módulo classificador, escolhe-se o rótulo correto para cada palavra. Ao lado de cada palavra do “corpus”, foi marcado o rótulo de categoria morfo-sintática correspondente, totalmente escrito em maiúsculas. O rótulo foi mantido separado da palavra por apenas um espaço em branco, de acordo com a figura 4.7.

em PREP cada PIND cruzamento N serao VSER instalados
VPP lacos N sensores N (PAR fios N detectores N para
PREP veiculos N colocados ADJ nas CONT vias N) PAR ,
VIRG oito NC por PREP cruzamento N , VIRG ateh PREP o
ART final N de PREP convenio N de PREP pesquisa N . PTO

Figura 4.7 - Marcação das Palavras

Mesmo utilizando este recurso, foram gastas 44 horas de trabalho de uma pessoa, para realizar a marcação manual destas 20.982 palavras.

Realizada a marcação, este “corpus” foi dividido em duas partes:

- a primeira parte, com 20.000 palavras, é denominada de “corpus” de treinamento;
- a segunda parte, com as restantes 982 é denominada de “corpus” de teste.

Como resultado da execução destas etapas, obtém-se um “corpus” de treinamento e um “corpus” de teste marcados.

4.3.3 Corpus de Treinamento e Corpus de Teste

O “corpus” de treinamento tem um papel vital no desenvolvimento do sistema rotulador. É o conhecimento contido neste “corpus” que será modelado no rotulador: ele terá os seus parâmetros estimados a partir deste “corpus”. É assim que serão incorporadas as estruturas lingüísticas.

O “corpus” de teste servirá para medir a precisão do sistema rotulador depois de treinado. A partir dos resultados obtidos na marcação deste “corpus” é que serão feitos os ajustes e retreinamento do sistema rotulador.

4.4 O Dicionário

O dicionário usado pelo sistema rotulador contém informações das classes de ambigüidade relativas a cada uma das palavras nele descritas. Ele é gerado a partir do “corpus” marcado manualmente. Analisa-se todo o “corpus” marcado e descobrem-se todos os rótulos que foram usados para marcar cada uma das palavras. Convém lembrar que, todos estes rótulos possíveis, para uma mesma palavra, formam a sua classe de ambigüidade.

O método usado para construir este dicionário é bastante simples. O “corpus” está marcado com um rótulo para cada palavra, conforme figura 4.8.

| | | | | | | | | | |
|-------------|------------|----------|----------|------------|------------|--------------|-----|------------|------|
| o | ART | primeiro | NO | semaforo | N | ' | CH | ' | CH |
| inteligente | ADJ | ' | CH | ' | CH | desenvolvido | ADJ | no | |
| CONT | brasil | NP | terah | VTER | , | VIRG | no | CONT | |
| proximo | ADJ | semestre | N | , | VIRG | um | ART | software | |
| N | de | PREP | operacao | N | automatica | ADJ | que | PR | , |
| VIRG | entre | PREP | outras | PIND | vantagens | N | , | | |
| VIRG | permitirah | VTD | a | elaboracao | N | diaria | ADJ | | |
| de | PREP | planos | N | de | PREP | trafego | N | diferentes | |
| ADJ | a | PREP | cada | PIND | minuto | N | , | para | PREP |
| estabelecer | VTD | o | ART | fluxo | N | ideal | ADJ | de | |
| PREP | veiculos | N | . | PTO | | | | | |

Figura 4.8 - Palavras no Corpus Marcado

De posse deste “corpus”, faz-se a sua ordenação por palavras. Desta forma, todas as ocorrências de uma mesma palavra ficarão dispostas seqüencialmente e se poderá saber todos os rótulos que, em algum momento, foram atribuídos a ela, como foi demonstrado na figura 4.9.

| | |
|----------|-------------|
| a | ART |
| a | CONT |
| a | PPOA |
| a | PREP |

Figura 4.9 - Corpus Ordenado

A seguir, é só agrupar todos os rótulos possíveis, para uma palavra em uma classe de ambigüidade. Então, de posse disto, o dicionário é construído, contendo as palavras e suas respectivas classes de ambigüidade. O resultado pode ser visto na seguinte figura:

| | |
|----------------|---------------------------|
| a | ART_CONT_PPOA_PREP |
| abacate | N |

Figura 4.10 - Duas Entradas do Dicionário

4.5 Arquitetura do Sistema Rotulador

O sistema rotulador realiza a análise de um “corpus” de treinamento e modela os padrões lingüísticos nele presentes. Após ter modelado este conhecimento, o rotulador pode usá-lo para fazer a marcação automática das palavras de um outro “corpus” qualquer.

Este sistema é composto de três módulos principais: o módulo construtor de HMMs, o módulo classificador e o módulo de Viterbi. Sua arquitetura é apresentada na figura 4.11, onde são destacadas as etapas de treinamento e de marcação.

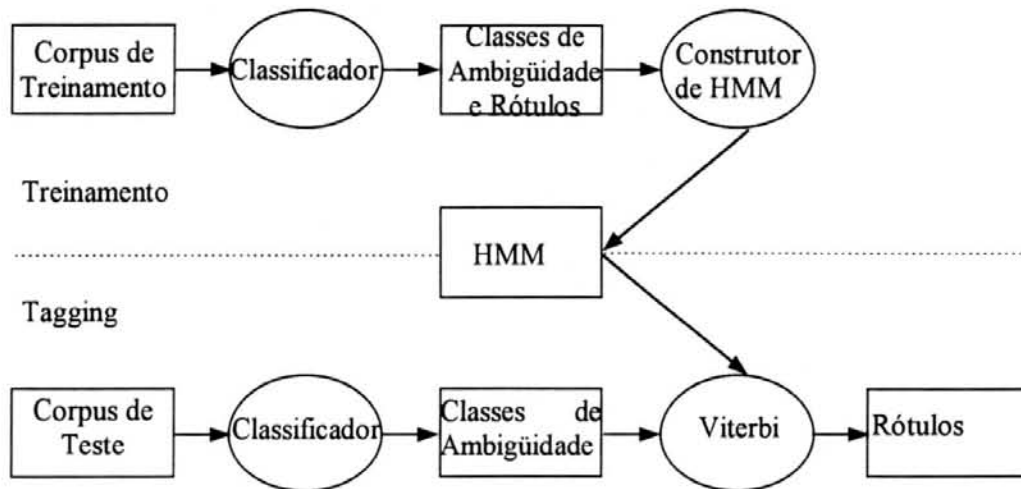


Figura 4.11 - A Arquitetura do Sistema Rotulador

Nas próximas seções, estes módulos são explicados em detalhes.

4.5.1 Módulo Classificador

O primeiro módulo do sistema rotulador, o classificador, recebe um “corpus” como entrada. Ele lê todas as sentenças do “corpus”, decompõe cada uma delas em palavras e atribui, para cada uma das palavras, uma classe de ambigüidade, conforme definido no dicionário.

Tendo sido feita a sua decomposição, a sentença ainda deve manter sua estrutura. Para tanto, definiu-se que uma sentença se estende até a ocorrência do próximo ponto final. Assim, o seu domínio irá abranger todas as palavras que estiverem entre um ponto final e outro.

Sentença 4.5: A mansão foi construída em 1879. Ela já pertenceu a um presidente.

Acima tem-se duas sentenças: a primeira - “A mansão foi construída em 1879” - e a segunda - “Ela já pertenceu a um presidente”.

Posteriormente, o classificador atribui para cada palavra, a classe de ambigüidade correspondente, através de uma procura no dicionário (figuras 4.12 e 4.13).

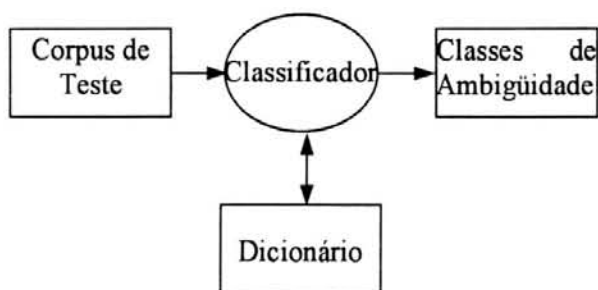


Figura 4.12 - O Módulo Classificador

| | |
|--------------------|---------------------------|
| a | ART_CONT_PPOA_PREP |
| menina | N_ADJ |
| corria | VI |
| rapidamente | ADV |
| . | PTO |

Figura 4.13 - Palavras Analisadas pelo Classificador

Uma característica que deve ser mencionada, é o fato de que cada palavra, no “corpus”, é tratada individualmente. Deste modo, estruturas formadas por mais de uma palavra, como as locuções, terão seus elementos tratados individualmente.

Sentença 4.6 - Logo que chegamos, estava escuro.

Nesta sentença, “logo que” é uma locução conjuntiva, mas é tratada como:

**Logo ADV que CONJ chegamos VI , VIRG estava VAUX escuro
ADJ . PTO**

Substantivos próprios também receberão o mesmo tratamento: “Joseh da Silva” equivale a Joseh NP da CONT Silva NP e “Organizacao Mundial de Saude” é Organizacao N Mundial ADJ de PREP Saude N. Outro caso é o de palavras como “d’agua” que é tratada como d/PREP ‘/CH agua/N. Da mesma forma, também as palavras compostas ou que tenham afixos separados por hífen, terão tratamento semelhante. Por exemplo, “ex-reitor” equivale a ex AF - TRAV reitor N; “guarda-sol” é tratado como guarda N - TRAV sol N.

Quando o sistema rotulador estiver na **fase de treinamento**, necessita-se de dados sobre os rótulos e as classes de ambigüidade relacionadas a cada palavra. Estes dados são obtidos usando a versão do “corpus” de treinamento marcado e a versão original (não marcada). O “corpus” de treinamento não marcado é usado como entrada para o módulo classificador que devolve, como saída, as classes de ambigüidade correspondentes a cada palavra. E o “corpus” de treinamento marcado é usado para obter o rótulo de cada palavra. Juntam-se estas duas versões do “corpus” de treinamento, e se obtém a entrada deste módulo, que é um “corpus” composto de classes de ambigüidade e rótulos de categorias morfo-sintáticas.

Na **fase de marcação**, após ser pré-processado, qualquer “corpus” deverá passar por este módulo, antes de ser analisado pelo rotulador. Receberá, então, as classes de ambigüidade correspondentes às suas palavras, que servirão de entrada para o módulo de Viterbi. Logo, o trabalho do rotulador será decidir qual o rótulo correto, dada a classe de ambigüidade, para a palavra em questão.

4.5.2 Módulo Construtor de HMMs

O segundo módulo é o construtor de HMMs. Ele é o responsável por construir o HMM a partir do “corpus” de treinamento marcado com rótulos e classes de ambigüidade. Através de uma análise dos padrões lingüísticos que ocorrem no “corpus de treinamento”, o construtor irá produzir, como resultado, um HMM.

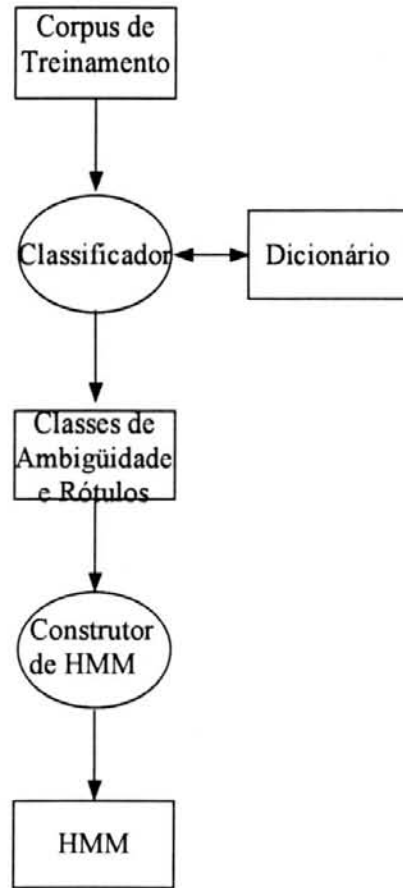


Figura 4.14 - A Arquitetura do Construtor

Para modelar as probabilidades contextuais, se utilizou o modelo de n-gramas, mais especificamente, o de bigramas. Para fazer o cálculo das estimativas destas probabilidades, está-se usando o algoritmo de Frequência Relativa, que faz esta estimativa a partir do Radiobras Corpus marcado, fórmula 4.1. E como utiliza-se classes de ambigüidade, o modelo da linguagem terá a seguinte forma, com C_i correspondendo à classe de ambigüidade da palavra w_i :

$$p(W,T) = \prod p(t_i, w_i | t_{i-1}) = \prod p(t_i, C_i | t_{i-1}).$$

Fórmula 4.1

Como se pode ver pela fórmula acima, substituiu-se o uso das palavras, w_i , pelo uso das classes de ambigüidade, C_i . Isto não apenas reduz o número de parâmetros a serem estimados no modelo, como também reduz o problema de

esparcidade dos dados. Além disto, é fácil de ser implementado usando HMMs. A fórmula da probabilidade contextual é descrita da seguinte forma:

$$p(t_i C_i | t_{i-1}) = \frac{f(t_{i-1}, t_i C_i)}{f(t_{i-1})}$$

Fórmula 4.2

Depois de feitas as estimativas das probabilidades, faz-se o refinamento destas. Este refinamento serve para evitar que se use probabilidades que tenham sido estimadas a partir de frequência muito baixas. Para o refinamento, usa-se o algoritmo de “Deleted Interpolation” [MER94].

$$p(t_i C_i | t_{i-1}) = \varepsilon \frac{f(t_{i-1}, t_i C_i)}{f(t_{i-1})} + (1 - \varepsilon) \frac{1}{N_T N_C}$$

Fórmula 4.3

onde N_T corresponde ao número de rótulos e N_C ao número de classes de ambigüidade. Este algoritmo especifica que, antes de aplicar o algoritmo de Frequência Relativa, um pedaço do texto de treinamento marcado seja separado. O restante do “corpus” é, então, usado para estimar as frequências. Escolhe-se um coeficiente ε , com valor entre 0 e 1, que maximize a probabilidade de emissão do texto que foi separado, no modelo interpolado (refinado). Este coeficiente é usado para ajustar as probabilidades, através da fórmula 4.3. Para maiores explicações, sobre o funcionamento do algoritmo, consultar [JEL80].

