Etiquetagem de Palavras para o Português do Brasil

Miriam L. Domingues[1], Eloi L. Favero[1], Ivo P. Medeiros[1]

[1]Centro Tecnológico – Universidade Federal do Pará (UFPA)
Programa de Pós Graduação em Engenharia Elétrica
Caixa Postal 8619, Belém – PA, Brasil, 66.075-900
{miriam,favero}@ufpa.br, ivo701@gmail.com

Resumo. *Neste trabalho, são investigados recursos de software e de corpus para desenvolver um etiquetador para o Português do Brasil que alcance uma acurácia superior a 99%. Centrado em uma solução híbrida, o artigo apresenta um estudo exploratório variando o componente probabilístico, o número de etiquetas e o número de regras.*

1. Introdução

A etiquetagem é a tarefa básica de rotular palavras de uma sentença com etiquetas morfossintáticas que as identificam como categorias gramaticais (substantivos, verbos, etc.) e podem, ainda, conter atributos refinados de cada categoria, por exemplo: gênero e número para um substantivo. A etiquetagem costuma apresentar o problema da ambigüidade que, em uma sentença, é normalmente resolvida com base no contexto em que se encontra a palavra. Para facilitar essa tarefa, são usados etiquetadores automáticos (*taggers*), que são implementados em duas abordagens principais: ***baseada em regras*** e ***probabilística***. Uma terceira abordagem chamada ***híbrida*** envolve a combinação de ambas.

Alguns etiquetadores conhecidos para o português brasileiro são o de Kinoshita et. al. (2006), que alcança uma acurácia de 95%, em média, e o etiquetador baseado em regras com *Constraint Grammar* (CG) de Eckhard Bick (1996) que possui acurácia superior a 99%, mas não está disponível para ser embutido em aplicações.

Este trabalho tem como objetivo encontrar subsídios para a construção de um etiquetador com acurácia mínima de 99% que possa ser embutido em ferramentas de Processamento de Linguagem Natural de uma plataforma de ensino virtual (Harb et al., 2003). A acurácia será pesquisada levando-se em consideração os fatores: método de etiquetagem, tamanho e a qualidade do corpus disponível para treinamento, conjunto de etiquetas e conjunto de regras. A metodologia utilizada consiste em: testar um etiquetador probabilístico, trabalhar sobre os fatores listados acima para maximizar a acurácia do etiquetador e desenvolver uma solução final centrada em uma abordagem híbrida com a implementação de duas estratégias: 1) inserção de regras para a correção de erros na etiquetagem; 2) modificação das etiquetas utilizadas, que passam a incorporar maior quantidade de atributos léxicos.

Neste artigo, a Seção 2 descreve os recursos utilizados no trabalho. A Seção 3 detalha o processo de etiquetagem. A Seção 4 relata os experimentos do estudo de caso, seus resultados e as soluções propostas e a Seção 5 apresenta as considerações finais.

2. Conjunto de Recursos

Os etiquetadores probabilísticos empregados neste estudo são o QTAG (Mason, 2006) e o TreeTagger (Schmid, 1994), ambos de uso livre e multilíngües. O corpus escolhido para treinamento e testes é o Bosque CF 7.4 do projeto Floresta Sintá(c)tica, que é a parte do Corpus CETENFolha já revista por analistas humanos (Linguateca, 2007). O CETENFolha é formado por textos do jornal Folha de São Paulo. Nesta pesquisa, o uso do Bosque, com apenas 80.078 palavras, se deu em razão de que, em experimentos realizados com a parte não revista, a taxa de acertos ficou abaixo de 90%, causada pela presença de muitos erros no Corpus. As etiquetas do Bosque se distribuem em 18 categorias gramaticais: adjetivo, advérbio, artigo, conjunção coordenativa, conjunção subordinativa, interjeição, nome ou substantivo, numeral, pronome determinativo, pronome independente, pronome pessoal, nome próprio, preposição, verbo finito, verbo no gerúndio, verbo no infinitivo, verbo no particípio e pontuação.

