Técnicas de Otimização da Precisão Empregadas no Etiquetador Tycho Brahe

Marcelo Finger*
mfinger@ime.usp.br

Departamento de Ciência da Computação
Instituto de Matemática e Estatística
Universidade de São Paulo

dedicado à memória de
Carlos Daniel Chacur Alves

Resumo

O Corpus Etiquetado Tycho Brahe do Português Europeu Histórico está sendo construído em paralelo com o desenvolvimento de ferramentas que permitam a sua etiquetagem automática. Em [AF99] discutimos os princípios básicos sobre os quais foi construído o etiquetador baseado no método Brill, utilizando técnicas de aprendizado computacional, e apresentamos os primeiros resultados.

Neste trabalho, mostramos as técnicas que foram usadas para aumentar a precisão do etiquetador. As modificações influiram tanto na arquitetura quanto nos algoritmos do sistema, e afetaram tanto a parte de treinamento quanto a de aplicação da etiquetagem. Discutimos os efeitos de cada mudança, apresentando os ganhos de precisão ao final de cada modificação.

Abstract

The Tycho Brahe Parsed Corpus of Historical Portuguese is being developed in parallel with the automated tools used for its construction. In [AF99], we described the principles used in the development of the Tycho Brahe automated part-of-speach tagger, based on a Brill's transformation rule learning technique.

*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*).

In this work, we show several techniques used to increase the precision of the initial tagger. Several modifications were proposed that affected both the architecture and the algorithms used in the tagger, in the learning phase and the application phase. We discuss the effects of each proposed modification, showing the gains, if any, caused by each one of them.

1 Introdução

A construção do Corpus Tycho Brahe de Português Histórico [Tyc99b] motivou o desenvolvimento de um etiquetador para o Português Europeu Clássico (PEC) e Português Europeu Moderno (PEM). Este etiquetador recebeu o mesmo nome que o corpus [Tyc99a].

Etiquetar um texto quer dizer classificar cada palavra em contexto a uma categoria morfo-sintática. Por exemplo, dada como entrada a seguinte sentença do corpus Tycho Brahe:

```
jejua o enfermo para recuperar a saúde
```

a saída do processo de etiquetagem seria a mesma sentença no formato $\langle palavra \rangle / \langle etiqueta \rangle$:

```
jejua/VB-P o/D enfermo/N para/P recuperar/VB a/D-F saúde/N
```

O conjunto de etiquetas do Corpus Tycho Brahe e seu significado está definido em [Tyc99a].

Em [AF99] discutimos os princípios básicos sobre os quais foi construído o etiquetador Tycho Brahe. Basicamente, trata-se de um etiquetador baseado em regras transformacionais, isto é, regras que sucessivamente transformam uma etiquetagem inicial. Tais regras são aprendidas na *fase de treinamento* a partir de um corpus etiquetado manualmente. Este etiquetador baseia-se em um método proposto por Brill [Bri93, Bri95]. Um exemplo de tal regra seria:

> Mude a etiqueta da palavra corrente de `VB` (verbo) para `N` (nome)
> se a etiqueta da palavra anterior é `D` (determinante).

Devido à riqueza morfológica do português, o método inicial de Brill não se aplica diretamente ao português e, para se tornar aplicável, teve de ser adaptado [Fin98, Alv99, AF99].

Os resultados iniciais obtidos para o Corpus Tycho Brahe indicavam que o etiquetador possuía uma precisão de 78,28%. Na data da escrita deste

artigo, o etiquetador encontra-se com 95,43% de precisão. Isto corresponde a uma diminuição de 4,55 vezes no número de erros cometidos pelo etiquetador.

É das técnicas e procedimentos que nos levaram a mais que quadruplicar a precisão do etiquetador que nos ocupamos no presente trabalho. Na Seção seguinte apresentamos a arquitetura do etiquetador de Brill e a inserção de um Refinador feita no Etiquetador Tycho Brahe. Na Seção 3 discutimos o aumento da precisão gerado pelo aumento do corpus de aprendizado. Nas Seções 4 e 5 discutimos estratégias para melhorar a eficiência do Refinador e do método de Brill. Nas Seções 6 e 7 discutimos o efeito da adição de um módulo de pós-correção e sua retroalimentação.