Após o modelo ter sido refinado, ele está pronto para utilizar o seu conhecimento adquirido para fazer a marcação de outros “corpora”.

4.5.3 Módulo de Viterbi

Depois de haver passado pelo módulo classificador e receber as classes de ambigüidade, o “corpus” é enviado ao módulo de Viterbi. Este módulo analisa as classes de ambigüidade encontradas em um “corpus”. Através da aplicação do algoritmo de Viterbi sobre o HMM treinado, descobre qual a seqüência de rótulos mais provável.

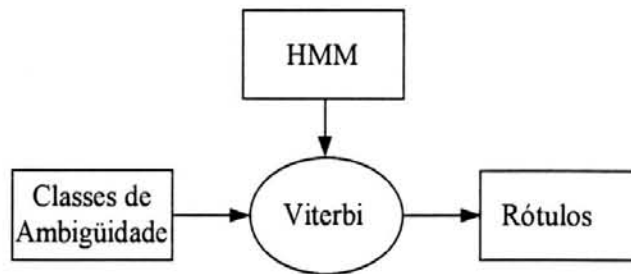


Figura 4.15 - A Arquitetura do Classificador

O algoritmo de Viterbi, percorre o HMM treinado, tentando descobrir qual é a seqüência de estados que tem maior probabilidade de ter gerado a seqüência de símbolos de entrada. Como o rotulador é um HMM, onde os estados equivalem aos rótulos e as classes de ambigüidade aos símbolos das transições, o algoritmo de Viterbi percorrerá o HMM para descobrir qual é a seqüência de rótulos mais provável para a sentença analisada. Pode haver mais de uma seqüência de rótulos (estados) possível para uma mesma sentença, mas o algoritmo irá selecionar o que apresenta maior probabilidade. Como está explicado na seção 6.3.6, este algoritmo tem complexidade linear, significando que o número de caminhos, que é um problema exponencial, com o uso do algoritmo de Viterbi, torna-se um problema linear.

Este processo de seleção da seqüência mais provável de rótulos para uma sentença é denominado de resolução de ambigüidades. A ambigüidade existente no “corpus” é resolvida, quando o algoritmo de Viterbi escolhe o rótulo mais correto, dentre os permitidos pela classe de ambigüidade. A seqüência de rótulos escolhida é aquela que apresenta maior probabilidade.

4.5.4 Testes de Funcionamento

Para ter certeza de que o sistema rotulador está funcionando de acordo com o desejado, existem alguns testes que podem ser feitos.

O primeiro deles, que serve para verificar a aquisição do conhecimento feita pelo rotulador, pode ser feito usando-se o próprio “corpus” de treinamento. Usa-se o sistema rotulador para marcar o “corpus” de treinamento não marcado. A seguir, comparam-se os resultados com os do “corpus” de treinamento manualmente marcado,

para ver como o sistema rotulador executou a sua tarefa. Teoricamente, o sistema rotulador deverá apresentar um resultado bastante alto, visto que o “corpus” que ele acabou de marcar é o mesmo que lhe transmitiu todo o seu conhecimento. Deve-se, porém, observar que os resultados deste teste não podem ser usados para medir a precisão do sistema rotulador, pois os dados analisados já eram previamente conhecidos. Eles servem apenas para comprovar se o rotulador incorporou corretamente o conhecimento a ele transmitido e está funcionando corretamente. Caso estes valores não sejam bons, o sistema rotulador deverá ser revisado.

O segundo teste tem por finalidade medir a precisão do sistema rotulador. Para tanto, será usado o “corpus” de teste para testar o sistema rotulador treinado. De que forma isto será feito? O “corpus” de teste na versão original (não marcada) será processado pelo módulo classificador, que gerará as classes de ambigüidade para ele. A seguir, o módulo de Viterbi será aplicado, gerando como resultado, a seqüência de rótulos mais provável para o “corpus” de teste. Após, a marcação feita automaticamente pelo sistema rotulador será comparada com a marcação manual. Pode-se, então, calcular a precisão que foi obtida pelo sistema rotulador. Além de servirem para o cálculo da precisão, as diferenças entre a marcação manual e a automática mostrarão as deficiências do sistema rotulador.

4.6 Marcação

Terminados os testes, o sistema rotulador pode ser usado para fazer a marcação automática de outros “corpora”, demonstrado na figura 4.16. Os “corpora” são, então, processados pelo classificador, de onde são obtidas as classes de ambigüidade.

Todas as palavras têm suas classes de ambigüidade associadas, e algumas delas podem ser classificadas como UNKNOWN (desconhecidas). Como se utiliza a informação contextual para resolver ambigüidades, a ocorrência de uma palavra desconhecida dificulta tanto a sua marcação quanto a marcação das palavras vizinhas.

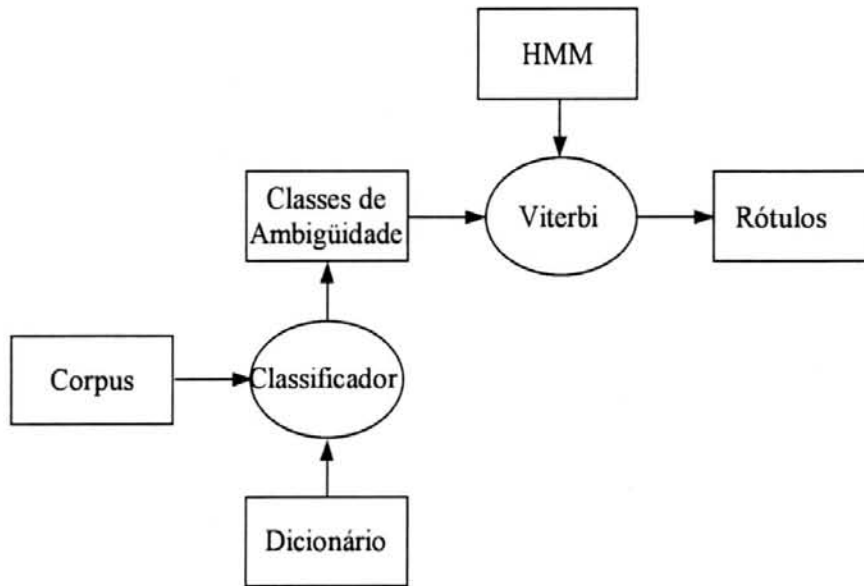


Figura 4.16 - Marcação de um Corpus

A tarefa do sistema rotulador, então, é descobrir qual é a seqüência de rótulos mais provável para esta seqüência de classes de ambigüidade. Para tanto, ele tem que ser capaz de tratar as palavras ambíguas e as palavras desconhecidas. A existência de mecanismos de tratamento de palavras ambíguas e de palavras desconhecidas, influenciará muito na precisão apresentada pelo sistema rotulador.

Para fazer a marcação destes “corpora”, é aplicado o algoritmo de Viterbi. E, como resultado, são determinados os rótulos correspondentes às palavras destes “corpora”.

4.6.1 Precisão

Para medir a precisão obtida pelo sistema rotulador construído, foi feito um teste usando o Radiobas Corpus, com 559 sentenças marcadas com rótulos de categorias morfo-sintáticas. Este “corpus” foi dividido em um “corpus” de treinamento com 13.475 palavras e um “corpus” de teste com 413 palavras. O “corpus” de treinamento foi usado para gerar um HMM de primeira ordem. Usando este sistema rotulador para marcar o “corpus” de teste, foi obtida uma precisão de 84,5% [VIL95].

Esta precisão obtida é bastante promissora, pois é equivalente à obtida por outros sistemas rotuladores citados na literatura e que usaram “corpus” de treinamento desta magnitude [KEM94b].

No próximo capítulo, será apresentada a avaliação feita sobre o rotulador. O objetivo desta avaliação é descobrir com qual padrão de comportamento o rotulador pode obter a maior precisão.

5. AVALIAÇÃO DO SISTEMA ROTULADOR

A marcação automática de categorias morfo-sintáticas usando métodos estatísticos ainda é algo muito recente. Portanto, há muito trabalho ainda a ser feito e lacunas a serem preenchidas.

Uma vez que os recursos e as ferramentas necessários para realizar esta marcação automática ainda são escassos na Língua Portuguesa, deve-se tentar aproveitar, da melhor maneira possível, aquilo que está disponível. Este problema, de escassez de recursos, está sendo defrontado não só nas pesquisas com a Língua Portuguesa, mas também com muitas outras línguas.

Tendo considerado estes aspectos, neste capítulo, é visto de que forma este trabalho pode contribuir para o esclarecimento de alguns pontos ainda obscuros, para aqueles que trabalham na área. São realizados três experimentos:

- “corpus” acentuado x “corpus” desacentuado - o primeiro aspecto a ser avaliado é a influência do número de palavras ambíguas na precisão do sistema rotulador;
- tamanho do “corpus” de treinamento - o segundo aspecto se refere ao requisito do tamanho do “corpus” de treinamento necessário para obter uma boa precisão;
- palavras desconhecidas - o último aspecto diz respeito à influência das palavras desconhecidas na precisão do sistema rotulador.

Nas próximas seções, estes aspectos serão detalhadamente descritos.

5.1 Treinamento com Corpus Acentuado

Devido ao fato de se estar trabalhando com textos não formatados, a representação dos acentos gráficos, quando existente, é bastante limitada. A maioria dos textos não apresenta qualquer tipo de acentuação. Devido a esta dificuldade de encontrar um “corpus” acentuado, se quer verificar se realmente existe a necessidade de trabalhar com tal “corpus”.

No Radiobras Corpus, o “corpus” com o qual se está trabalhando, a representação dos acentos gráficos é um tanto quanto limitada. Acentos no meio das palavras, por exemplo, não são representados, figura 5.1.

Eh muito pouco se comparado com o investimento que outros paises fazem em Ciencia e Tecnologia, mas segundo Daniel Ribeiro de Oliveira, da Secretaria de Planejamento do Governo Federal, esta realidade eh a mesma para os outros setores do Governo que disputam as verbas do orcamento da Uniao e que nao tem suas verbas vinculadas constitucionalmente as suas despesas, como eh o caso da Previdencia Social e da Educacao.

Para Luis Antonio Barreto de Castro, secretario-executivo do Programa de Apoio ao Desenvolvimento Cientifico e Tecnologico (PADCT), a questao vai ser critica por mais tempo e por isso eh necessario ter ideias criativas que possam alavancar o desenvolvimento na area. Segundo Castro eh preciso obter recursos alem do que a Uniao pode oferecer, e a empresa privada eh um parceiro natural para financiar pesquisas. No ambito do PADCT um grande numero de projetos jah estao sendo orientados pela relacao entre a universidade e a iniciativa privada.

Figura 5.1 - Corpus Acentuado

Os acentos gráficos são representados da seguinte maneira:

- acento agudo no final das palavras é representado pela adição de “h” ao final destas palavras, por exemplo, “é” é representado por “eh”;
- acento grave no final das palavras é representado pela adição do “a”, assim, “à” é representado por “aa”.

Bem mais fácil de ser encontrado, um “corpus” desacentuado é totalmente escrito sem a representação de acentos, figura 5.2.

o "seminario internacional de avaliacao e propostas para o desenvolvimento científico e tecnologico para o brasil", realizado em sao paulo, foi o palco de discussao da politica nacional de ciencia e tecnologia para o futuro. uma das conclusoes importantes tiradas durante o seminario e de que a pesquisa cientifica tem que se aproximar mais da sociedade e das necessidades do mercado gerando tecnologias que tornem mais competitiva a industria nacional. por outro lado o financiamento a ciencia e tecnologia deve vir de fontes alternativas ao governo.

Figura 5.2 - Corpus Desacentuado

Pelo que se pode observar no texto da figura acima, a ambigüidade lexical, que é um problema que ocorre muito freqüentemente, aumenta ainda mais quando se trabalha com um "corpus" desacentuado. Por exemplo, considerando a palavra "e", que aparece em duas sentenças extraídas do texto na figura 5.2.

Sentença 5.1 - "o "seminario internacional de avaliacao e propostas para o desenvolvimento ..."

Sentença 5.2 - "uma das conclusoes importantes tiradas durante o seminario e de que a pesquisa cientifica tem que se aproximar mais da sociedade"

Quando se trabalha com "corpus" desacentuado, estas duas palavras são indistingüíveis, pois têm a mesma grafia: "e" pode ser tanto uma conjunção (sentença 5.1) quanto um verbo (verbo ser, terceira pessoa do singular do presente do indicativo), sentença 5.2. Estas duas palavras, que normalmente não seriam confundidas, tornam-se ambíguas, pois são representadas da mesma maneira e não há como distingüí-las, a não ser pela análise do contexto. Uma situação aparentemente não ambígua, como a apresentada acima, torna-se ambígua em um "corpus" não acentuado. Assim, neste tipo de "corpus", além das palavras que já são ambíguas normalmente, há ainda aquelas que se tornam ambíguas pela ausência de acentos.

