Etiquetagem do Português Clássico Baseada em Corpus

Carlos Daniel Chacur Alves*[1] and Marcelo Finger**[2]

[1] DCC-IME-USP
chacur@ime.usp.br
[2] DCC-IME-USP
mfinger@ime.usp.br

Resumo A construção do Corpus Tycho Brahe do Português histórico etiquetado motivou a criação de ferramentas automáticas para a etiquetagem morfo-sintática de palavras de um texto. Para este fim, selecionamos o algoritmo de etiquetagem de Brill por estar entre os que obtêm os melhores resultados na etiquetagem do Inglês, além de representar o conhecimento explicitamente.
No entanto, mostramos teórica e experimentamente que tal método não é apropriado para línguas com a riqueza morfológica do Português. Propomos uma alteração do método de Brill que impõe uma estrutura ao conjunto de etiquetas utilizado e adiciona uma fase de refinamento ao método original.
Apresentamos medidas e resultados obtidos com a versão atual do Corpus Tycho Brahe e discutimos os diversos problemas encontrados no desenvolvimento de nossa abordagem.

1 Introdução

O Corpus Tycho Brahe está sendo construído como parte de um projeto que pretende estudar a evolução do Português Europeu Clássico (PEC) para o Português Europeu Moderno (PEM), contendo textos de autores nascidos no período que compreende os séculos XVI, XVII, XVIII e XIX. Em particular, o projeto procura estabelecer como a mudança prosódica dos padrões rítmicos do Português levou a mudanças sintáticas, em especial na colocação dos pronomes clíticos. O Corpus Tycho Brahe está sendo construído para levantar dados sobre a parte sintática da evolução do Português Europeu, e pretendemos colocá-lo a disposição na Internet[1].

Esse corpus histórico compreende, atualmente, dois tipos de textos:

* Trabalho financiado pela Fapesp, processo 97/12986-3.
** Trabalho parcialmente financiado pelo CNPq, bolsa 300597/95-9 (PQ) e pelo Projeto Temático Fapesp 98/3382-0 (*Rhythmic Patterns, Parameter Setting and Language Change*).

[1] A escolha do nome Tycho Brahe está ligada a esta divulgação do corpus. Tycho Brahe foi um astrônomo dinamarquês que viveu entre os anos 1546 e 1601, pioneiro

1. Textos originais, ortograficamente transcritos, com informações sobre as edições, autores, datas e comentários dos editores.
2. Os mesmos textos, com cada palavra marcada morfo-sintaticamente (sem as informações extra-lingüísticas mencionadas acima).

No futuro, o corpus deverá incluir também esses textos com a estrutura sintática anotada.

A construção de um corpus cujo número de palavras etiquetadas é da ordem de milhões não seria viável sem o apoio de métodos automáticos para o processamento dos dados. Sendo assim, faz parte do projeto o desenvolvimento de etiquetadores morfo-sintáticos, ferramentas indispensáveis quando se deseja anotar textos razoavelmente grandes.

É preciso ressaltar que estamos construindo um etiquetador morfo-sintático com o claro objetivo de *auxiliar a construção de um corpus*. A metodologia de desenvolvimento do etiquetador segue em paralelo com a do corpus, da seguinte maneira:

1. Primeiramente, os lingüistas encarregados do corpus etiquetaram 21.000 palavras manualmente.
2. Estas palavras serviram para "treinar" o etiquetador uma primeira vez.
3. Novos textos foram trazidos para o formato eletrônico e etiquetados com nosso etiquetador.
4. Tais textos foram corrigidos manualmente pelos lingüistas, estendendo-se o tamanho do corpus.
5. Os novos textos podem ser usados para "treinar" um novo etiquetador, retornando ao passo 2.

Esse fato diferencia o desenvolvimento do nosso etiquetador do desenvolvimento de outros etiquetadores automáticos para o Inglês, que puderam contar com corpora contendo da ordem de 1.000.000 de palavras já etiquetadas para testar e refinar seus algoritmos. Por exemplo, o etiquetador de Charniak [5] foi testado sobre o Brown Corpus e o etiquetador de Brill [3] utilizou o Penn Treebank Wall Street Journal Corpus (PTWSJC). Problemas advindos do fato de não termos um grande corpus ao nosso dispor durante o desenvolvimento do etiquetador serão discutidos na Seção 6.