2 Arquitetura do Etiquetador Tycho Brahe

O etiquetador Tycho Brahe é baseado no método desenvolvido por Eric Brill, mas fez-se necessário adaptar este método e introduzir modificações. O método de Brill [Bri95] para etiquetagem morfo-sintática de palavras é um método baseado em aprendizado computacional. O programa aprendedor gera uma série de regras contextuais que serão usadas na etiquetagem. A fase de etiquetagem está representada na Figura 1.

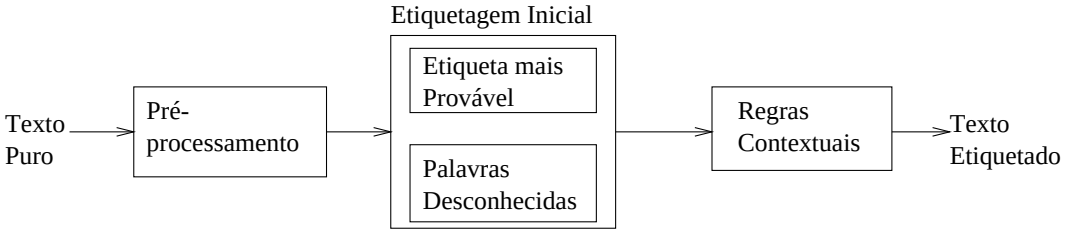

Figura 1: Etiquetagem de textos segundo o método Brill

O processo de etiquetagem é iniciado submetendo o texto a um pré-processamento, que realiza a formatação e filtragem do texto. Em seguida, no *etiquetador inicial*, cada palavra é associada a uma etiqueta; palavras conhecidas são associadas à etiqueta mais freqüente no corpus de treinamento. Por fim, vem a fase principal da etiquetagem, em que as regras contextuais, também aprendidas na fase de treinamento, modificam a etiquetagem inicial de acordo com o contexto em que cada palavra aparece no texto. Ao final, temos um texto etiquetado.

Existe um módulo especial para etiquetagem inicial de palavras desconhecidas, o qual dispõe de regras também aprendidas na fase de treinamento. Desta forma, na fase de treinamento aprende-se dois conjuntos de regras: regras de etiquetagem de palavras desconhecidas e regras de transformação contextual.

Devido à riqueza morfológica do português, o número de etiquetas para o português é da ordem de 300, muito maior que o número de etiquetas para o inglês utilizadas no Penn-Treebank Wall Street Journal Corpus [San90], da ordem de 30. Como foi demonstrado em [Fin98, Alv99], a complexidade do treinamento pelo método de Brill tem complexidade $O(N^4)$, onde N é número de etiquetas. Nossas estimativas teóricas mostram que o treinamento para o português é mais de 300 vezes mais lento que para o inglês, assumindo-se que estamos treinando com um corpus de referência de mesmo tamanho. Na prática, ficou comprovada a impraticabilidade do método de Brill ser aplicado diretamente para o português, pois o tempo de treinamento de português foi de 14 dias[1], enquanto que para o nosso método descrito abaixo, o treinamento durou 42 horas.

Para evitar o problema da complexidade excessiva do método de Brill, propusemos uma reestruturação das etiquetas. As etiquetas do Corpus Tycho Brahe são compostas de uma parte básica e de um sufixo. Por exemplo, a etiqueta

`D-UM-F-P`

possui parte básica `D`, designando um determinante, e sufixos designando os traços de indefinido (`UM`), feminino (`F`) e plural (`P`).

O número de etiquetas básicas presentes no nosso conjunto de etiquetas é de 33, da mesma ordem do tamanho do conjunto de etiquetas do inglês. Portanto, o treinamento de etiquetagem de apenas etiquetas básicas voltou a ser factível, contornando assim o problema da complexidade.