Intuitivamente, um "corpus" acentuado apresenta um número bem menor de palavras ambíguas do que um "corpus" desacentuado. Porém, é bem mais difícil de ser encontrado. Mas será que este maior número de palavras ambíguas em um

“corpus” desacentuado vai influenciar a precisão do sistema? Este é o primeiro fato que se quer analisar - a influência do número de palavras ambíguas sobre a precisão do sistema rotulador. Para testar esta hipótese, é feita uma experiência que compara os resultados de um sistema rotulador treinado com duas versões de um mesmo “corpus” de treinamento: o “corpus” de treinamento acentuado e o “corpus” de treinamento desacentuado. Os resultados obtidos serão apresentados na seção 5.4

5.2 Tamanho do Corpus de Treinamento

Quando se usam métodos estatísticos, recomenda-se que se use o maior “corpus” marcado possível, para que se consiga obter uma boa precisão e não se enfrente o problema de esparcidade dos dados. Mas qual é o tamanho de “corpus” necessário? Na literatura da área, a maioria dos trabalhos usa ou sugere que sejam usados “corpora” de treinamento marcados com pelo menos 1.000.000 de palavras.

No entanto, como, para a maioria das línguas, não há nem sequer “corpora” marcados disponíveis, este padrão de 1.000.000 de palavras se apresenta como uma barreira intransponível. Isto ocorre porque a tarefa da marcação de um “corpus” é extremamente dispendiosa e requer muito esforço e tempo de trabalho; não é uma tarefa trivial. Necessita-se de uma pessoa que conheça bem a língua em questão, de preferência um linguísta, para realizar a marcação manual do “corpus”. Além disto, esta é uma tarefa bastante cansativa. Considerando-se todos estes aspectos, este padrão de 1.000.000 de palavras apresenta-se excessivamente alto. Mas será que é preciso mesmo um “corpus” marcado desta magnitude para conseguir boas estimativas?

Este problema foi enfrentado no começo deste trabalho, onde um grande empecilho foi a falta de um “corpus” marcado para a Língua Portuguesa. Como não havia nenhum, a solução foi começar a marcar um manualmente. Marcou-se, então, 20.982 palavras do Radiobras Corpus, como explicado no capítulo 4. Porém, as seguintes questões começaram a se impor: “Que tamanho deve ter um “corpus” marcado? Há alguma influência do tamanho do “corpus” de treinamento sobre a precisão do sistema? Será que com um “corpus” de treinamento de 20.000 palavras é

possível obter uma boa precisão?”. Assim, esta segunda experiência trata destas questões, onde se tenta estabelecer, dentro dos recursos disponíveis, qual o padrão *ótimo* para fazer o treinamento do sistema rotulador para a Língua Portuguesa.

5.3 Dicionário Fechado x Dicionário Aberto

O tratamento das palavras desconhecidas sempre foi um ponto crítico para a performance dos sistemas rotuladores. Como se está utilizando a informação contextual, a marcação de cada palavra depende da marcação das que estão próximas a ela. Portanto, a ocorrência de uma palavra desconhecida afeta não somente a marcação dela própria, como também a marcação das palavras vizinhas (sentença 5.3).

Sentença 5.3: Maria casa hoje.

Supondo que a palavra “Maria” não esteja definida no dicionário, o módulo classificador a rotula como uma palavra desconhecida - UNKNOWN:

Maria UNKNOWN casa N_VI hoje ADV . PTO

Potencialmente, esta palavra pode ser rotulada, pelo módulo de Viterbi, como qualquer um dos rótulos abertos, seção 2.2: NP, N, ADJ, VAUX, VPP, VI, VTD, VTI. Este fato não contribui em nada para a resolução da ambigüidade de “casa”, que tanto pode ser rotulada como um substantivo quanto como um verbo intransitivo.

Mas qual é exatamente a influência das palavras desconhecidas sobre o sistema rotulador? Para responder a esta pergunta, dois dicionários diferentes são utilizados: o dicionário aberto e o dicionário fechado. O dicionário aberto é definido tendo por base todas as palavras do “corpus” de treinamento e suas respectivas classes de ambigüidade. Já ao fechado, são acrescentadas, ainda, as palavras e classes de ambigüidade do “corpus” de teste. Assim, ambos foram construídos com todas as palavras do “corpus” de treinamento, porém apenas um deles, o dicionário fechado, tem também as palavras do “corpus” de teste.

Estes dicionários são utilizados pelo módulo classificador. É este módulo que atribui a classe de ambigüidade para as palavras, de acordo com a

definição do dicionário. Quando o classificador utiliza o dicionário fechado para processar o “corpus” de teste, não há a possibilidade da ocorrência de palavras desconhecidas, porque todas as palavras do “corpus” de teste estão definidas no dicionário fechado. Por outro lado, quando o classificador utilizar o dicionário aberto, este permitirá a ocorrência de palavras desconhecidas, que são aquelas que não estão definidas no dicionário.

No primeiro passo desta experiência, o “corpus” de teste não marcado é processado pelo módulo classificador, que gera as classes de ambigüidade para as palavras deste “corpus”, usando o dicionário aberto. No segundo passo, o mesmo “corpus” de teste é processado pelo classificador, só que, desta vez, utiliza-se o dicionário fechado. A seguir, estas duas versões do “corpus” de teste são processadas pelo módulo de Viterbi. A precisão obtida pelo sistema rotulador, em cada uma das versões do “corpus” de teste, será apresentada na próxima seção.

5.4 Resultados

Para realizar a avaliação proposta por este trabalho, são definidas três variáveis, explicadas nas seções 5.1, 5.2 e 5.3:

- 1) o **“corpus” de treinamento** - são utilizadas duas versões do “corpus” de treinamento de 20.000 palavras:
 - o “corpus” acentuado e
 - o “corpus” desacentuado
- 2) o **tamanho do “corpus” de treinamento** - o “corpus” de treinamento marcado é usado para gerar 5 subconjuntos: o primeiro com 2.500 palavras, o segundo com 5.000 palavras, o terceiro com 10.000 palavras, o quarto com 15.000 e o quinto com 20.000.
- 3) o **dicionário** - são usadas duas versões deste dicionário: o dicionário aberto e o dicionário fechado.

Estas variáveis são combinadas, testadas e avaliadas, como mostra a tabela 5.1. Ao todo, são realizados 20 testes e os resultados obtidos são apresentados em detalhes nas próximas seções.

Tabela 5.1 - Testes Efetuados

| Teste | Qualidade do Corpus | | Tamanho do Corpus | | | | | Dicionários | |
|-------|---------------------|--------------|-------------------|-------|--------|--------|--------|-------------|---------|
| | Acentuado | Desacentuado | 2.500 | 5.000 | 10.000 | 15.000 | 20.000 | Aberto | Fechado |
| 1 | X | | X | | | | | X | |
| 2 | X | | | X | | | | X | |
| 3 | X | | | | X | | | X | |
| 4 | X | | | | | X | | X | |
| 5 | X | | | | | | X | X | |
| 6 | X | | X | | | | | | X |
| 7 | X | | | X | | | | | X |
| 8 | X | | | | X | | | | X |
| 9 | X | | | | | X | | | X |
| 10 | X | | | | | | X | | X |
| 11 | | X | X | | | | | X | |
| 12 | | X | | X | | | | X | |
| 13 | | X | | | X | | | X | |
| 14 | | X | | | | X | | X | |
| 15 | | X | | | | | X | X | |
| 16 | | X | X | | | | | | X |
| 17 | | X | | X | | | | | X |
| 18 | | X | | | X | | | | X |
| 19 | | X | | | | X | | | X |
| 20 | | X | | | | | X | | X |

5.4.1 Corpus de Treinamento Acentuado

Para este experimento, é utilizado o “corpus” de treinamento **acentuado**, com 20.000 palavras marcadas, a partir do qual são gerados cinco “corpora” de treinamento com tamanhos diferentes. Cada um é utilizado para treinar um HMM:

- o **HMM1** é treinado com **2.500** palavras;

- o **HMM2** é treinado utilizando **5.000** palavras;
- o **HMM3** treinado utilizando **10.000** palavras;
- o **HMM4** treinado com **15.000** e
- o **HMM5** com **20.000**.

5.4.1.1 Usando o Dicionário Aberto

A seguir, o “corpus” de teste não marcado é processado pelo módulo classificador, usando o **dicionário aberto**. Deste processamento, resultam as classes de ambigüidade para o “corpus” de teste, que são apresentadas na seção 5.4.1.1.

Cada um dos cinco HMMs é usado para fazer a marcação do “corpus” de teste acentuado, chegando aos resultados apresentados na tabela 5.2.

Tabela 5.2 - Resultados no Corpus Acentuado, Dicionário Aberto

| HMM | Corpus de Treinamento | Rótulos Corretos | Rótulos Errados | Total de Rótulos | Precisão |
|------|-----------------------|------------------|-----------------|------------------|----------|
| HMM1 | 2.500 | 644 | 338 | 982 | 65,59% |
| HMM2 | 5.000 | 672 | 310 | 982 | 68,43% |
| HMM3 | 10.000 | 697 | 285 | 982 | 70,98% |
| HMM4 | 15.000 | 731 | 251 | 982 | 74,44% |
| HMM5 | 20.000 | 732 | 250 | 982 | 74,54% |

Se pode observar que, à medida que o tamanho do texto de treinamento aumenta, a precisão do sistema rotulador também aumenta. A diferença entre a precisão do HMM1, 65,59%, e a do HMM5, 74,54%, representa um aumento de 8,95%, que é um valor bastante significativo, figura 5.3.

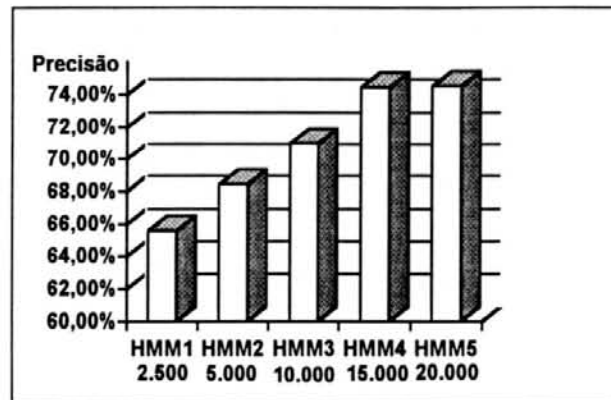


Figura 5.3 - Dicionário Aberto: Tamanho do Corpus de Treinamento Acentuado x Precisão

Outro aspecto bastante importante é a diferença de precisão entre o HMM4 e o HMM5: apenas 0,1% (74,44% - 74,54% respectivamente).

5.4.1.2 O Corpus Acentuado com o Dicionário Aberto

O “corpus” de teste acentuado usado tem 982 palavras marcadas. Quando é processado pelo classificador, usando o dicionário aberto, são encontradas 35 classes de ambigüidade, que estão listadas na tabela 5.3. Das 982 palavras, 264 são ambíguas, apresentando mais de um rótulo e 173 são desconhecidas (figura 5.4).

As palavras ambíguas e as desconhecidas representam, respectivamente, 26,88% e 17,62% das palavras do “corpus”, e equivalem juntas a quase metade das palavras, o que é um valor bastante alto, figura 5.4. Entre as palavras ambíguas, 160 têm apenas dois rótulos e 104 têm mais de dois rótulos, representando, respectivamente, 16,29% e 10,59% do total de palavras do “corpus” de teste. Das 718 palavras restantes, 173 são desconhecidas e 545 (55,50%) têm um único rótulo (figura 5.5).

Tabela 5.3 - Dicionário Aberto: Classes de Ambigüidade do Corpus Acentuado

| Classes de Ambigüidade | Frequência |
|-------------------------------|-------------------|
| ADJ - NP | 2 |
| ADJ - VPP | 2 |
| ADJ - VTD | 1 |
| ADJ - VTI | 1 |
| ADV - ART - PREP | 22 |
| ADV - CONJ | 3 |
| ADV - CONJ - PREP | 2 |
| ADV - N | 3 |
| ADV - PREP | 4 |
| ART - CARD | 15 |
| ART - CONT | 3 |
| ART - PD - PPOA | 33 |
| CONJ - N - PPOA | 2 |
| CONJ - PR - PREP | 18 |
| CONJ - PREP | 2 |
| CONJ - PREP - VTD | 15 |
| CONT - PD | 1 |
| CONT - VTD | 9 |
| N - NP | 14 |
| N - PREP | 64 |
| N - VAUX | 1 |
| N - VSER | 1 |
| N - VTD - VTI | 1 |
| N - VTI | 1 |
| NP - VSER | 7 |
| PD - VAUX - VTI | 1 |
| VAUX - VPP - VSER | 1 |
| VAUX - VTD | 3 |
| VAUX - VTD - VTI | 5 |
| VAUX - VTI | 2 |
| VI - VTD | 2 |
| VI - VTD - VTI | 1 |
| VPP - VTI | 1 |
| VTD - VTI | 2 |
| VTD - VTI | 1 |

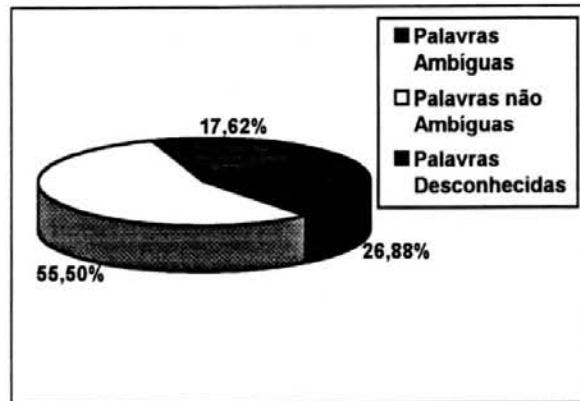


Figura 5.4 - Dicionário Aberto: Palavras Ambíguas e Palavras Desconhecidas no Corpus Acentuado

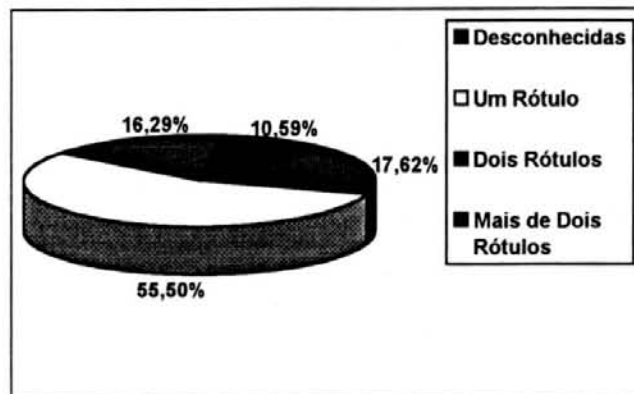


Figura 5.5 - Dicionário Aberto: Rótulos por Palavra no Corpus de Teste Acentuado

Das 173 palavras desconhecidas, 8 são advérbios, 17 são substantivos próprios, 23 são adjetivos, 56 são verbos e 69 são substantivos, representando, respectivamente, 4,62%, 9,83%, 13,30%, 32,37%, 39,88% (figura 5.6).