Existem vários métodos na literatura para a etiquetagem de palavras a partir de sua categoria morfo-sintática (os chamados *part-of-speech taggers*). Tais etiquetadores podem estar baseados em métodos estatísticos de cadeias de Markov [5], regras de transformação [1, 3] ou por minimização de entropia [10]. Todos esses métodos foram originalmente desenvolvidos para a língua inglesa e dentre estes, um dos que obteve o melhor desempenho foi o método de regras de

---

na observação e coleta sistemática de dados astronômicos, que mais tarde foram utilizados pelo seu assistente Johannes Kepler para o desenvolvimento das famosas leis que levam seu nome.

transformação de Brill [3]. Por esse motivo, e por sua representação explícita do conhecimento, escolhemos o método de Brill para a etiquetagem do Português.

No entanto, o método de Brill não pode ser diretamente aplicado ao Português. Isso se deve ao fato de o Português ser uma língua morfologicamente muito mais rica que o Inglês, o que faz com que o número de etiquetas necessárias para a marcação do Português [4] seja muitas vezes superior ao número de etiquetas utilizadas para o Inglês [12].

Na Seção 2 mostramos uma estimativa da complexidade de um etiquetador baseado em regras de transformação para o Português. Esta estimativa indica não ser factível a mera transposição do método de Brill para linguagens de morfologia rica. Por esse motivo, introduzimos uma nova filosofia de etiquetagem [6] e um novo método de etiquetagem em três fases, descritos na Seção 3, que adapta o método de Brill para as línguas morfologicamente ricas. Pelo novo método, as etiquetas possuem estrutura interna, composta por um *núcleo* e um *complemento*. As duas primeiras fases de nosso método nada mais são do que o método de Brill aplicado ao núcleo das etiquetas. A terceira parte é um refinador do núcleo de cada etiqueta, descrito na Seção 3. A Seção 4 apresenta os resultados obtidos até então, enquanto a Seção 5 descreve problemas encontrados no processo de refinamento e sugere formas de contorná-los. Por fim, apresentamos conclusões sobre o trabalho de construção de um corpus histórico concomitantemente com o desenvolvimento de ferramentas para a sua criação.

É importante salientar que o método de etiquetagem que será descrito a seguir já foi implementado nas linguagens C e Perl e roda em plataformas Unix. No entanto, por nos encontrarmos ainda na segunda iteração do ciclo acima descrito, a fase de testes ainda não foi concluída.

2 A Complexidade da Etiquetagem

Dentre os diversos paradigmas que existem para a etiquetagem morfo-sintática, decidimos utilizar o proposto por Eric Brill [3], já que esse apresenta algumas vantagens sobre os métodos estocásticos ou baseados em redes neuronais, como a representação explícita do conhecimento aprendido sob a forma de regras, além de obter alto percentual de acerto quando treinado com um texto suficientemente grande.

Nessa seção apresentaremos uma *estimativa* da complexidade do algoritmo de Brill em termos do número de etiquetas. Enfatizamos que se trata apenas de uma estimativa, pois ela é baseada na complexidade do pior caso. Felizmente, esta estimativa foi confirmada por experimentos, que estão descritos na Seção 4.

2.1 O Algoritmo de Aprendizado de Regras proposto por Brill

As regras (ou transformações) nada mais são do que uma maneira de se codificar as informações que são aprendidas pelo etiquetador. Tais regras, que são geradas a partir de moldes, são compostas por duas partes: as *ações* e as *condições*.

As primeiras são regras que efetivamente atribuem ou alteram a etiqueta de uma palavra, como mostra o exemplo abaixo:

Troque a etiqueta de VERBO para SUBSTANTIVO.

As condições são regras que basicamente descrevem quando uma ação deve ser aplicada (em outras palavras, elas descrevem situações que devem ser satisfeitas para o acionamento das ações a elas associadas). Um exemplo de regra desse tipo é:

A palavra anterior é um ARTIGO.

Segundo o paradigma de etiquetagem baseado em transformações, há aprendizado de regras em dois contextos. Inicialmente, são aprendidas regras para etiquetagem de palavras desconhecidas, ou seja, de palavras que não pertencem ao léxico[2] que será utilizado pelo etiquetador. Há também o aprendizado de regras contextuais, que é feito de acordo com o algoritmo abaixo:

Entrada: O corpus de treinamento (referência + texto marcado pelo etiquetador inicial), conjunto de etiquetas, conjunto de moldes e a função objetivo.
Saída: A lista de regras contextuais.
O que faz: Constrói a lista de regras contextuais a partir da avaliação de cada uma das possíveis candidatas segundo a função objetivo.