Na nossa arquitetura, a etiquetagem inicial e as regras contextuais se referem apenas à parte básica das etiquetas. Para obtermos a etiquetagem completa precisamos submeter a saída do etiquetador de Brill a um *Refinador*, que é um programa contendo informações lingüísticas, tanto morfosintáticas quanto conceituais, que realiza o processo de complementar com um sufixo a etiqueta básica obtida no método de Brill. Por exemplo, no caso da palavra `umas`, o Refinador ao receber na sua entrada a etiquetagem `umas/D` deve fornecer na saída `umas/D-UM-F-P`. A arquitetura de etiquetagem ficou como ilustrado na Figura 2.

[1] Na realidade, o experimento foi interrompido antes do seu término, devido a uma falta de energia elétrica. O experimento nunca foi repetido.

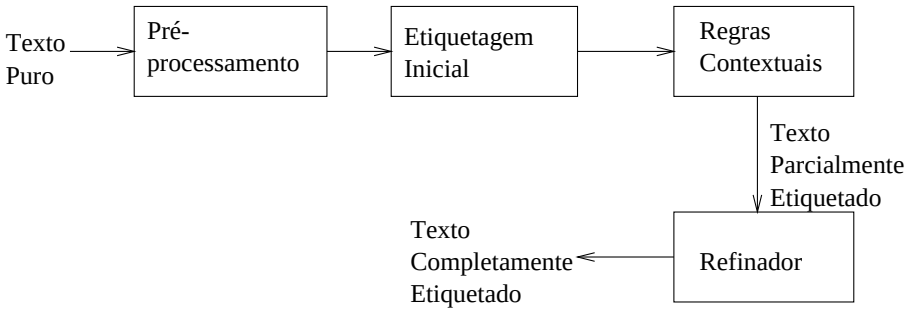

Figura 2: Arquitetura do Etiquetador Tycho Brahe

Para medirmos a precisão do etiquetador, necessitamos de um corpus etiquetado de teste, diferente do corpus de treinamento, manualmente verificado. Aplica-se o etiquetador a este texto, sem as etiquetas. A *precisão* da etiquetagem é dada por:

$$Precisão = \frac{N.\ de\ etiquetas\ coincidentes}{N.\ de\ etiquetas\ total}$$

Ao inserirmos um módulo a mais, a precisão do sistema decai pois o Refinador não tem eficiência de 100%. Este é o preço que se paga para viabilizar o treinamento. Nas nossas medidas iniciais [AF99], o Refinador tinha uma eficiência de 91,17%, reduzindo em pouco menos de 9% o resultado da etiquetagem de etiquetas básicas pelo método de Brill.

3 Aumento do Corpus de Treinamento

Um dos principais problemas de se ter um programa baseado em aprendizagem computacional é que não sabemos qual o tamanho de um corpus de treinamento que seja *estatisticamente relevante.* O nosso primeiro esforço de melhorar a eficiência do etiquetador focalizou a obtenção de um corpus de treinamento maior.

O desenvolvimento do Corpus Tycho Brahe se deu em paralelo com o da ferramenta de etiquetagem. Por isso, não tínhamos disponível, a princípio, um corpus de treinamento muito grande, apenas 21.000 palavras manualmente etiquetadas. O desenvolvimento simultâneo do corpus e do etiquetador se dá em ciclos de treinamento/reetiquetagem/verificação manual, para que o resultado possa ser usado no próximo ciclo de treinamento. Até o momento, tivemos três ciclos de treinamento, indicados na Tabela 1.

Ciclo	*N. de palavras do corpus de treino*	*N. de palavras do corpus de teste*	*Precisão*
1	21.000	5.000	78,28%
2	50.000	20.000	84,23%
3	130.000	45.000	88,24%

Tabela 1: Ciclos de treinamento do etiquetador Tycho Brahe

Como vemos, em cada iteração houve uma multiplicação do tamanho do corpus de treinamento da ordem de 2 a 3 vezes, e o ganho em precisão foi de 6% no primeiro caso e de 4% no segundo. Do ponto de vista de correção de erros, vemos que houve quase uma redução à metade da porcentagem de erros cometidos na versão inicial e após o terceiro ciclo.

Isto indica que foi uma decisão acertada nos dedicarmos, no primeiro momento, ao aumento do tamanho do corpus de treinamento. Este aumento do corpus de treinamento não teria sido possível sem a colaboração dos lingüistas ligados ao Corpus Tycho Brahe. Com 130.000 palavras para treino, passamos a nos dedicar à melhoria do Refinador.