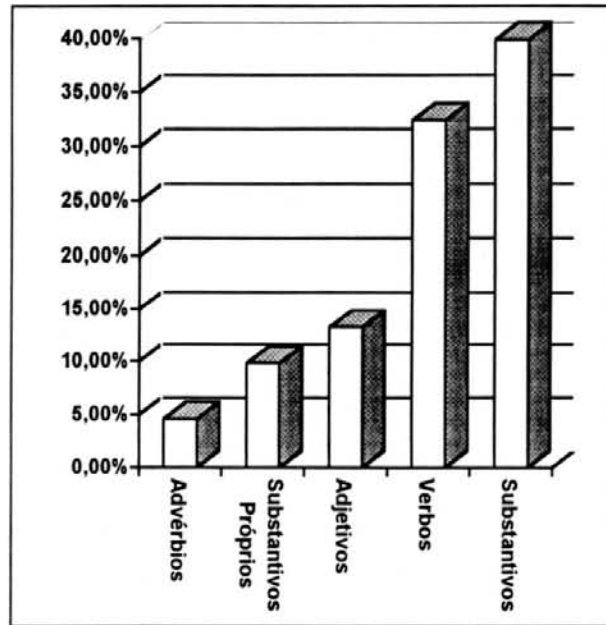


Figura 5.6 - Dicionário Aberto: Palavras Desconhecidas no Corpus Acentuado

Os substantivos respondem pela maioria das palavras desconhecidas (86 no total, com os substantivos próprios). São seguidos pelos verbos, sendo que, dos 56 12 são verbos no participípio. Há ainda os adjetivos e os advérbios.

5.4.1.3 Usando o Dicionário Fechado

Os mesmo 5 HMMs da experiência anterior, são usados nesta. Só que, o “corpus” de teste não marcado, é processado pelo módulo classificador, usando o **dicionário fechado**. Cada um dos HMMs é, então, usado para fazer a marcação deste “corpus” de teste (tabela 5.4).

Tabela 5.4 - Resultados no Corpus Acentuado, Dicionário Fechado

| HMM | Corpus de Treinamento | Rótulos Corretos | Rótulos Errados | Total | Precisão |
|------|-----------------------|------------------|-----------------|-------|----------|
| HMM1 | 2.500 | 764 | 218 | 982 | 77,81% |
| HMM2 | 5.000 | 798 | 184 | 982 | 81,26% |
| HMM3 | 10.000 | 827 | 155 | 982 | 84,22% |
| HMM4 | 15.000 | 863 | 119 | 982 | 87,88% |
| HMM5 | 20.000 | 864 | 118 | 982 | 87,98% |

Observando estes resultados, se pode verificar que a precisão aumentou 10,17%, de 77,81% do HMM1 até 87,98% do HMM5, figura 5.7.

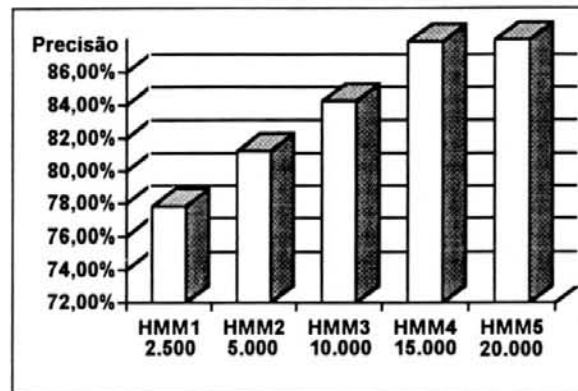


Figura 5.7 - Dicionário Fechado: Tamanho do Corpus de Treinamento Acentuado x Precisão

5.4.1.4 O Corpus Acentuado com o Dicionário Fechado

No “corpus” de teste acentuado processado pelo classificador, usando o dicionário fechado, são encontradas 267 palavras pertencentes a 37 classes de ambigüidade (tabela 5.5). Estas palavras ambíguas representam 27,19% das palavras do “corpus” (figura 5.8), entre as quais 163 têm apenas dois rótulos e 106 têm mais de dois rótulos, representando, respectivamente, 16,60% e 10,59% do total de palavras do “corpus” de teste. As restantes 715 têm um único rótulo, 72,81% (figura 5.9).

Tabela 5.5 - Dicionário Fechado: Classes de Ambigüidade do Corpus Acentuado

| Classes de Ambigüidade | Frequência |
|-------------------------------|-------------------|
| ADJ - ADV - N | 1 |
| ADJ - CONJ - N - ORD - PREP | 2 |
| ADJ - N | 17 |
| ADJ - NP | 2 |
| ADJ - VPP | 2 |
| ADJ - VTD | 1 |
| ADJ - VTI | 1 |
| ADV - ART - PREP | 22 |
| ADV - CONJ | 3 |
| ADV - CONJ - PREP | 2 |
| ADV - N | 3 |
| ADV - PREP | 5 |
| ART - CARD | 15 |
| ART - CONT | 3 |
| ART - PD - PPOA | 33 |
| CONJ - N - PPOA | 2 |
| CONJ - PR - PREP | 18 |
| CONJ - PREP | 2 |
| CONJ - PREP - VTD | 15 |
| CONT - PD | 1 |
| N - NP | 13 |
| N - PREP | 64 |
| N - VAUX | 1 |
| N - VSER | 1 |
| N - VTD - VTI | 1 |
| N - VTI | 1 |
| NP - VSER | 7 |
| VAUX - VPP - VSER | 1 |
| VAUX - VSER | 1 |
| VAUX - VTD | 2 |
| VAUX - VTD - VTI | 6 |
| VAUX - VTI | 2 |
| VI - VTD | 1 |
| VI - VTD - VTI | 1 |
| VPP - VTD | 1 |
| VPP - VTI | 5 |
| VTD - VTI | 10 |

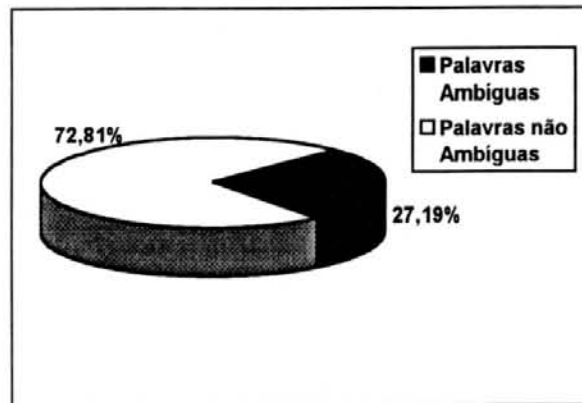


Figura 5.8 - Dicionário Fechado: Palavras Ambíguas e Palavras Desconhecidas no Corpus Acentuado

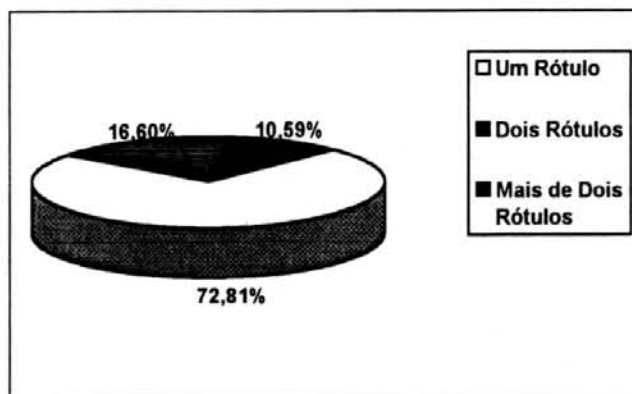


Figura 5.9 - Dicionário Fechado: Rótulos por Palavra no Corpus de Teste Acentuado

5.4.1.5 Dicionário Aberto x Dicionário Fechado

Fazendo uma comparação entre o “corpus” de teste processado com o dicionário aberto e o processado com o dicionário fechado, pode-se ver que não há um aumento muito grande no número de palavras ambíguas, que no primeiro é de 264 e no segundo de 267 - 3 palavras com apenas dois rótulos. O aumento mais significativo ocorre nas palavras com um único rótulo: 545 com o dicionário aberto e 715 com o dicionário fechado (tabela 5.6) - 170 palavras. Isto significa que a maioria das palavras desconhecidas do primeiro “corpus” se enquadram entre as palavras com um único rótulo do segundo (figura 5.10).

Tabela 5.6 - Corpus Acentuado: Dicionário Aberto x Dicionário Fechado

| | Dicionário Aberto | | | Dicionário Fechado | | |
|--------------------------------|-------------------|-----------------|----------|--------------------|-----------------|----------|
| | Rótulos Corretos | Rótulos Errados | Precisão | Rótulos Corretos | Rótulos Errados | Precisão |
| HMM1 | 644 | 338 | 65,59% | 764 | 218 | 77,81% |
| HMM2 | 672 | 310 | 68,43% | 798 | 184 | 81,22% |
| HMM3 | 697 | 285 | 70,98% | 827 | 155 | 84,21% |
| HMM4 | 731 | 251 | 74,44% | 863 | 119 | 87,88% |
| HMM5 | 732 | 250 | 74,54% | 864 | 118 | 87,98% |
| Palavras Desconhecidas | 173 | | | - | | |
| Palavras com 1 Rótulo | 545 | | | 715 | | |
| Palavras Ambíguas | 264 | | | 267 | | |
| Palavras com 2 Rótulos | 160 | | | 163 | | |
| Palavras com mais de 2 Rótulos | 104 | | | 104 | | |
| Total | 982 | | | 982 | | |

O aumento na precisão com o uso do dicionário fechado também é bastante significativo, o que permite observar a influência das palavras desconhecidas sobre a precisão obtida, como pode ser visto na figura 5.11.

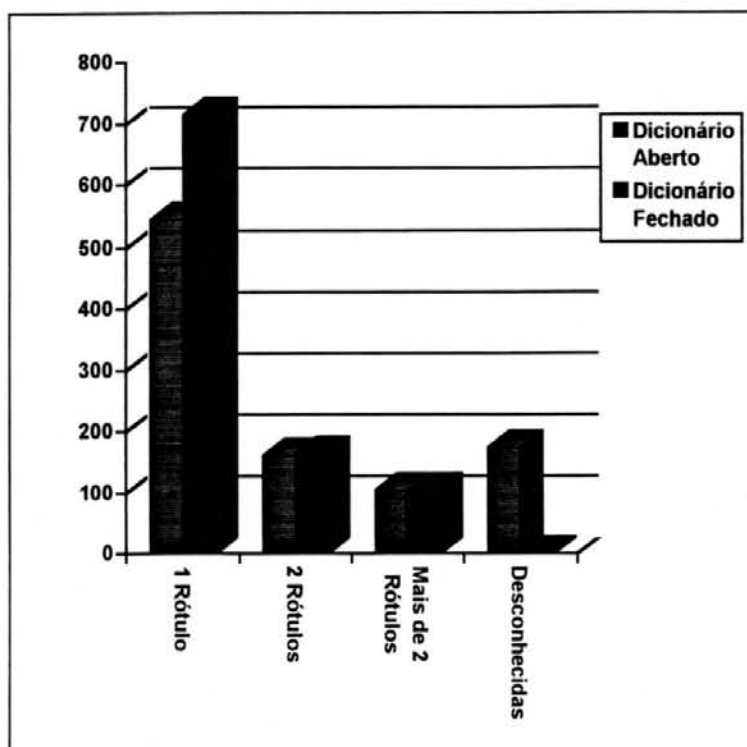


Figura 5.10 - Classes de Ambigüidade no “Corpus” Acentuado - Dicionário Aberto x Dicionário Fechado

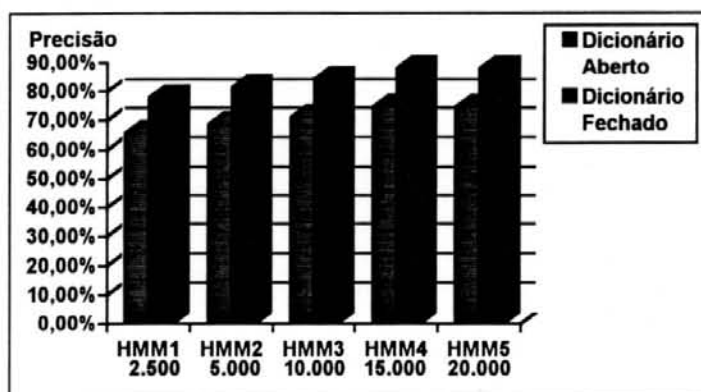


Figura 5.11 - Precisão no Corpus Acentuado - Dicionário Aberto x Dicionário Fechado

5.4.2 Corpus de Treinamento Desacentuado

Neste teste, usa-se o “**corpus**” de treinamento desacentuado, com 20.000 palavras marcadas, para gerar os cinco “corpora” de treinamento (2.500, 5.000, 10.000, 15.000 e 20.000 palavras). Cada um destes “corpora” é utilizado para treinar um HMM:

- o **HMM6** é treinado com **2.500** palavras;
- o **HMM7** é treinado com **5.000** palavras;
- o **HMM8** é treinado utilizando **10.000** palavras;
- o **HMM9** **15.000** palavras e
- o **HMM10** **20.000**.