1. **enquanto** regras ainda podem ser aprendidas
2. **para** X $\leftarrow$ $etiq_1$ **até** $etiq_t$
3. **para** Y $\leftarrow$ $etiq_1$ **até** $etiq_t$
4. **para** posição $\leftarrow$ 1 **até** n
5. **para** M $\leftarrow$ $molde_1$ **até** $molde_m$
6. **para cada** COND[3]de M
7. avaliar sua performance
8. **se** esta é a regra com melhor resultado até agora **então**
9. armazená-la como tal.
10. **se** a melhor regra obteve resultado maior ou igual a MÍNIMO[4]**então**
11. adicioná-la ao final da lista de regras
12. atualizar o corpus de treinamento
13. **senão** encerrar o aprendizado

[2] O léxico é um arquivo com as palavras e suas possíveis etiquetas.
[3] Condição.
[4] MÍNIMO é uma constante que deve ser definida para interromper o aprendizado. Valores altos para esta constante impedem que um determinado número de regras seja aprendido, diminuindo o tempo gasto pela fase de aprendizado.

Observe que o algoritmo gera uma lista de regras onde a ordem é importante, já que a escolha da i-ésima regra depende da aplicação das regras escolhidas anteriormente.

A Figura 1 ilustra o processo de aprendizado descrito pelo algoritmo acima.

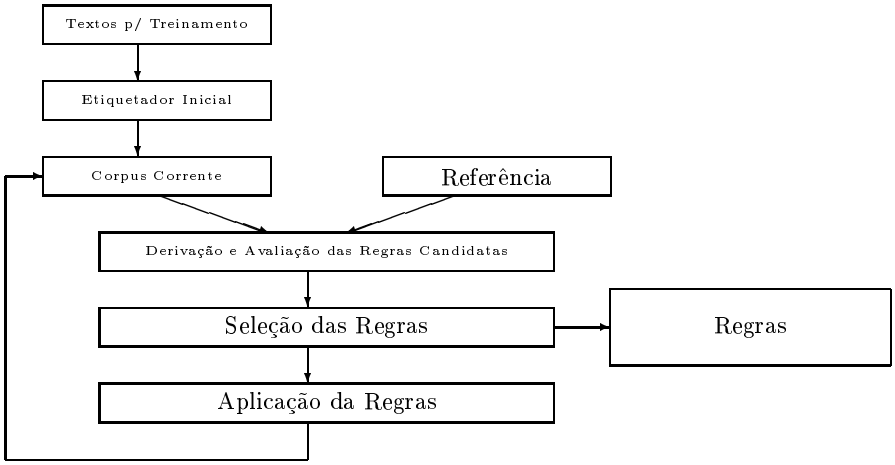

Figura1. Aprendizado de Transformações Dirigido por Erros

Note que o primeiro passo para o aprendizado das regras contextuais é a marcação do texto pelo etiquetador inicial. Para o nosso sistema, utilizamos o etiquetador inicial proposto por Brill, a menos de pequenas alterações no algoritmo de etiquetagem inicial.

2.2 A Complexidade do Algoritmo de Brill

De acordo com o método proposto por Brill, há duas fases distintas para que possamos etiquetar um texto: inicialmente, devemos "treinar" o etiquetador, para que em seguida ele possa ser utilizado para a etiquetagem dos textos propriamente dita. Como a fase de aprendizado gasta muito mais tempo do que a etiquetagem, é com ela que devemos nos preocupar. Segundo Brill [1], o algoritmo de aprendizado tem complexidade, no pior caso, dada por:

$$O(\mid n \mid \times \mid op \mid \times \mid amb \mid),$$

onde $\mid n \mid$ é o tamanho do corpus de treinamento (número de palavras), $\mid op \mid$ é o número de operações do tipo *altere a etiqueta corrente de* $etiq_1$ *para* $etiq_2$ (ações) e $\mid amb \mid$ é o número de condições.

Para facilitar a compreensão, consideremos o seguinte exemplo, onde o único molde para ações é *altere a etiqueta corrente de* $etiq_1$ *para* $etiq_2$ e os moldes para as condições são os seguintes:

1. a palavra anterior tem etiqueta $etiq$;
2. a palavra anterior tem etiqueta $etiq_1$ e a palavra posterior tem etiqueta $etiq_2$;
3. as duas palavras posteriores têm etiquetas $etiq_1$ e $etiq_2$;
4. uma das três palavras anteriores tem etiqueta $etiq$;
5. a palavra a duas posições antes da posição corrente tem etiqueta $etiq$,

onde $etiq$, $etiq_1$ e $etiq_2$ são variáveis sobre o conjunto de etiquetas utilizado.