4 Melhorias no Refinador

Durante o aumento do corpus, o Refinador sofreu algumas melhorias. Com 130.000 palavras, a precisão da etiquetagem de etiquetas básicas obtidas pelo etiquetador de Brill era de 95,21%. Isto indica que a eficiência do Refinador era da ordem de 92,6%, provocando uma queda de 7,5% na eficiência final da etiquetagem. Decidimos melhorar a eficiência da etiquetagem.

Analisando os erros cometidos, vimos que a maior parte dos erros eram devidos a refinamentos incorretos de tempos verbais. Além disso, verificamos que o etiquetador de etiquetas básicas cometia erros facilmente detectáveis, mas que não eram corrigidos pelo Refinador, pois este nunca alterava a etiqueta básica.

4.1 O Desconjugador de Verbos

Para tratar do problema do refinamento dos tempos verbais, sofisticamos a parte do Refinador que faz a análise morfológica de verbos. Criamos um programa que faz a inversão das declinações verbais. Ou seja, dada uma palavra e assumindo-se que seja um verbo conjugado, o programa detecta

a raiz verbal, a pessoa e o tempo verbal. A este programa chamamos de *Desconjugador Verbal*.

Para a confecção do desconjugador, nos baseamos em duas fontes. A primeira delas foi o programa conjugador de verbos de [Kar99] utilizado na geração do dicionário *br.ispell* do programa emacs. A outra fonte foi o Breviário da Conjugação de Verbos [Rei78].

O desconjugador funciona basicamente como um identificador de sufixos e um verificador de conjugações irregulares e exceções. Ele possui um dicionário de verbos, possibilitando a desambiguação de algumas formas verbais ambíguas. O programa também gera uma lista de verbos "deduzidos", ou seja, verbos que não constam do dicionário, mas cuja forma infinitiva foi inferida durante o funcionamento. A verificação manual desta lista de verbos permite uma atualização rápida do dicionário de verbos, o que aumenta a precisão do desconjugador.

4.2 Pré-correção do Refinamento

Um dos maiores problemas do Refinador, na forma como foi concebido inicialmente, era o fato de que, se a etiqueta básica estava errada na sua entrada, ela certamente permaneceria errada após o refinamento.

No entanto, existem uma série de palavras em português que são muito freqüentes e que possuem uma única categoria morfo-sintática (na maioria esmagadora de sua ocorrências). Um exemplo de tais palavras está ilustrado na Tabela 2.

Palavra	*Categoria*
e	conjunção coordenativa
me, te, lhe	pronome clítico
em, de, com, sem, sob	preposição

Tabela 2: Palavras com uma única categoria morfo-sintática

O algoritmo de Brill, por tratar-se de um método ignorante de fatos lingüísticos, algumas vezes produzia erros nestas palavras. Para evitar tais erros, adicionamos ao refinador um módulo de *pré-correção* que modifica a etiqueta básica de palavras que patentemente estão erradas. Tal problema era mais sério nos ciclos iniciais, quando tínhamos poucas palavras para o treinamento, mas de certa forma esta pré-correção nunca insere erros e foi mantida.

Com o retorno fornecido pelos lingüistas que verificam a saída do etiquetador, foram adicionadas ao módulo de pré-correção regras mais sofisticadas. Os resultados das alterações do Refinador ajudaram, e muito, a melhorar a precisão da etiquetagem, conforme indicado na Tabela 3.

Corpus de Treinamento	*Precisão do Refinamento*	*Precisão da Etiquetagem*
130.000 palavras	97,19%	92,53%

Tabela 3: Precisão após as melhorias do Refinador

5 Modificações no Algoritmo de Brill

Com o sucesso na alteração do Refinador, começamos a melhorar também o próprio processo de etiquetagem pelo método Brill. Duas modificações foram bem sucedidas, uma modificação na fase de etiquetagem e outra na fase de aprendizado.