5.4.2.1 Usando o Dicionário Aberto

O módulo classificador utiliza o **dicionário aberto** para processar o “corpus” de teste. O resultado deste processamento é enviado ao módulo de Viterbi, que utiliza cada um dos HMMs para fazer a marcação deste “corpus” de teste. Os resultados são apresentados na tabela 5.7.

Tabela 5.7 - Resultados no Corpus Desacentuado, Dicionário Aberto

| HMM | Corpus de Treinamento | Rótulos Corretos | Rótulos Errados | Total de Rótulos | Precisão |
|-------|-----------------------|------------------|-----------------|------------------|----------|
| HMM6 | 2.500 | 638 | 344 | 982 | 64,97% |
| HMM7 | 5.000 | 667 | 315 | 982 | 67,92% |
| HMM8 | 10.000 | 692 | 290 | 982 | 70,47% |
| HMM9 | 15.000 | 727 | 255 | 982 | 74,03% |
| HMM10 | 20.000 | 729 | 253 | 982 | 74,24% |

A precisão obtida pelo HMM6, treinado com 2.500 palavras, tem uma diferença de 9,27% com a obtida pelo HMM10, treinado com 20.000. A figura 5.12 apresenta um gráfico no qual se pode ver como a precisão aumenta, à medida que se aumenta o tamanho do “corpus” de treinamento.

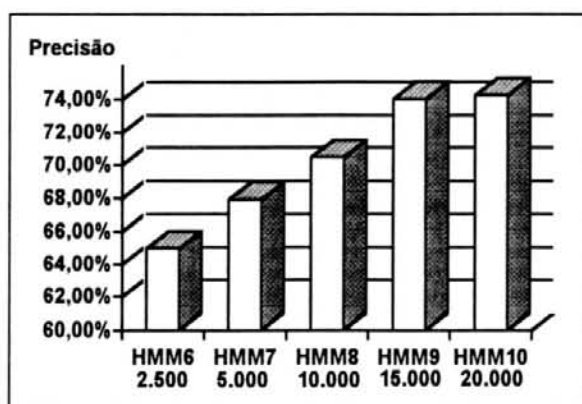


Figura 5.12 - Dicionário Aberto: Tamanho do Corpus de Treinamento Desacentuado x Precisão

5.4.2.2 O Corpus Desacentuado com o Dicionário Aberto

O módulo classificador utiliza o dicionário aberto para processar o “corpus” de teste desacentuado. São encontradas 298 palavras ambíguas (figura 5.13), pertencentes a 43 classes de ambigüidade (tabela 5.8).

Tabela 5.8 - Dicionário Aberto: Classes de Ambigüidade do Corpus Desacentuado

| Classes de Ambigüidade | Frequência |
|-------------------------------|-------------------|
| ADJ - ADV - N | 1 |
| ADJ - CONJ - N - ORD - PREP | 2 |
| ADJ - N | 16 |
| ADJ - NP | 2 |
| ADJ - VPP | 2 |
| ADJ - VTD | 1 |
| ADJ - VTI | 1 |
| ADV - ART - CONT - PREP | 22 |
| ADV - CONJ | 3 |
| ADV - CONJ - PREP | 2 |
| ADV - N | 3 |
| ADV - PREP | 6 |
| AF - N | 1 |
| ART - CARD | 15 |
| ART - CONT | 3 |
| ART - PD - PPOA | 33 |
| ART - PPOA | 4 |
| ART - PREP | 3 |
| CONJ - N - PPOA | 2 |
| CONJ - PREP | 2 |
| CONJ - PREP - VTD | 15 |
| CONJ - PR - PREP | 18 |
| CONJ - VSER | 22 |
| CONT - PD | 1 |
| CONT - VTD | 9 |
| NP - VSER | 7 |
| N - NP | 14 |
| N - PREP | 64 |
| N - VAUX | 2 |
| N - VSER | 1 |
| N - VTD - VTI | 1 |
| N - VTI | 1 |
| PD - VAUX - VTI | 1 |
| VAUX - VPP - VSER | 1 |
| VAUX - VSER | 1 |
| VAUX - VTD | 2 |
| VAUX - VTD - VTI | 5 |
| VAUX - VTI | 2 |
| VI - VTD | 2 |
| VI - VTD - VTI | 1 |
| VPP - VTI | 1 |
| VTD - VTER | 1 |
| VTD - VTI | 2 |

Das palavras do “corpus”, 298 são ambíguas e 173 são desconhecidas, representando, respectivamente, 30,35% e 17,62% do total de palavras do “corpus”. São 194 as palavras que têm dois rótulos (19,76%) e 104 as que têm mais de dois rótulos (10,59%), figura 5.14. Por se tratar do mesmo “corpus” de teste, as palavras desconhecidas são as mesmas, tanto no “corpus” acentuado quanto no desacentuado, como pode ser visto na seção 5.4.1.2. Assim, das 173 palavras desconhecidas, 8 são advérbios (4,62%), 17 são substantivos próprios (9,83%), 23 são adjetivos (13,30%), 56 são verbos (32,37%) e 69 são substantivos (39,88%), (figura 5.15).

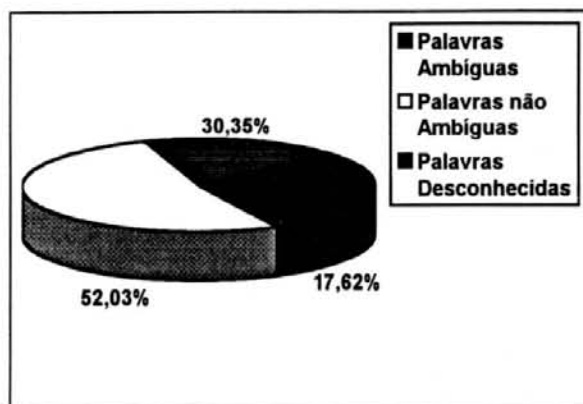


Figura 5.13 - Dicionário Aberto: Palavras Ambíguas e Desconhecidas no Corpus Desacentuado

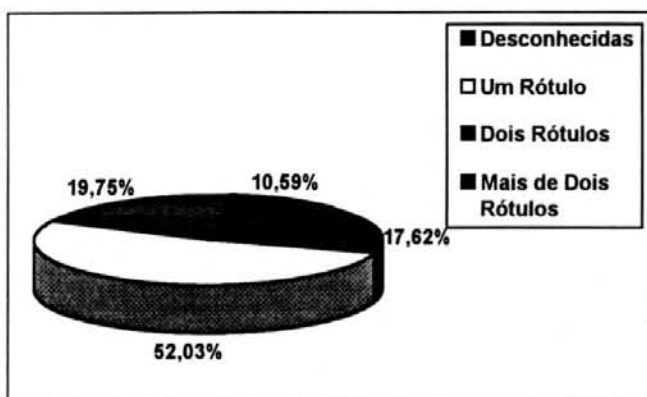


Figura 5.14 - Dicionário Aberto: Rótulos por Palavra no Corpus de Teste Desacentuado

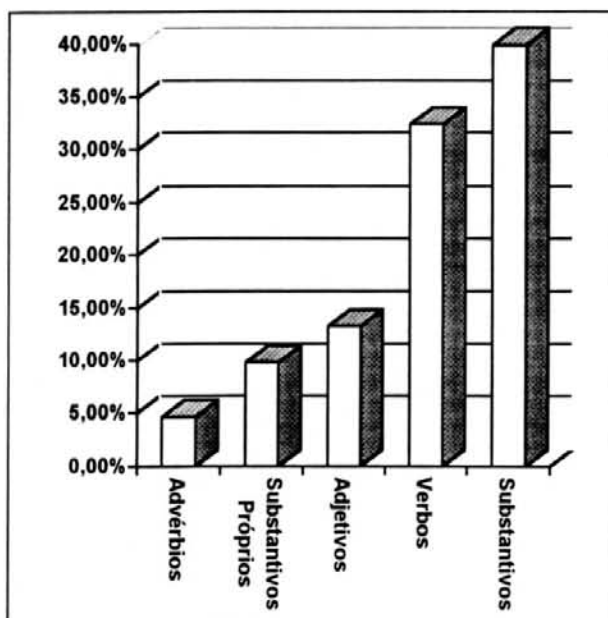


Figura 5.15 - Dicionário Aberto: Palavras Desconhecidas no Corpus Desacentuado

5.4.2.3 Usando o Dicionário Fechado

Após o classificador utilizar o **dicionário fechado** para processar o “corpus” de teste desacentuado, cada um dos HMMs é usado para marcá-lo. Os resultados obtidos podem ser vistos na tabela 5.9.

Tabela 5.9 - Resultados no Corpus Desacentuado, Dicionário Fechado

| HMM | Corpus de Treinamento | Rótulos Corretos | Rótulos Errados | Total de Rótulos | Precisão |
|-------|-----------------------|------------------|-----------------|------------------|----------|
| HMM6 | 2.500 | 768 | 214 | 982 | 78,21% |
| HMM7 | 5.000 | 799 | 183 | 982 | 81,36% |
| HMM8 | 10.000 | 834 | 148 | 982 | 84,93% |
| HMM9 | 15.000 | 864 | 118 | 982 | 87,98% |
| HMM10 | 20.000 | 866 | 116 | 982 | 88,19% |

A precisão obtida por cada um dos HMMs pode ser vista no gráfico da figura 5.16.

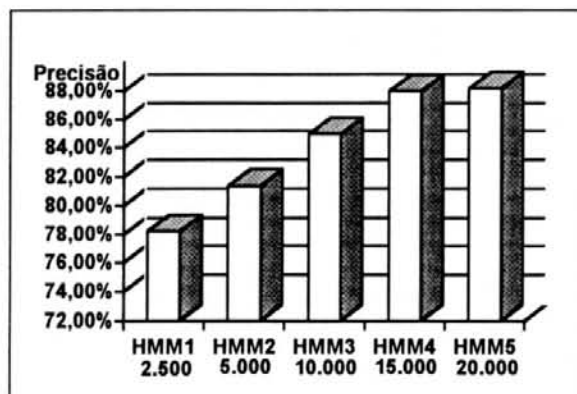


Figura 5.16 - Dicionário Fechado: Tamanho do Corpus de Treinamento x Precisão

Como se pode ver, há um aumento bastante significativo da precisão, de 9,98% - do HMM6 com 78,21% ao HMM10 com 88,19%. O maior aumento da precisão ocorre entre o HMM7 e o HMM8, onde o crescimento é de 3,57%. Este aumento apresenta uma tendência para a estabilização e entre o HMM9 e o HMM10 há um crescimento de apenas 0,21% na precisão.

5.4.2.4 O Corpus Desacentuado com o Dicionário Fechado

O mesmo “corpus” de teste desacentuado foi processado pelo classificador, só que desta vez foi utilizado o dicionário fechado. Há 313 palavras ambíguas (figura 5.17) que pertencem à 44 classes de ambigüidade (tabela 5.10), representando 31,87 % das palavras do “corpus” (figura 5.17). Das 313 palavras ambíguas, 208 (21,18%) têm apenas dois rótulos e 105 (10,69%) têm mais de dois rótulos (figura 5.18).

Tabela 5.10 - Dicionário Fechado: Classes de Ambigüidade do Corpus Desacentuado

| Classes de Ambigüidade | Frequência |
|-------------------------------|-------------------|
| ADJ - ADV - N | 1 |
| ADJ - CONJ - N - ORD - PREP | 2 |
| ADJ - N | 17 |
| ADJ - NP | 2 |
| ADJ - VPP | 2 |
| ADJ - VTD | 1 |
| ADJ - VTI | 1 |
| ADV - ART - CONT - PREP | 22 |
| ADV - CONJ | 3 |
| ADV - CONJ - PREP | 2 |
| ADV - N | 3 |
| ADV - PREP | 6 |
| AF - N | 1 |
| ART - CARD | 15 |
| ART - CONT | 3 |
| ART - PD - PPOA | 33 |
| ART - PPOA | 4 |
| ART - PREP | 3 |
| CONJ - N - PPOA | 2 |
| CONJ - PREP | 2 |
| CONJ - PREP - VTD | 15 |
| CONJ - PR - PREP | 18 |
| CONJ - VSER | 22 |
| CONT - PD | 1 |
| CONT - VTD | 9 |
| NP - VSER | 7 |
| N - NP | 14 |
| N - PREP | 64 |
| N - VAUX | 2 |
| N - VSER | 1 |
| N -VTD -VTI | 1 |
| N - VTI | 1 |
| PD - VAUX - VTI | 1 |
| VAUX -VPP - VSER | 1 |
| VAUX - VSER | 1 |
| VAUX - VTD | 2 |
| VAUX - VTD - VTI | 6 |
| VAUX - VTI | 2 |
| VI - VTD | 2 |
| VI - VTD - VTI | 1 |
| VPP - VTD | 1 |
| VPP - VTI | 5 |
| VTD - VTER | 1 |
| VTD - VTI | 10 |