Nesse caso, o número de ambientes de condições (em função do número de etiquetas a serem utilizadas) para cada um dos moldes acima é dado por:

RRC	*# de ambientes*
1	t
2	$t \cdot t$
3	$t \cdot t$
4	t
5	t

Logo, o número total de ambientes para as condições é:

$$t + t \cdot t + t \cdot t + t + t = 2 \cdot t^2 + 3 \cdot t$$

Claramente, a complexidade de cada iteração do algoritmo de aprendizado para esse conjunto de moldes em função do número de etiquetas que serão utilizadas é:

$$O(t^2 \cdot (2 \cdot t^2 + 3 \cdot t) \cdot \mid n \mid),$$

ou seja, depende da quarta potência do número de etiquetas.

A seguir, faremos um comparativo entre a eficiência do algoritmo de aprendizado no caso acima para diferentes conjuntos de etiquetas. Para isso, consideraremos o conjunto de etiquetas do Penn Treebank Wall Street Journal Corpus (PTWSJC) [12] e do Corpus Tycho Brahe (CTB) [7].

Para o PTWSJC, com 36 etiquetas, temos:

$$\mid op \mid \cdot \mid env \mid \cdot n \ = 36^2 \cdot (2 \cdot 36^2 + 3 \cdot 36) \cdot n = 3.499.200 \cdot n$$

Para o CTB, com 176[5] etiquetas, temos:

[5] Esse é o número de etiquetas atualmente, sem considerarmos as etiquetas formadas pelas aglutinações dos clíticos (ex.: TR+CL, VB-R!CL, etc.). Como o corpus está em processo final de preparação, esse conjunto ainda não está "fechado", podendo mudar no decorrer desse processo.

$$| op | \cdot | env | \cdot n = 176^2 \cdot (2 \cdot 176^2 + 3 \cdot 176) \cdot n = 1.935.101.696 \cdot n$$

Comparando, o tempo gasto na fase de aprendizado utilizando-se os conjuntos de etiquetas citados acima e supondo que os corpora de treinamento têm mesmos tamanhos, temos que:

$$\frac{Tempo(TB)}{Tempo(PTWSJ)} = \frac{1.935.101.696 \cdot n}{3.499.200 \cdot n} \approx 553,$$

ou seja, para corpora de mesmos tamanhos, o aprendizado regras contextuais, para o conjunto de etiquetas do PTWSJC é, no pior caso, aproximadamente 553 vezes mais lento do que o aprendizado utilizando-se o conjunto de etiquetas do CTB.

Observe que o problema de eficiência tende a se tornar mais grave quando é utilizado um maior número de moldes e/ou quando os moldes analisam muitas palavras/etiquetas.

Segundo Brill [2], a fase de aprendizado de regras contextuais, utilizando-se o conjunto de etiquetas do PTWSJC e um corpus de treinamento com 500.000 palavras leva aproximadamente 1 dia em uma SUN Sparc-10. Sendo assim, utilizando-se um corpus com o mesmo tamanho e uma máquina similar, teríamos um tempo de treinamento de 553 dias (ou aproximadamente 1 ano e meio !!!) ao usar o conjunto de etiquetas do CTB.

Observe mais uma vez que os cálculos acima são apenas de uma estimativa, que não considera otimizações feitas na implementação. Contudo, podemos concluir que utilizar o método Brill "puro" para construir um etiquetador para textos em Português baseado nas etiquetas do CTB é computacionalmente inviável.

3 O Refinamento

3.1 O Projeto das Etiquetas

Para resolver o problema apresentado acima, ou seja, diminuir o tempo gasto pelo etiquetador na fase de aprendizado quando o número de etiquetas utilizadas é grande, extendemos o método original para que a nova versão apresentasse maior eficiência computacional[6]. A solução encontrada foi a seguinte: em vez de utilizarmos etiquetas monolíticas (ou seja, compostas por apenas uma parte), decidimos utilizar etiquetas compostas, onde a classe gramatical seria representada pelo núcleo da etiqueta e os demais traços morfo-sintáticos como número, tempo verbal, gênero, etc., seriam representados por um refinamento desse núcleo. Por exemplo, em vez de termos a etiqueta

[6] O método apresentado a seguir não altera a complexidade computacional do algoritmo de aprendizado. Contudo, o tempo de processamento esperado é menor do que seria se utilizássemos o método sem modificações.

DUMFP,

que é uma etiqueta que representa um artigo indefinido, feminino, no plural (ex.: *umas*), temos a etiqueta composta

D-UM-F-P.