Modificação na Etiquetagem. O método de Brill, durante a etiquetagem, é capaz de atribuir qualquer etiqueta a uma palavra, dependendo apenas das regras contextuais. A idéia de modificação da etiquetagem busca *limitar o conjunto de etiquetas que podem ser atribuídas a uma palavra.*

As palavras de um texto a ser etiquetado são divididas em dois tipos: palavras conhecidas, que são aquelas que ocorreram pelo menos uma vez no corpus de treinamento; caso contrário é uma palavra desconhecida.

A limitação imposta foi a seguinte: se durante a etiquetagem encontramos uma palavra conhecida, então apenas será permitido que as regras de etiquetagem atribuam a esta palavra alguma das etiquetas vistas durante o treinamento. Desta forma, algumas regras de etiquetagem ficam bloqueadas se tentarem atribuir a uma palavra conhecida um etiqueta nunca vista associada a esta palavra. As palavras desconhecidas não sofrem desta restrição.

Esta restrição só pode ser usada quando o corpus de treinamento é "suficientemente grande", pois com poucas palavras de treino é possível que muitas das possibilidades palavra/etiqueta simplesmente não tenham ocorrido no treinamento.

Não podemos dizer que tal restrição constitui uma verdadeira modificação no método de Brill. Apesar desta restrição não constar da sua descrição em [Bri95], encontramos esta possibilidade como um parâmetro de compilação no código original de Brill.

Esta modificação causou um incremento de 0,19% na precisão final (ou seja, nas etiquetas completas) do etiquetador.

Modificações no Treinamento. O aprendedor de regras usado pelo método Brill funciona da seguinte maneira. Um grande número de regras é gerado de acordo com vários moldes de regras. Cada uma das regras é aplicada ao corpus de treinamento e comparada com uma referência (assume-se que a referência esteja correta). A cada regra associa-se uma pontuação da seguinte maneira:

- $+1$ para cada erro de etiquetagem corrigido pela regra.
- -1 para cada erro de etiquetagem introduzido pela regra.

A regra escolhida é aquela com a melhor pontuação. Esta regra é armazenada e aplicada ao corpus de treinamento, e o processo continua até que as regras não produzam pontuações acima de um limite prefixado, que pode ser 0.

Desta forma, regras que causam alguns erros ainda assim podem ser consideradas como boas regras. A idéia é de considerar uma punição $-N$ maior para cada erro, $N > 1$. Desta forma, regras que quase não erram estarão sendo beneficiadas contra regras que produzem alguns erros.

Treinamos o etiquetador para vários valores de $N \geq 1$ de punição, e obtivemos os dados que estão na Tabela 4. Como o treinamento se refere às etiquetas básicas, a medida principal é a precisão nas etiquetas básicas.

Punição	*Precisão em etiquetas básicas*	*Precisão em etiquetas completas*
1	95,21	92,53
2	95,63	92,94
4	95,75	93,04
8	96,50	93,78
100	96,05	93,36

Tabela 4: Valores de precisão para diferentes punições

Todos os valores da Tabela 4 se referem ao treinamento com 130.000 palavras. Não sabemos explicar por que o resultado para punição $N = 8$ é tão superior aos demais, mas este não nos parece um fenômeno estável, pois para outros testes que fizemos, a punição máxima foi ora para $N = 4$,

ora para $N = 100$. No entanto, consistentemente, valores de $N > 1$ sempre deram resultado melhor que $N = 1$.

O valor de $N = 100$ tem um significado especial. Como nenhuma de nossas regras corrige mais que 100 erros, ao fixarmos a punição em 100 estamos de fato proibindo a seleção de regras que causem qualquer tipo de erros. Decidimos trabalhar com valores de punição de 100. Em futuros ciclos do sistema, pretendemos realizar novos testes. O ganho de precisão neste caso foi de 0,83%.

6 O Pós-Corretor

A correção manual de textos automaticamente etiquetados feita por lingüistas, levou-os a notar certos erros freqüentes na etiquetagem. Estes erros, muitas vezes, eram claramente dependentes do contexto em que certas palavras ocorriam.