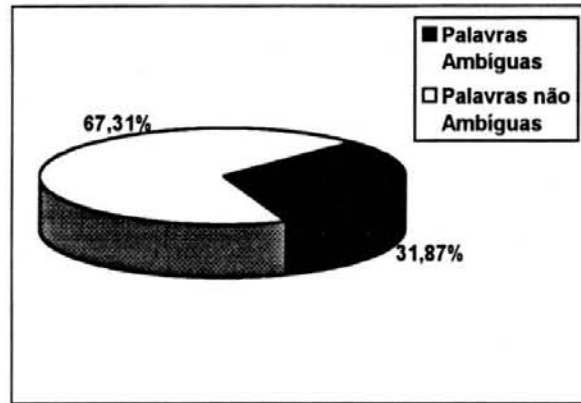


Figura 5.17 - Dicionário Fechado: Palavras Ambíguas no Corpus Desacentuado

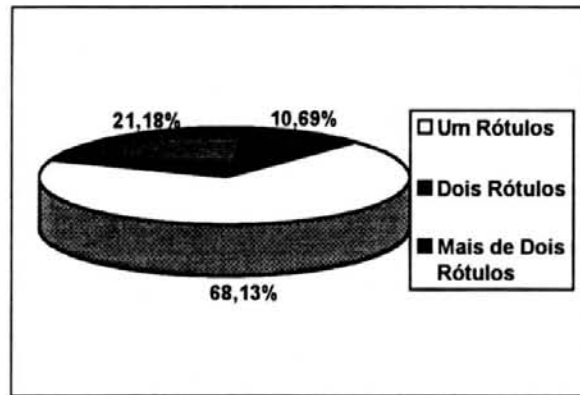


Figura 5.18 - Dicionário Fechado: Rótulos por Palavra no Corpus de Teste Desacentuado

5.4.3 Dicionário Aberto x Dicionário Fechado

Comparando-se os resultados obtidos com o “corpus” de teste processado com o dicionário aberto e o processado com o dicionário fechado, a precisão do primeiro fica entre 64,97% (HMM6) e 74,24% (HMM10) e a do segundo, entre 78,21% (HMM6) e 88,19% (HMM10), com uma diferença de 13,24% no limite inferior (HMM6) e 13,95% no superior (HMM10) (figura 5.20). Com relação às palavras ambíguas, pode-se ver que utilizando o dicionário fechado, não há um aumento muito grande no seu número em relação ao dicionário aberto - 15 palavras. É com as palavras com um único rótulo que ocorre o maior aumento: 511 com o dicionário aberto e 669 com o dicionário fechado (tabela 5.11). Da mesma forma que com o “corpus” acentuado, a maioria das palavras desconhecidas encontradas com o dicionário aberto se enquadram entre as palavras com um único rótulo do dicionário fechado (figura 5.19).

Tabela 5.11 - Corpus Desacentuado: Dicionário Aberto x Dicionário Fechado

| | Dicionário Aberto | | | Dicionário Fechado | | |
|--------------------------------|-------------------|-----------------|----------|--------------------|-----------------|----------|
| | Rótulos Corretos | Rótulos Errados | Precisão | Rótulos Corretos | Rótulos Errados | Precisão |
| HMM6 | 638 | 344 | 64,97% | 768 | 214 | 78,21% |
| HMM7 | 667 | 315 | 67,92% | 799 | 183 | 81,36% |
| HMM8 | 692 | 290 | 70,47% | 834 | 148 | 84,93% |
| HMM9 | 727 | 255 | 74,03% | 864 | 118 | 87,98% |
| HMM10 | 729 | 253 | 74,24% | 866 | 116 | 88,19% |
| Palavras Desconhecidas | 173 | | | - | | |
| Palavras com 1 Rótulo | 511 | | | 669 | | |
| Palavras Ambíguas | 298 | | | 313 | | |
| Palavras com 2 Rótulos | 194 | | | 208 | | |
| Palavras com mais de 2 Rótulos | 104 | | | 105 | | |
| Total | 982 | | | 982 | | |

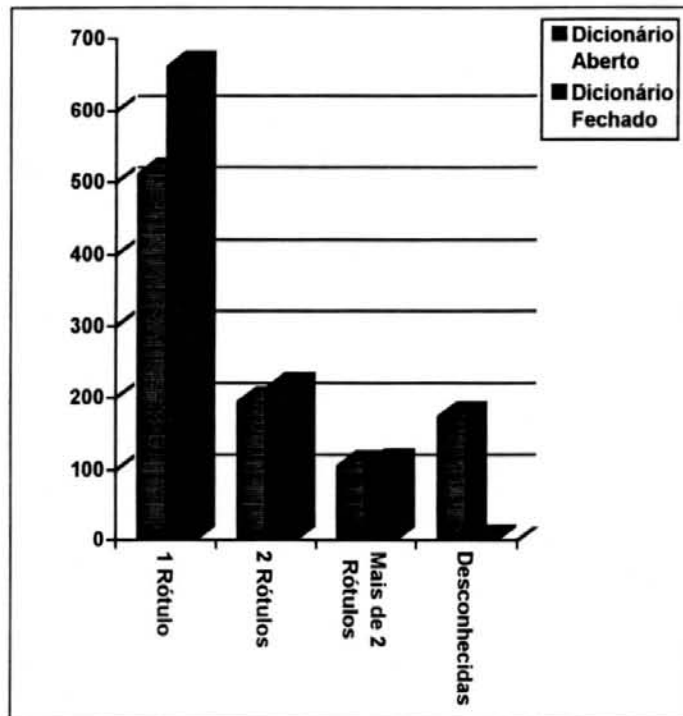


Figura 5.19 - Classes de Ambigüidade - Dicionário Aberto x Dicionário Fechado

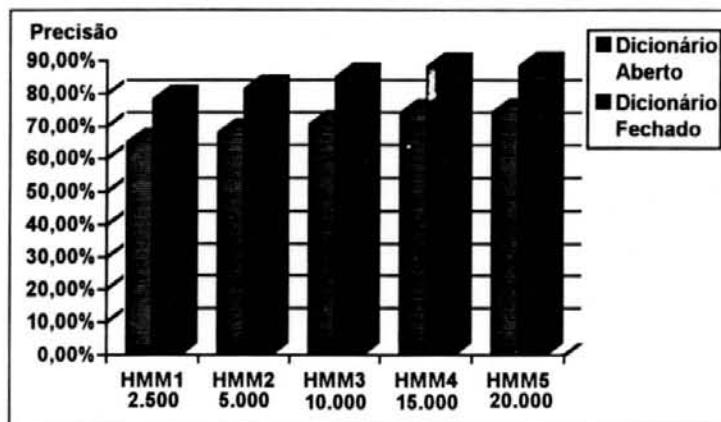


Figura 5.20 - Corpus Acentuado - Dicionário Aberto x Dicionário Fechado

5.4.4 Corpus Acentuado x Corpus Desacentuado

Os resultados obtidos com um “corpus” de treinamento acentuado e um desacentuado podem ser vistos na tabela 5.12.

Analisando-se a tabela com relação à precisão, pode-se perceber que, à medida que o tamanho do “corpus” de treinamento vai aumentando, a variação da precisão diminui (figura 5.21). Há um crescimento bastante grande no início, com o “corpus” de 2.500 palavras e que vai se estabilizando à medida que o tamanho do “corpus” atinge as 20.000 palavras (figura 5.22). Este comportamento é observado tanto com o uso do “corpus” acentuado quanto com o desacentuado.

E, como é de se esperar, o número de palavras ambíguas no “corpus” desacentuado é bem maior que no “corpus” acentuado, sendo que a maior parte se concentra nas palavras que têm apenas dois rótulos (figura 5.23). Todavia, apesar de haver menos palavras ambíguas no “corpus” acentuado, não se observa nenhum aumento na precisão (tabela 5.12 e figura 5.21). Com isto pode-se concluir que o número de palavras ambíguas existentes não influi na precisão do sistema rotulador, dado.

Tabela 5.12 - Corpus Acentuado x Corpus Desacentuado

| | Corpus Acentuado | | | Corpus Desacentuado | | |
|--------------------------------|------------------|-------------------|--------------------|---------------------|-------------------|--------------------|
| | HMM | Dicionário Aberto | Dicionário Fechado | HMM | Dicionário Aberto | Dicionário Fechado |
| | HMM1 | 65,59% | 77,81% | HMM6 | 64,97% | 78,21% |
| | HMM2 | 68,43% | 81,22% | HMM7 | 67,92% | 81,36% |
| | HMM3 | 70,98% | 84,21% | HMM8 | 70,47% | 84,93% |
| | HMM4 | 74,44% | 87,88% | HMM9 | 74,03% | 87,98% |
| | HMM5 | 74,54% | 87,98% | MM10 | 74,24% | 88,19% |
| Palavras Desconhecidas | | 173 | - | | 173 | - |
| Palavras com 1 Rótulo | | 545 | 715 | | 511 | 669 |
| Classes de Ambigüidade | | 35 | 37 | | 43 | 44 |
| Palavras Ambíguas | | 264 | 267 | | 298 | 313 |
| Palavras com 2 Rótulos | | 160 | 163 | | 194 | 208 |
| Palavras com mais de 2 Rótulos | | 104 | 104 | | 104 | 105 |
| Total | | 982 | 982 | | 982 | 982 |

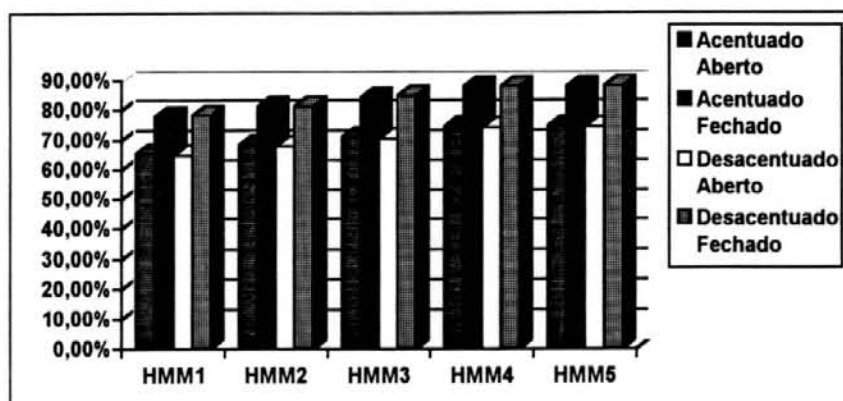


Figura 5.21 - Precisão: Corpus Acentuado x Corpus Desacentuado

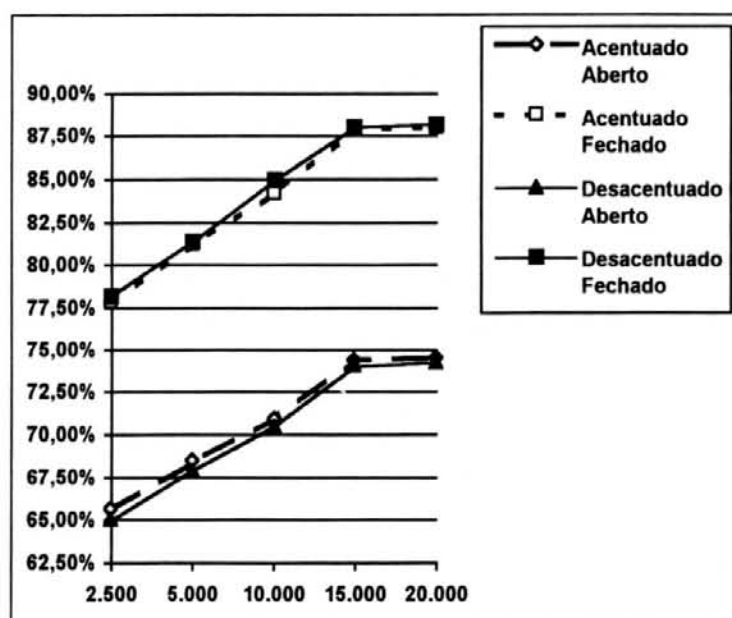


Figura 5.22 - Variação da Precisão

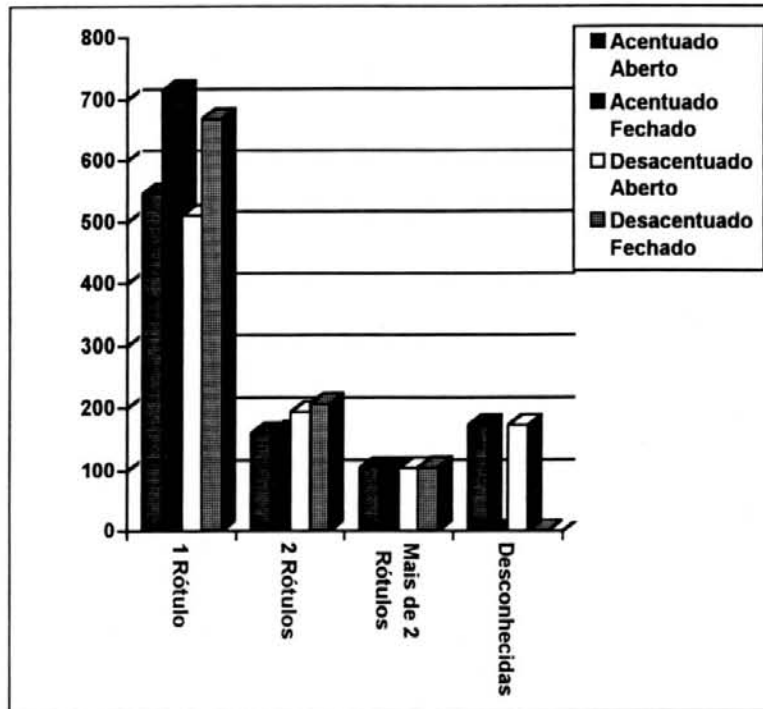


Figura 5.23 - Classes de Ambigüidade: Corpus Acentuado x Corpus Desacentuado

5.4.5 Conclusões a Respeito dos Resultados

Como já foi visto no capítulo 4, o “corpus” de treinamento disponível para este trabalho tem apenas 20.000 palavras marcadas. Diante de um “corpus” tão pequeno quando comparado ao “Brown Corpus” ou ao “PennTrebek”, com seus milhões de palavras, surgem diversas questões como: “Será que com um “corpus” de treinamento de apenas 20.000 palavras é possível obter uma boa precisão? Qual é a relação entre o tamanho do “corpus” de treinamento e a precisão?”. Procurando na literatura, nenhuma resposta concreta foi encontrada.