Para facilitar o trabalho dos lingüistas que etiquetam textos a mão ou corrigem textos etiquetados automaticamente, algumas etiquetas tem complementos padrões que são omitidos. Por exemplo, uma etiqueta D com o complemento UM indica um determinante indefinido *masculino singular*, e a etiqueta N indica um nome *masculino singular*.

O conjunto de etiquetas utilizado pelo Corpus Tycho Brahe está descrito em [9], em que todas as etiquetas obedecem ao padrão acima. Esta estrutura foi inicialmente proposta em [6]. O conjunto de etiquetas conta com 176 etiquetas, com 36 núcleos distintos.

Essa alteração na *estrutura* das etiquetas nos induziu a acrescentar uma fase no processo de etiquetagem dos textos, já que, em vez de utilizarmos a etiqueta completa na fase de aprendizado, utilizaremos apenas os núcleos, diminuindo a quantidade de etiquetas envolvidas no processo de aprendizado para 36 (coincidentemente, esse é o número de etiquetas do *Penn Treebank Wall Street Journal Corpus*, utilizado por Brill em parte de seus experimentos).

Portanto, após a aplicação das regras aprendidas, haverá uma nova fase na etiquetagem que terá como objetivo exatamente fazer o refinamento dos núcleos para que o texto resultante esteja marcado com as etiquetas completas.

Dessa forma, a etiquetagem de um texto apresentará 3 fases:

Fase I. O texto recebe uma marcação inicial na qual apenas os núcleos das etiquetas serão utilizados. O processo de etiquetagem inicial é o mesmo que o proposto por Brill.

Fase II. Através da aplicação das regras contextuais aprendidas anteriormente, a marcação inicial é alterada para que melhores resultados sejam obtidos (observe que, assim como ocorre na fase anterior, apenas os núcleos das etiquetas serão utilizados). Esta fase também é idêntica à proposta por Brill.

Fase III. É feito um refinamento da marcação de cada uma das palavras (quando for o caso, já que algumas etiquetas não podem ser refinadas) para completar as etiquetas. Esse processo não envolve aprendizado de regras e sim classificação das palavras em função de suas características morfológicas.

As Figuras 2 e 3 apresentam algumas etiquetas do CTB e do PTWSJC, respectivamente. Note a diferença estrutural no formato das etiquetas de cada um dos corpora.

A seguir, mostraremos como esse refinamento pode ser feito de forma eficiente, isto é, sem comprometer a performance dos processos de aprendizado das regras contextuais e etiquetagem.

Etiqueta	Classe Gramatical
CC	Conjunção Coordenativa
DT	Determinante
EX	"*there*" existencial
FW	Palavra Estrangeira
JJ	Adjetivo
JJS	Adjetivo, Superlativo
MD	Modal

Figura2. Algumas Etiquetas do Corpus WSJ Penn Treebank

Etiqueta	Classe Gramatical
ET	Verbo Estar, Infinitivo
VB-PP	Verbo, Particípio Passado
PRO$-F-P	Pronome Possessivo, Feminino, Plural
ADJ-R	Adjetivo Comparativo/Consecutivo de Grau
Q	Quantificadores
NUM	Numeral Cardinal
FW	Palavras Estrangeiras

Figura3. Algumas Etiquetas do Corpus Tycho Brahe

3.2 O Algoritmo de Refinamento

Nessa seção, apresentaremos, em pseudo-código, o algoritmo de refinamento das etiquetas. Vale ressaltar que a utilização desse método só tem sentido caso as etiquetas tenham sido projetadas para serem refinadas.

Entrada: Um texto marcado com os núcleos das etiquetas utilizadas.
Saída: O texto de entrada com a etiquetagem completa.
O que faz: Para cada palavra do texto de entrada, verifica se a etiqueta corrente pode ou não ser refinada. Em caso afirmativo, investiga se há alguma característica na palavra que permita a determinação do complemento da etiqueta. Caso não seja encontrada nenhuma tal característica, utiliza um complemento padrão (caso esta etiqueta deva ser necessariamente refinada, pois pode haver casos onde o refinamento não é obrigatório). Se a etiqueta corrente não admitir refinamento, nada é feito.

1. **para cada** palavra do texto **faça**
2. determine a etiqueta corrente
3. **se** a etiqueta pode ser refinada **então**
4. determine qual o refinamento a ser utilizado
5. etiqueta corrente ← etiqueta corrente + complemento
6. **senão**
7. não faça nada

3.3 Análise do Algoritmo de Refinamento

Vamos, então, analisar a complexidade do algoritmo mostrado na Seção 3.2. Para tanto, calcularemos o número de iterações de cada uma das passagens do algoritmo no pior caso:

Passagem	*# de execuções*
1	n
2	n
3	n
4	k_1
5	k_1
6	k_2
7	k_2

onde:

- n é o número de palavras do texto a ser etiquetado,
- $k_1 + k_2 = n$,
- k_1 é o número de vezes que o teste da linha 3 do algoritmo sucede, ou seja, o número de palavras cujas etiquetas devem ser refinadas,
- k_2 é o número de vezes que o teste da linha 3 do algoritmo falha, ou seja, o número de palavras cujas etiquetas não podem ser refinadas.