Por exemplo, em português, o artigo definido *a* nunca ocorre antes de verbos finitos. Desta forma a seguinte etiquetagem estará claramente errada:

```
... a/D-F sustentaram/VB-P ...
```

Ocorre que, durante a etiquetagem das etiquetas básicas, a informação sobre a finitude do verbo não está presente, e o programa não poderá aprender tal regra. No entanto, tal erro é facilmente detectável e sua correção também é obvia. No caso do exemplo anterior, a etiqueta da palavra *a* deve ser alterada para pronome clítico e portanto a etiquetagem correta deve ser:

```
... a/CL sustentaram/VB-P ...
```

Este exemplo foi retirado diretamente do Corpus Tycho Brahe, e como este, há vários outros erros, os quais podem ser corrigidos. Desta forma foi criado um *pós-corretor* a ser inserido no sistema após o refinamento. A arquitetura do sistema ficou conforme ilustrado na Figura 3.

A implementação deste pós-corretor foi fortemente influenciada pela filosofia de implementação de Gramáticas de Restrição [Kar95], se bem que o paradigma de gramáticas de restrição não foi implementado. Basicamente, tentamos identificar situações contextuais patentemente não-gramaticais que eram resultantes da etiquetagem. Para estas situações de erro identificado, em geral tínhamos uma ou poucas opções de correção. No caso da correção ser única, esta era diretamente aplicada. No caso de mais de uma possibilidade, duas opções foram adotadas:

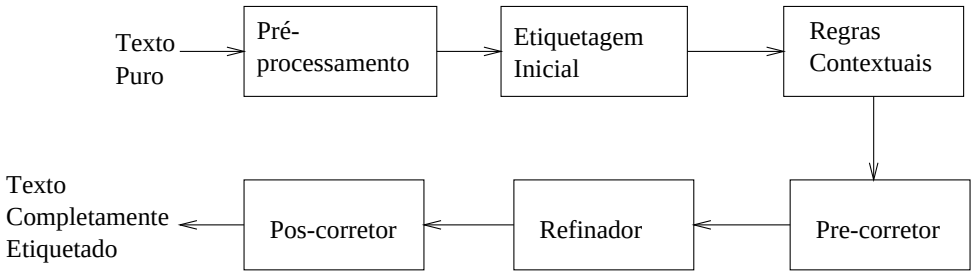

Figura 3: Arquitetura Final do Etiquetador Tycho Brahe

- um contexto mais abrangente era explorado na tentativa de desambiguar a etiquetagem;
- caso não fosse encontrado um tal contexto, a etiqueta mais provável é atribuída à palavra mal-etiquetada.

A geração destas regras contextuais deve muitíssimo às opiniões emitidas pelos lingüistas que trabalharam no Corpus Tycho Brahe. Muitas das opiniões por eles emitidas levaram à novas alterações no pré-corretor, o que adicionou 0,68% à precisão final do refinamento.

O pós-corretor adicionou 0,78% à precisão final. O ganho acumulado desta atividade de elaboração de regras foi de 1,46%. E novas regras continuam surgindo, à medida que os lingüistas vão corrigindo manualmente mais textos automaticamente etiquetados.

Com estas alterações, a precisão final da etiquetagem ficou em 95,43%.

7 Retroalimentação do Etiquetador

Uma vez que o processo de refinação agora faz a correção de erros das etiquetas básicas, tivemos a idéia de conectar a sua saída como entrada da aplicação de regras contextuais. Ou seja, decidimos retroalimentar a saída do pós-corretor a entrada do sistema, como se este fosse o resultado da etiquetagem inicial, para verificarmos o ganho do sistema.

Obviamente, para que esta conexão seja possível, temos que filtrar os sufixos, pois a aplicação de regras contextuais se dá no nível das etiquetas básicas apenas. A arquitetura retroalimentada do sistema é mostrada na Figura 4.