Analisando-se os testes realizados pode-se ver que realmente o tamanho do “corpus” de treinamento tem influência sobre a precisão obtida. Quanto maior o “corpus” utilizado maior a precisão. Assim, neste experimento, a maior precisão é obtida usando os “corpora” de treinamento com 20.000 palavras. Todavia, pode-se notar também, que, apesar do grande crescimento inicial, esta precisão começa a se estabilizar, figura 5.22, e entre os “corpora” de 15.000 palavras e os de 20.000 não há um aumento tão grande na precisão. Isto ocorre apesar de se utilizar 5.000 palavras a mais, com um aumento de 33% no tamanho do “corpus” de 15.000 palavras.

Apesar de se utilizar um “corpus” com apenas 20.000 palavras, obtém-se resultados bastante satisfatórios, com a precisão do HMM5 e do HMM10 variando entre 74,24% e 88,19%. Estes resultados sugerem que se está bastante próximo do tamanho ideal, onde mesmo utilizando um “corpus” maior, a precisão não se altera muito. Este é um resultado bastante animador, pois, certamente, um “corpus” marcado desta magnitude é bem mais acessível para a maioria das línguas do que um com 1.000.000 de palavras.

Comparando-se, agora, os testes realizados com um “corpus” acentuado e um desacentuado pode-se perceber que, ao contrário do que se poderia imaginar, o número de palavras ambíguas não influi na precisão do sistema rotulador, figura 5.21. A diferença do número de palavras ambíguas no “corpus” acentuado fechado e no “corpus” desacentuado fechado é de 46 palavras. Ou seja, a falta de uma representação para os acentos resulta em um aumento no número de palavras ambíguas. No entanto, apesar deste aumento no número de palavras ambíguas, a diferença de precisão

observada é tão pequena que não representa uma variação significativa. Deste resultado, conclui-se que o tratamento dado pelo sistema às palavras ambíguas é bastante eficiente e, que a ambigüidade existente, pode ser resolvida com informações sobre o contexto mais próximo. Porém, observa-se que os resultados obtidos se devem em parte ao fato de o “corpus” acentuado utilizado ter somente acentuação parcial. Talvez, com um “corpus” totalmente acentuado os resultados possam ser diferentes.

Portanto, o uso de um “corpus” desacentuado, para realizar o treinamento deste sistema, não compromete os resultados obtidos, visto que não acarreta variações significativas na sua precisão, dado o nível de acentuação do “corpus” acentuado.

Maior cuidado deve ser dispensado às palavras desconhecidas. Como se pode observar pelos resultados obtidos, elas representam um fator muito forte na precisão total obtida pelo rotulador. Em conseqüência, deve-se tentar construir mecanismos que possam tratar deste tipo de palavras, pois, do contrário, pode-se obter uma precisão bastante baixa. A existência de uma palavra desconhecida não apenas dificulta a sua marcação, mas também a marcação das palavras vizinhas, o que afeta significativamente a precisão do sistema.

A existência de um tratamento eficiente para estas palavras é fundamental para que se obtenha uma boa precisão. Neste trabalho, está-se usando um método bastante simples para o seu tratamento, que especifica que uma palavra desconhecida será rotulada com uma das possíveis categorias, de acordo com o contexto no qual está inserida. Porém, como se pode ver, a diferença entre a precisão obtida com o dicionário aberto e a obtida com o dicionário fechado é bastante grande. Isto indica a necessidade de um método mais elaborado para o tratamento das palavras desconhecidas.

Alguns mecanismos foram propostos na literatura, como o usado em Cutting [CUT92], que é dividido em três estágios:

- no primeiro estágio, pesquisa-se por radicais/temas e categorias morfo-sintáticas associadas, em um dicionário manualmente construído;

- o segundo estágio realiza a análise dos sufixos das palavras que não foram encontradas neste dicionário;
- se a palavra não está no dicionário manualmente construído e o seu sufixo não foi reconhecido, usa-se, então, uma classe de ambigüidade “default”, que contém todas as categorias abertas da língua.

Schmid [SCH94a] utiliza um método semelhante ao de Cutting, com os três estágios, onde, no segundo, utiliza-se um dicionário de sufixos construído automaticamente a partir do “corpus” marcado.

Brill [BRI92b], na primeira versão de seu sistema, atribui, para as palavras desconhecidas, o rótulo de substantivo próprio se começarem por uma letra maiúscula e de substantivo no restante dos casos. Já na segunda versão do sistema, faz uma compilação das terminações mais comuns para cada categoria morfo-sintática e atribui o rótulo correspondente para a palavra desconhecida.

Com base nos experimentos realizados, pode-se afirmar que, o comportamento que resultou na maior precisão foi com a utilização do “corpus” de treinamento de 20.000 palavras e com o dicionário fechado. Assim, após a análise destes aspectos, pode-se ver que, apesar de não se dispor de um “corpus” marcado com 1.000.000 de palavras, se pode obter bons resultados usando um “corpus” com apenas 20.000. Por fim, reforça-se, ainda, a necessidade da existência, no sistema rotulador, de mecanismos para o tratamento de palavras desconhecidas.

6. CONCLUSÕES E TRABALHOS FUTUROS

O mundo inteiro vem presenciando a expansão da era tecnológica. Nesta era, a informação, escrita e falada, trafega livremente, através de meios como o computador. Devido a esta abundância de informações, surge a necessidade de ferramentas que possam lidar com elas. Para trabalhar com a informação na forma escrita, estas ferramentas devem apresentar uma característica indispensável: a capacidade de trabalhar com textos de domínio irrestrito. Para incorporar esta característica, utilizam-se métodos estatísticos. A aplicação destes métodos, ao Processamento de Linguagem Natural, tem obtido resultados bastante impressionantes. Várias áreas têm sido beneficiadas com isto, como a da tradução automática e a da psicolingüística. Uma área que tem se destacado bastante é a da aplicação de métodos estatísticos na marcação automática de textos usando categorias morfo-sintáticas. Para executar esta tarefa, são construídos sistemas conhecidos como **Rotuladores de Categorias Morfo-Sintáticas** (do inglês: "Part-of-Speech Taggers").

Neste trabalho, foi apresentado um rotulador de categorias morfo-sintáticas para a Língua Portuguesa. Este sistema é capaz de fazer a análise de textos irrestritos, utilizando métodos estatísticos. Apresenta uma característica muito importante que é a capacidade da aquisição automática do conhecimento, a partir de um "corpus": o sistema analisa um "corpus", descobre quais são os padrões lingüísticos que ocorrem nele e os modela automaticamente em um "Hidden Markov Model". Após ter sido feita a aquisição do conhecimento, este sistema está apto a analisar outros "corpora" e reconhecer neles os padrões lingüísticos adquiridos. Ele também é capaz de tratar de palavras desconhecidas e de palavras ambíguas. A entrada deste rotulador são as palavras de um "corpus". Retornando, como saída, os rótulos correspondentes a cada uma das palavras.

Um sistema assim pode ser utilizado nas mais diversas aplicações. Entre elas, uma aplicação que tem se mostrado bastante importante é no processo de análise sintática (parsing): o rotulador é usado como um pré-analisador. Ele é usado para eliminar, ou pelo menos reduzir substancialmente, a ambigüidade das palavras, antes de

se fazer a análise sintática propriamente dita. Assim, pode-se aliviar parte do trabalho do analisador sintático.

A principal contribuição deste trabalho foi realizar a avaliação de aspectos importantes do sistema rotulador, de forma a encontrar um padrão de comportamento, no qual a precisão seja máxima. Para tanto, foram definidas algumas variáveis, sobre as quais se centrou esta avaliação. A primeira variável definida foi a influência das palavras ambíguas. Para testar este ponto se verificou se a precisão do sistema aumentava ou não com o uso de um “corpus” acentuado, onde há um menor número de palavras ambíguas. A segunda variável diz respeito ao tamanho ideal do “corpus” de treinamento: até que ponto o aumento do tamanho do “corpus” de treinamento implica um aumento da precisão? E a terceira variável foi a influência das palavras desconhecidas na precisão do sistema rotulador.

A precisão máxima do sistema foi obtida utilizando um “corpus” de treinamento com 20.000 palavras e o dicionário fechado. Estes resultados a que se chegou com esta avaliação indicam que não é necessário dispor de um “corpus” marcado com 1.000.000 de palavras para obter bons resultados. Com um “corpus” de 20.000 palavras já se pode obter uma precisão bastante satisfatória (a precisão varia de 64,97% a 88,19%). Além disto, determinou-se que o sistema é bastante eficiente no tratamento de palavras ambíguas. Portanto, a utilização de um “corpus” de treinamento acentuado não está relacionada com a precisão do sistema, sendo considerada como ponto supérfluo. Os resultados também apontam a necessidade da construção de mecanismos mais elaborados de tratamento de palavras desconhecidas para a obtenção de melhores resultados.

Outra contribuição importante deste trabalho, foi a construção de um “corpus” marcado para o Português. São poucas as línguas, além do Inglês, que possuem este tipo de “corpus”. E, como já mencionado no capítulo 2, são inúmeras as aplicações existentes para semelhante recurso. Além de ser necessário para a realização de estudos nesta área, um “corpus” marcado pode ser aplicado, por exemplo, para facilitar o ensino da morfologia de uma língua [GÜV94]. Outra aplicação é o estudo da associação existente entre as palavras. Estas associações podem ocorrer por vários

motivos, como, por exemplo, devido às restrições léxico-sintáticas ou ainda, a nível de relações semânticas [CHU90].

Este foi um trabalho pioneiro para a Língua Portuguesa, visto que, até o momento nenhum estudo deste tipo havia sido feito. Pretende-se que, com os estudos feitos, se possa ter contribuído para o desenvolvimento e aprimoramento dos trabalhos nesta área, que, apesar de serem recentes, têm apresentado resultados bastante promissores.

6.1 Trabalhos Futuros

Este foi apenas o primeiro passo de um estudo ao qual se pretende adicionar, futuramente, resultados de novas avaliações, de diferentes aspectos dos sistemas rotuladores. O próximo passo a ser feito é a adição de mais informações aos rótulos morfo-sintáticos, tais como gênero e número. Com isto, se quer testar se um aumento na qualidade das informações contribui para melhorar a precisão e o desempenho do sistema rotulador.

Além disto, pretende-se, gradativamente, ampliar o tamanho do Radiobras Corpus. Assim, com um “corpus” maior a disposição, novos estudos poderão ser implementados.

Há, ainda, um sistema rotulador de categorias morfo-sintáticas baseado em árvores de decisão [SCH94b], que está sendo treinado para o Português. Se pretende realizar estudos comparativos entre este rotulador e o rotulador estatístico, de forma a descobrir as vantagens de cada um frente à Língua Portuguesa.

Pretende-se, ainda, utilizar os resultados obtidos por este trabalho e ampliá-los, de modo a construir uma gramática probabilística para o Português. Este trabalho dará prosseguimento a trabalhos como o de [JOS94], [BOD93], [BLA93] e [STK94b], onde um “corpus” é utilizado como uma gramática estocástica e, associada a cada palavra deste “corpus”, se tem uma estrutura sintática.

7. BIBLIOGRAFIA

- [BLA 93] BLACK, E. et al. Towards History-based Grammars: Using Richer Models for Probabilistic Parsing. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ANNUAL MEETING (ACL-93), 31., 1993, Columbus. **Proceedings ...** Cambridge:MIT Press, 1993.
- [BOD 93] BOD, R. Using an Annotated Corpus as a Stochastic Grammar. In: EUROPEAN CHAPTER CONFERENCE OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (EACL-93), 6., 1993, Utrecht. **Proceedings ...** Cambridge:MIT Press, 1993.
- [BRI 90] BRILL, E. et al. Deducing linguistic structure from the statistics of large corpora. In: DARPA SPEECH AND NATURAL LANGUAGE WORKSHOP, 1990, Hidden Valley, PA. **Proceedings ...** [S.l.:s.n.], 1990. p. 275-282.
- [BRI 92] BRILL, E.; MARCUS, M. Tagging an Unfamiliar Text with Minimal Human Supervision. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, AAAI Workshop, 1992. **Proceedings ...** [S.l.:s.n.], 1992.
- [BRI 92a] BRILL, E. A Simple Rule-Based Part of Speech Tagger. In: DARPA SPEECH AND NATURAL LANGUAGE WORKSHOP, 1992. **Proceedings ...** [S.l.:s.n.], 1992. p. 112-116.
- [BRI 92b] BRILL, E.; MARCUS, M. Automatically Acquiring Phrase Structure Using Distributional Analysis. In: DARPA SPEECH AND NATURAL LANGUAGE WORKSHOP, 1992. **Proceedings ...** [S.l.:s.n.], 1992.