Sendo assim, a complexidade do algoritmo de refinamento é dada por:

$$O(n),$$

ou seja, é linear no número de palavras do texto a ser etiquetado.

Como, para cada uma das etiquetas a serem refinadas, há um número finito e constante de possibilidades para refinamento, o passo 4 do algoritmo acima é feito em tempo constante ($O(1)$). O mesmo acontece em relação ao passo 3. Observe também, que as etiquetas não fazem parte da entrada do algoritmo.

Conforme pode ser observado no algoritmo acima, para determinar o refinamento que será utilizado, observa-se dois itens: a etiqueta corrente da palavra e sua morfologia. Basicamente, são analisados os prefixos[7] e sufixos[8] das palavras em regras como:

1. Se a palavra corrente estiver marcada com a etiqueta N e terminar com a letra *s*, alterar sua etiqueta para N-P.[9]

[7] Estamos nos referindo às primeiras letras da palavra, independentemente do fato desse prefixo existir (ou não) na língua portuguesa.

[8] Analogamente, estamos nos referindo às últimas letras da palavra, independentemente do fato desse sufixo existir (ou não) na língua portuguesa.

[9] Esse método falha para a palavra *lápis* e similares; numa futura versão isto será sanado através da utilização de um dicionário a ser compilado a partir do próprio corpus.

2. Se a palavra corrente estiver marcada com a etiqueta VB e terminar com o prefixo *ando*, alterar sua etiqueta para VB-G.
3. Se a palavra corrente for *estivera*, alterar sua etiqueta para ET-RA.

Note que, ao contrário do que ocorre no método proposto por Eric Brill, as regras não são aprendidas e sim codificadas no módulo do sistema que chamamos de refinador. Sendo assim, o sistema que desenvolvemos possui a arquitetura apresentada na Figura 4.

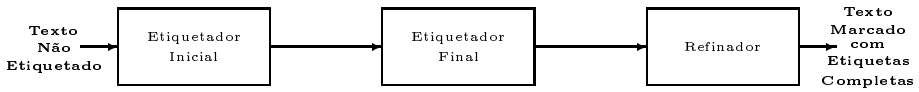

Figura4. Arquitetura do Etiquetador

4 Resultados Obtidos

Algumas experiências já puderam ser realizadas, mesmo com o pequeno corpus disponível (21.000 palavras etiquetadas à mão) na primeira iteração do processo. Ficou claro que, de fato, o método proposto tem fase de aprendizado mais rápida do que o método original, conforme pode ser visto na tabela abaixo:

	Regras Contextuais	Palavras Desconhecidas	Proporção
Etiq. Completas	1h	4h	4:1
Apenas os Núcleos	15h	81h	5.4:1

Quanto à performance do refinador, tivemos os seguintes resultados:

Apenas os Núcleos	Etiq. Completas	Refinador
85.86%	78.28%	91.17%

Esses testes foram realizados utilizando-se um corpus de aproximadamente 5.000 palavras.

Outros métodos foram utilizados para a anotação morfo-sintática de textos do Português, conforme pode ser visto em [8] e [13]. Contudo, eles foram aplicados ao Português Contemporâneo, que não apresenta uma série de problemas observados no Português Clássico (ver próxima seção para maiores detalhes).

Nas próximas iterações, contando com um corpus maior, pretendemos estender esses experimentos.

5 Ambigüidades no Refinamento e Outros Problemas

Durante o desenvolvimento do refinador, nos deparamos com alguns problemas. Dentre eles, podemos citar:

1. O refinamento das etiquetas de palavras homógrafas.
   Há, em Português, formas homógrafas de diferentes tempos verbais, como pode ser observado abaixo:
   (a) i. Quando você terminar, *avisa*/VB-I (verbo avisar, 2ª pessoa do imperativo afirmativo, singular)...
   ii. Ele *avisa*/VB-P (verbo avisar, 3ª pessoa do presente indicativo, singular)...
   (b) i. Todo mundo sabe que nós *cantamos*/VB-P (verbo cantar, 1ª pessoa do presente do indicativo, plural)...
   ii. Todo mundo sabe que nós *cantamos*/VB-D (verbo cantar, 1ª pessoa do pretérito perfeito do indicativo, plural)...
   (c) i. Quando eu *falar*/VB-SR (verbo falar, 1ª pessoa do futuro do subjuntivo, singular)...
   ii. Para eu *falar*/VB (verbo falar, infinitivo)...