Os resultados desta retroalimentação foram patéticos. Na primeira reiteração, obtivemos uma melhora de apenas 0,02%, ou seja, apenas 9 etiquetas

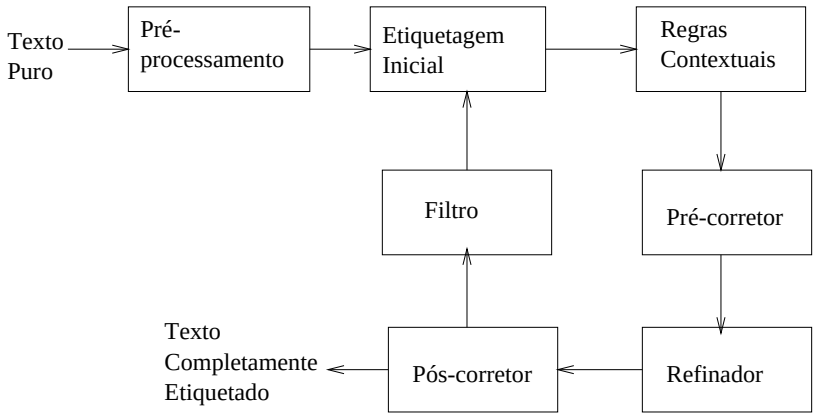

Figura 4: Arquitetura do Etiquetador Retroalimentada

foram corrigidas em 45.000. Na segunda reiteração, não houve nenhuma correção. A precisão final do sistema ficou em 95,45%.

Pelo exposto acima, decidimos abandonar a idéia de reiteração. A arquitetura final do sistema permaneceu aquela da Figura 3.

8 Conclusões

As técnicas de aumento de precisão aqui apresentadas nos mostram dois pontos básicos:

- O aumento no tamanho do corpus de treinamento é a atividade que mais traz aumentos na precisão da etiquetagem.
- Uma vez saturada o aumento por retreinamento, devemos anexar conhecimento lingüístico ao processo para otimizá-lo ainda mais.

Esta segunda fase depende da intensa colaboração de lingüistas no processo de detecção de erros sistemáticos e de sugestão de meios para corrigi-los.

Como trabalhos futuros, podemos citar:

- O retreinamento do corpus para maiores valores ainda e o exame de um valor de punição ideal, conforme sugerido na Seção 5.
- A adaptação do método aqui apresentado para ortografias antigas do português.

- A utilização deste etiquetador no processo de análise sintática automática.

Agradecimentos

Este trabalho não teria sido possível sem a colaboração do grupo de lingüistas do Corpus Tycho Brahe: Charlotte Galves, Helena Britto, Cristiane Namiuti, Taís Meneghati e Patrícia Abdo.

Referências

[AF99] Carlos D. C. Alves and Marcelo Finger. Etiquetagem do português clássico baseada em córpora. In *IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR99)*, Évora, Portugal, 21–22 September 1999.

[Alv99] C. D. Chacur Alves. Etiquetagem de textos em português baseada em corpus. Master's thesis, Departamento de Ciência da Computação, Instituto de Metemática e Estatística da USP, 1999.

[Bri93] E. Brill. *A Corpus-Based Approach to Language Learning*. PhD thesis, Department of Computer and Information Science, University of Pennsylvania, 1993.

[Bri95] E. Brill. Transformation-Based Error Driven Learning and Natural Language: A case Study in Part of Speech Tagging. *Computational Linguistics*, 21(4):543–565, 1995.

[Fin98] M. Finger. Tagging a morphologically rich language. In *Proceeding of the first Workshop on Text, Speech and Dialogue (TSD'98)*, pages 39–44, Brno, Czech Republic, 1998.

[Kar95] F. Karlsson. *Constraint Grammar*. Springer Verlag, 1995.

[Kar99] Ricardo Ueda Karpischek. Conjugador de verbos da língua portuguesa, outubro 1999. Disponível em `http://www.ime.usp.br/~ueda/`.

[Rei78] Otelo Reis. *Breviário da Conjugação de Verbos*. Livraria Francisco Alves Editora, 37^a edição revista e atualizada edition, 1978.

[San90] B. Santorini. *Part of Speech Tagging Guidelines for the Penn Treebank Project.*, third revision edition, 1990. Updated in 1995 by Robert MacIntyre.

[Tyc99a] General manual: The Tycho Brahe Parsed Corpus of Historical Portuguese. Available at `http://www.ime.usp.br/~tycho/corpus/manual`, 1999.

[Tyc99b] Home-page do projeto *Rythimic Patterns, Parameter Setting & Language Change*. Available at `http://www.ime.usp.br/~tycho`, 1999.