- [BRI 93] BRILL, E. Automatic Grammar Induction and Parsing Free-Text: A Transformation-Based Approach. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ANNUAL MEETING (ACL-93), 31., 1993, Columbus. **Proceedings ...** Cambridge:MIT Press, 1993.
- [BRI 93a] BRILL, E. **A Corpus-Based Approach to Language Learning**. Philadelphia: University of Pennsylvania, 1993. PhD Dissertation.
- [BRI 94] BRILL, E. Some Advances in Transformation-Based Part-of-Speech Tagging. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI-94), 12., 1994, Seattle. **Proceedings ...** [S.l.:s.n.], 1994.
- [BRS 94] BRISCOE, E.J.B. Prospects for Practical Parsing of Unrestricted Text: Robust Statistical Parsing Techniques. In: EUROPEAN SUMMERSCHOOL IN LOGIC, LANGUAGE AND INFORMATION (ESSLLI'94), 1994, Copenhagen. Advanced Course CA4.
- [CHA 93] CHANG, C.H.; CHEN, C.D. HMM-based Part-of-Speech Tagging for Chinese Corpora. In: WORKSHOP ON VERY LARGE CORPORA: ACADEMIC AND INDUSTRIAL PERSPECTIVES, 1993. **Proceedings ...** [S.l.:s.n.], 1993. p. 40-47.
- [CHD 95] CHANOD, J. P.; TAPANAINEN, P. **Creating a tagset, lexicon and guesser for a French tagger**. 1995. (Disponível via WWW em <http://xxx.lanl.gov/cmp-lg/>).
- [CHK 93] CHARNIAK, E. **Statistical Language Learning**. Cambridge: Bradford Book, 1993.

- [CHE 93] CHEN, K.H.; CHEN, H.H. Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and Its Automatic Evaluation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ANNUAL MEETING (ACL-94), 32., 1994, Las Cruces. **Proceedings ...** Cambridge:MIT Press, 1994. p. 234-240.
- [CHN 93] CHEN, H.H.; CHEN, Y.S. Approximate N-Gram Markov Model for Natural Language Generation. In: QUALICO, 1994. **Proceedings ...** [S.l.:s.n.], 1994.
- [CHU 88] CHURCH, K. W. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In: CONFERENCE ON APPLIED NATURAL LANGUAGE PROCESSING (ANLP-88), 2., 1988, Austin. **Proceedings ...** Cambridge:MIT Press, 1988. p. 136-143.
- [CHU 90] CHURCH, K. W.; HANKS, P. Word Association Norms, Mutual Information and Lexicography **Computational Linguistics**, Cambridge, v. 16, n.1, p 22-29, 1990.
- [CHU 91] CHURCH, K. W.; GALE, W.A. A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probabilities of English Bigrams. **Computer Speech and Language**, n. 5, p. 19-54, 1991.
- [CHU 93] CHURCH, K. W.; MERCER, R.L.; Introduction to the Special Issue on Computational Linguistics Using Large Corpora. **Computational Linguistics**, Cambridge, v. 19, n. 1, p. 1-24, 1993.
- [CLO 93] CLOEREN, J. Towards a Cross-Linguistic Tagset. In: WORKSHOP ON VERY LARGE CORPORA: ACADEMIC AND INDUSTRIAL PERSPECTIVES, 1993, Ohio State University. **Proceedings ...** [S.l.:s.n.], 1993. p. 30-39.

- [CUT 92] CUTTING, D. et al. A practical part-of-speech tagger. In: CONFERENCE ON APPLIED NATURAL LANGUAGE PROCESSING (ANLP-92), 3., 1992, Trento. **Proceedings ...** Cambridge:MIT Press, 1992. p. 133-140.
- [CUT 93] CUTTING, D.; PEDERSEN, J. **The Xerox Part-of-Speech Tagger**. [S.l.:s.n.], 1993. p. 1-6. Technical Report.
- [DAG 94] DAGAN, I.; PEREIRA, F.; LEE, L. Similarity-Based Estimation of Word Cooccurrence Probabilities. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ANNUAL MEETING (ACL-94), 32., 1994, Las Cruces. **Proceedings ...** Cambridge:MIT Press, 1994. p. 272-278.
- [DER 88] DEROSE, S. Grammatical category disambiguation by statistical optimization. **Computational Linguistics**, Cambridge, v. 14, n. 1, p. 31-39, 1988.
- [DUN 93] DUNNING, T. Accurate Methods for the Statistics of Surprising and Coincidence. **Computational Linguistics**, Cambridge, v. 19, n.1, p. 61-74, 1993.
- [ELW 94] ELWORTHY, D. Automatic Error Detection in Part of Speech Tagging. In: CONFERENCE ON APPLIED NATURAL LANGUAGE PROCESSING (ANLP-94), 4., 1994, Stuttgart. **Proceedings ...** Cambridge:MIT Press, 1994.
- [ELW 94a] ELWORTHY, D. Does Baum-Welch Re-estimation Help Taggers? In: CONFERENCE ON NEW METHODS IN LANGUAGE PROCESSING (NeMLaP), 1994, Manchester. **Proceedings ...** [S.l.:s.n.], 1994.

- [ELW 95] ELWORTHY, D. Tagset Design and Inflected Languages. In: EUROPEAN CHAPTER CONFERENCE OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (EACL-95), 7., 1995, Dublin. **Proceedings ...** Cambridge:MIT Press, 1995.
- [GÜV 94] GÜVENİR, A.; KEMAL O. **Using a Corpus for Teaching Turkish Morphology**. Ankara: Bilkent University, 1994. Technical Report.
- [HIN 89] HINDLE, D. Acquiring Disambiguation Rules from Text. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ANNUAL MEETING (ACL-89), 27., 1989, Vancouver. **Proceedings ...** Cambridge:MIT Press, 1989. p. 118-125.
- [JEL 80] JELINEK, F.; MERCER, R. L. Interpolated Estimation of Markov Source Parameters from Sparse Data. In: WORKSHOP ON PATTERN RECOGNITION IN PRACTICE, 1980. **Proceedings ...** [S.l.:s.n.], 1980. p. 381-397.
- [JOS 94] JOSHI, A. K.; SRINIVAS, B. Disambiguation of Super Parts of Speech (or Supertags): Almost Parsing. In: COLING, 1994, Kyoto. **Proceedings ...** [S.l.]:H. Karlgren Ed., 1994.
- [KEM 94] KEMPE, A. **A Stochastic Tagger and an Analysis of Tagging Errors**. Stuttgart: University of Stuttgart, 1994. Technical Report.
- [KEM 94a] KEMPE, A. Probabilistic Tagging with Feature Structures. In: COLING 94, 1994, Kyoto. **Proceedings ...** [S.l.:s.n.] H. Karlgren Ed., 1994.
- [MAG 94] MAGERMAN, D. **Natural Language Parsing as Statistical Pattern Recognition**. Stanford: Stanford University, 1994. PhD Dissertation.

- [MAG 95] MAGERMAN, D. **Statistical Language Learning - Review**. 1995. (Disponível via WWW em <http://xxx.lanl.gov/cmp-lg/>).
- [MAR 90] MARCKEN, C.G. de. Parsing the LOB corpus. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ANNUAL MEETING (ACL-90), 1990, Pittsburgh. **Proceedings ...** Cambridge:MIT Press, 1990. p. 243-251.
- [MAQ 94] MARQUES, N. M. C.; LOPES, J. G. P. POLARIS: A Portuguese Lexicon Acquisition and Retrieval Interactive System. In: CONFERENCE ON PRACTICAL APPLICATIONS OF PROLOG, 1994. **Proceedings ...** [S.l.:s.n.], 1994.
- [MAQ 95] MARQUES, N. M. C. **YAHT - Yet Another HMM Tagger-Modelos de Markov Escondidos Aplicados à Classificação de Largos Corpora de Textos**. Monte da Caparica: Universidade Nova de Lisboa, 1995. Technical Report.
- [MER 91] MERIALDO, B. Tagging text with a probabilistic model. In: ICASSP, 1991, Toronto. **IEEE Proceedings ...** [S.l.:s.n.], 1991. p. 809-812.
- [MER 94] MERIALDO, B. Tagging English Text with a Probabilistic Model. **Computational Linguistics**, Cambridge, v. 20 ,n. 2, p. 155-171, 1994.
- [MIL 94] MILLER, S. et al. Hidden Understanding Models of Natural Language. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ANNUAL MEETING (ACL-94), 32., 1994, Las Cruces. **Proceedings ...** Cambridge:MIT Press, 1994. p. 25-32.
- [NAK 89] NAKAMURA, M.; SHIKANO, K. A study of English word prediction based on neural network. In: ICASSP, 1989, Glasgow. **Proceedings ...** [S.l.:s.n.], 1989. p. 731-734.

- [NAK 90] NAKAMURA, M. et al. Neural Network Approach to Word Category Prediction for English Texts. In: COLING-90, 1990, Helsinki University. **Proceedings ...** [S.l.] H. Karlgren Ed., 1990. p.213-218.
- [RAM 94] RAMSHAW, L.; MARCUS, M. **Exploring the Statistical Derivation of Transformational Rule Sequences for Part-of-Speech Tagging.** 1994. (Disponível via WWW em http://xxx.lanl.gov/cmp-lg/cmp-lg_9406011).
- [ROC 95] ROCHE, E.; SCHABES, Y. Deterministic Part-of-Speech Tagging with Finite-State Transducers. **Computational Linguistics**, Cambridge, v. 21, n. 2, p. 227-253, 1995.
- [SCA 92] SHABES Y. Statistical versus Rule-Based Methods for Text Analysis. In: EUROPEAN SUMMER SCHOOL ON LANGUAGE AND SPEECH COMMUNICATION, 1992, University of Utrecht. Tutorial presented at the European Summer School on Language and Speech Communication, 1992.
- [SCH 94] SCHIMID, H. **Part-of-Speech Tagging with Neural Networks.** 1994. (Disponível via WWW em <http://xxx.lanl.gov/cmp-lg/>).
- [SCH 94a] SCHIMID, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: CONFERENCE ON NEW METHODS IN LANGUAGE PROCESSING (NeMLaP), 1994, Manchester. **Proceedings ...** [S.l.:s.n.], 1994.
- [SCH 95] SCHIMID, H. **Improvements in Part-of-Speech Tagging with Application to German.** Stuttgart: Universidade de Stuttgart, 1995. Technical Report.

- [SCZ 94] SCHÜTZE, H.; SINGER, Y. Part-of-Speech Tagging Using a Variable Context Markov Model. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ANNUAL MEETING (ACL-94), 32., 1994, Las Cruces. **Proceedings ...** Cambridge:MIT Press, 1994.
- [SPR 94] SPROAT, R. A Stochastic Finite-State Word-Segmentation Algorithm for Chinese. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ANNUAL MEETING (ACL-94), 32., 1994, Las Cruces. **Proceedings ...** Cambridge:MIT Press, 1994. p. 66-73.
- [STK 94] STOLKE, A.; OMOHUNDRO, S. M. **Best-First Model Merging for Hidden Markov Model Induction**. Berkeley: University of California at Berkeley, ICSI, 1994. Technical Report.
- [STK 94a] STOLKE, A.; SEGAL, J. **Precise n-gram Probabilities from Stochastic Context-free Grammars**. Berkeley: University of California at Berkeley, ICSI, 1994. Technical Report.
- [SU 94] SU, K.Y.; WU, M.W.; CHANG, J.S. A Corpus-based Approach to Automatic Compound Extraction. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ANNUAL MEETING (ACL-94), 32., 1994, Las Cruces. **Proceedings ...** Cambridge:MIT Press, 1994. p. 242-247.
- [TAP 94] TAPANAINEN, P.; JÄRVINEN, T. Syntactic Analysis of Natural Language Using Linguistic Rules and Corpus-Based Patterns. In: COLING, 1994, Kyoto. **Proceedings ...** [S.l.]:H. Karlgren Ed., 1994. v.1, p. 629-634.
- [TAP 94a] TAPANAINEN, P.; VOUTILAINEN, A. Tagging Accurately - Don't Guess if you Know. In: CONFERENCE ON APPLIED NATURAL LANGUAGE PROCESSING (ANLP-94), 4., 1994, Stuttgart. **Proceedings ...** Cambridge:MIT Press, 1994. p. 47-52.

- [VIL 95] VILLAVICENCIO, A. et al. Part-of-Speech Tagging for Portuguese Texts. In: SIMPÓSIO BRASILEIRO DE INTELIGÊNCIA ARTIFICIAL (SBIA95), 1995, Campinas. **Proceedings ...** Berlin:Springer-Verlag, 1994.
- [VOU 95] VOUTILAINEN, A. A syntax-based part-of-speech analyser. In: EUROPEAN CHAPTER CONFERENCE OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (EACL-95), 7., 1995, Dublin. **Proceedings ...** Cambridge:MIT Press, 1995.




*Avaliando um Rotulador Estatístico de Categorias Morfo-Sintáticas para a
Língua Portuguesa*


por

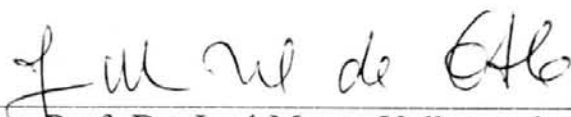
Aline Villavicencio

Dissertação apresentada aos Senhores:

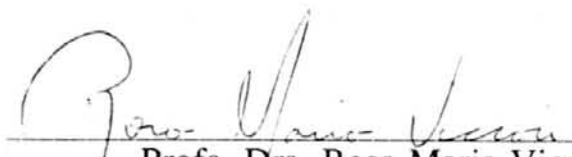

Prof. Dr. José Gabriel Pereira Lopes (UNL/PORTUGAL)

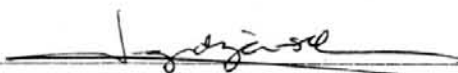

Profa. Dra. Vera Lúcia Strube de Lima (PUCRS)


Prof. Dr. Wilson José Leffa


Prof. Dr. José Mauro Volkmer de Castilho

Vista e permitida a impressão.
Porto Alegre, 9/10/95.


Profa. Dra. Rosa Maria Viccari,
Orientador.


p/ Prof. Flávia Tech Wagner
Coordenador do Curso de Pós Graduação
em Ciência da Computação - CPG
Instituto de Informática - UFRGS