   Embora uma solução definitiva ainda não tenha sido encontrada para o caso geral, há algumas situações onde o problema foi resolvido. Por exemplo, para determinar se uma forma verbal está no futuro do subjuntivo ou no infinitivo, utilizamos o seguinte critério: caso o verbo em questão seja precedido pela palavra *quando* ou *se*, será marcado como futuro do subjuntivo. Caso contrário, será marcado como infinitivo.
   Problemas semelhantes acontecem em relação a outras classes gramaticais. Por exemplo:
   (a) Pela manhã, ele regou o jardim. O *mesmo*/N foi feito durante a tarde.
   (b) Ela virá *mesmo*/ADV que algo ocorra.
   (c) O *mesmo*/ADJ homem saiu.
   (d) O *mesmo*/ADJ homem que/WPRO foi ao supermercado acaba de sair.
   (e) Ele fez o *mesmo*/ADJ-R trabalho que/C você.

   Como pode ser visto, a palavra *mesmo* pode ser marcada de diversas maneiras. Contudo, apenas duas delas geram problemas, já que apresentam etiquetas com os mesmos núcleos (formas *d* e *e*). Para resolver esse problema, verificamos qual é a marca recebida pela palavra *que* que precede o advérbio *mesmo*. Se for sucedido por um *que* marcado com C, será marcado com a etiqueta ADJ-R; caso contrário, será marcado com a etiqueta ADJ.
2. A variação na grafia de uma mesma palavra.
   Há, no período estudado, palavras que apresentam mais de uma forma de grafia. Além de dificultar a implementação do refinador (isso torna mais complicado determinar um traço na morfologia da palavra que caracterize sua classe gramatical), afeta as desempenhos dos etiquetadores inicial e final. Pode acontecer que uma palavra conhecida (que pertence ao léxico que será utilizado na etiquetagem) seja tratada como desconhecida.

Consequentemente, a palavra, que seria marcada com a etiqueta mais freqüentemente associada a ela no corpus de treinamento, será marcada como nome comum ou nome próprio. O mesmo acontece com as palavras produzidas por erros de grafia.

3. O conjunto de etiquetas utilizado na marcação dos textos ainda não está completamente definido.
   Isso faz com que muitas vezes tenhamos que alterar a implementação do refinador, além de reetiquetar os trechos do corpus afetados pelas alterações no conjunto de etiquetas.

6 Conclusões: o Desenvolvimento Simultâneo do Corpus e Etiquetador

A conclusão principal deste estágio de trabalho é que não estamos construindo um *etiquetador em si*, mas uma *ferramenta de apoio à construção de um corpus etiquetado*. Dessa forma, nossas preocupações no momento divergem daquelas encontradas por outros desenvolvedores de etiquetadores.

Conforme descrevemos na introdução, diferentemente de outros etiquetadores existentes, nosso etiquetador morfo-sintático está sendo desenvolvido em paralelo com a construção do corpus sobre o qual ele será testado. Tal procedimento pode ser descrito como "tentar navegar em um navio enquanto ele é construído". Por outro lado, essa abordagem é crucial para a criação do Corpus Tycho Brahe, que não estaria sendo construído sem as ferramentas automáticas em questão.

Essa abordagem de desenvolvimento paralela corpus-etiquetador nos tráz problemas
interessantes, os quais ainda estamos vivenciando. Esses problemas advém da escassez de palavras etiquetadas para treinarmos nosso etiquetador. Nosso primeiro conjunto de treino possuia apenas 21.000 palavras etiquetadas a mão. Dessas, por motivos metodológicos, apenas 11.000 foram usadas para treinar o etiquetador (primeira fase); o resto foi utilizado para testar o etiquetador e para o aprendizado de regras para etiquetagem de palavras desconhecidas.

Em sua tese, Eric Brill considera que um corpus "muito pequeno" possuiria 40.000 palavras. Dessa forma, não é surpreendente que os resultados obtidos na primeira iteração treino-etiquetagem-correção não tenham sido estupendos. No entanto, não sabemos precisar a razão da baixa precisão obtida, que podem ser justificadas de duas maneiras:

(a) O número de palavras (11.000) não é *estatisticamente relevante* para o treinamento do etiquetador.
(b) O método de Brill, por se basear no aprendizado da etiqueta baseado na vizinhança próxima, não se aplica ao Português do século XVII, que é uma língua com uma ordem de palavras muito mais livre que o Inglês moderno.

O problema com a hipótese (a) é que não sabemos quantas palavras seriam *estatisticamente relevantes*, nem sabemos como calcular tal valor. Some-se a

isso o fato de estarmos tratando com textos históricos, de autores de épocas, convenções ortográficas e estilos distintos, o que nos diferencia muito dos corpora nos quais outros métodos citados foram desenvolvidos.

O problema da hipótese (b) é que absolutamente todos os métodos conhecidos para etiquetagem de palavras fazem uma busca localmente restrita e, portanto, todos os outros métodos estão sujeitos ao mesmo problema.

7 Trabalhos Futuros

O primeiro trabalho que estaremos realizando será a validação da hipótese (a) acima. Para isso, necessitamos aumentar o número de palavras etiquetadas no corpus, para podermos retreinar o etiquetador. Se a precisão do etiquetador aumentar com os novos treinamentos, teremos um indicador que realmente estávamos lidando com um problema de falta de relevância estatística. Poderemos então, experimentalmente, determinar o limite da relevância estatística.

Caso não haja uma melhora na precisão com o aumento do corpus de treinamento, teremos de buscar novos métodos de etiquetagem.

Essa experiência será de grande importância para o desenvolvimento da segunda fase do Corpus Tycho Brahe, que pretende apresentar estrutura sintática dos textos que, por hora, temos apenas na versão pura e etiquetada morfo-sintaticamente.

Outros pontos que pretendemos investigar é como utilizar o refinador para acrescentar um mecanismo de recuperação de erros cometidos na segunda fase do processo de etiquetagem, bem como a utilização de métodos para pré-etiquetagem que possam melhorar a marcação inicial dos textos e evitar que erros cometidos nessa fase possam se propagar para fases posteriores ([11] mostra um exemplo de como um texto pode ser pré-etiquetado).

Referências

1. Brill, E. *A Corpus-Based Approach to Language Learning.* Ph.D. Dissertation, Department of Computer and Information Science, University of Pennsylvania, 1993.
2. Brill, E. *README.TRAINING.* Arquivo pertencente à distribuição do etiquetador, disponível na URL `http://www.cs.jhu.edu/brill`, 1993.
3. Brill, E. *Transformation-Based Error Driven Learning and Natural Language. A case Study in Part of Speech Tagging.* Computational Linguistics, 21(4):543-565, 1995.
4. Britto, H., C. Galves, I. Ribeiro, M. Augusto, A. P. Scher *Morphological Annotation System for Automatic Tagging of Electronic Textual Corpora: from English to Romance Languages.* 6th International Symposium on Social Communication, Applied Linguistics Center, Santiago de Cuba, 1998.
5. Charniak, E. *Statistical Language Learning.* MIT Press, 1993.

6. Finger, M. *Tagging a Morphologically Rich Language.* Proceedings of the First Workshop on Text, Speech and Dialogue (TSD'98), pp. 39–44, Brno, Czech Republic, 1998.
7. Home-page do projeto *Rythimic Patterns, Parameter Setting & Language Change*, disponível na URL `http://www.ime.usp.br/~tycho`.
8. Marques, N. C., Lopes, J. G. *A Neural Network approach to Part-of-Speech Tagging.* Proceedings of the Second Workshop on Computational Processing of Written and Spoken Portuguese, Curitiba, Brazil, 1996.
9. *General Manual: The Tycho Brahe Parsed Corpus of Historical Portuguese*, disponível na URL `http://www.ime.usp.br/~tycho/corpus/manual`.
10. Ratnaparkhi, A. *Maximum Entropy Models for Natural Language Ambiguity Resolution.* Ph.D. Dissertation, Institute for Research in Congnitive Science, University of Pennsylvania, 1998.
11. Reis, R., Almeida, J. J. *Etiquetador Morfo-sintáctico para o Português.* Actas do XVII Congresso da Associação Portuguesa de Lingüística, Braga, 1996.
12. Santorini, B. *Part of Speech Tagging Guidelines for the Penn Treebank Project.* 3a. revisão, 2a. impressão, 1990 (versão atualizada em 1995 por Robert MacIntyre).
13. Villavicencio, A., Marques, N. M. C., Lopes, J. G. P., Villavicencio, F. *Part-of-Speech Tagging for Portuguese Texts.* Proceedings of the 12th Brazilian Conference on Artificial Intelligence (SBIA'95). Lecture Notes in Computer Science 991, 1995.