3. Processo de Etiquetagem das Palavras

O processo inicia com a geração de arquivos de recursos léxicos do português brasileiro para cada etiquetador. As informações para esses arquivos são retiradas do corpus de treinamento pré-etiquetado. Na etiquetagem, foram empregados dois softwares desenvolvidos em Java (cada um deles embutindo uma das ferramentas escolhidas), que possuem os seguintes módulos: 1) *Pré-processamento*: identifica e reúne/separa palavras que, no CETENFolha, ora formam um único *token* do tipo: *ao contrário de* (ao_contrário_de); ora formam mais de um *token* do tipo: *da* (de a), *à* (a a), etc. 2) *Etiquetagem*: processa o QTAG/TreeTagger para etiquetar as palavras. 3) *Pós-processamento*: identifica grupos de palavras que, no CETENFolha, se apresentam de duas formas: ora aparecem reunidas em um único *token* e possuem uma única etiqueta do tipo: *do_que conj-s*, ora aparecem separadas e são etiquetadas individualmente, por exemplo: *de prp o pron-det que pron-indp*; aplica regras que fazem a junção/separação dessas palavras conforme o contexto e atribui a(s) etiqueta(s) correspondente(s). 4) *Comparação e cálculo da acurácia*: compara os pares de *palavras-etiquetas* obtidos com os pares previamente etiquetados correspondentes e identifica os pares iguais (acertos) e os pares diferentes (erros). A acurácia é calculada pela soma de todos os acertos/erros dividida pelo total de pares *palavras_etiquetas* submetidos à etiquetagem.

4. Estudo de Caso

Para a avaliação dos etiquetadores e considerando o pequeno tamanho do *corpus*, foram realizados dois tipos de experimentos por ferramenta: *Tipo 1*: corpus dividido em 20 subconjuntos de sentenças. Cerca de 95% das sentenças (19 subconjuntos) é usado para treinamento e 5% (1 subconjunto), para teste. O experimento foi repetido 20 vezes, alternando-se os subconjuntos de treinamento/teste, os quais são disjuntos. Foi calculada a acurácia média das 20 repetições. *Tipo 2*: o corpus foi dividido em 20 subconjuntos de sentenças, com 100% destas tomado para treinamento e 5%, para teste. Também repetido 20 vezes e obtida a acurácia média, como nos experimentos do Tipo 1.

Os resultados da etiquetagem dos 20 subconjuntos de teste contendo 4.004 palavras em média, são mostrados na Figura 1. As palavras mais freqüentes no total de erros com o QTAG foram: *a* (12,49%), *que* (4,84%) e *o* (2,61%) e as categorias

gramaticais: *n* (32,77%), *art* (14,32%) e *prop* (8,09%) e com o TreeTagger foram: *a* (5,72%), *que* (5,10%) e *um* (1,74%) e as categorias gramaticais: *n* (35,98%), *adj* (12,10%) e *prop* (11,50%).

4.1. Inserção de Regras

Para melhorar a acurácia da etiquetagem, foram construídas regras para resolver a ambigüidade nos casos não resolvidos pelos etiquetadores probabilísticos. Sobre as ocorrências das palavras e etiquetas mais problemáticas, foi feito um estudo de mineração de dados para a obtenção de regras com algoritmos de classificação da ferramenta Weka (Witten, 2005). As regras com melhor acurácia foram extraídas e codificadas em Java, constituindo um novo módulo do software de etiquetagem, posterior ao módulo de pós-processamento. Muitas dessas regras foram refinadas, para corrigir erros que ainda ocorreram devido a exceções nos padrões de seqüências de etiquetas. Foram implementadas 679 regras, que resolvem a ambigüidade em uma ou em uma série de ocorrências de palavras e etiquetas, em uma janela de sete posições. Um exemplo dessas regras, programado em Java, possui o seguinte formato:

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

Os resultados obtidos mostram que, com a inserção de regras, a acurácia aumenta, em média, 2,24% com o QTAG e 0,35% com TreeTagger nos experimentos do Tipo 1. Nos do Tipo 2, aumenta, em média, 0,76% com o QTAG e 0,16% com o TreeTagger (Figura 1).

4.2. Modificação nas Etiquetas

Na análise da Seção anterior, foi observado que em muitos casos não é possível formular uma regra baseando-se apenas nas palavras/etiquetas vizinhas, a menos que se tenha uma informação a mais nas etiquetas. Diante disso, propôs-se um novo conjunto de etiquetas agregando um maior número de atributos léxicos disponíveis no corpus. Após alguns experimentos, esse novo conjunto ficou definido por 161 etiquetas, originadas da combinação de 18 categorias com informações associadas de gênero, número, caso e pessoa, tempo e modo verbais. Novos experimentos, somente com o TreeTagger, mostraram que a modificação nas etiquetas aumenta a acurácia, em média, 1,53% nos experimentos do Tipo 1 e 0,57% nos do Tipo 2 em comparação com os resultados iniciais. A aplicação de um conjunto de 200 regras atualizadas com as etiquetas modificadas, aumenta a acurácia, em média, 1,63% em experimentos do Tipo 1 e 0,59% em experimentos do Tipo 2 em relação aos resultados iniciais (Figura 1).

5. Considerações Finais

Neste trabalho, foram buscadas soluções para desenvolver um etiquetador com acurácia acima de 99%. Centradas em uma abordagem híbrida, duas estratégias foram exploradas para aumentar a acurácia no processo de etiquetagem: o uso de regras e a modificação

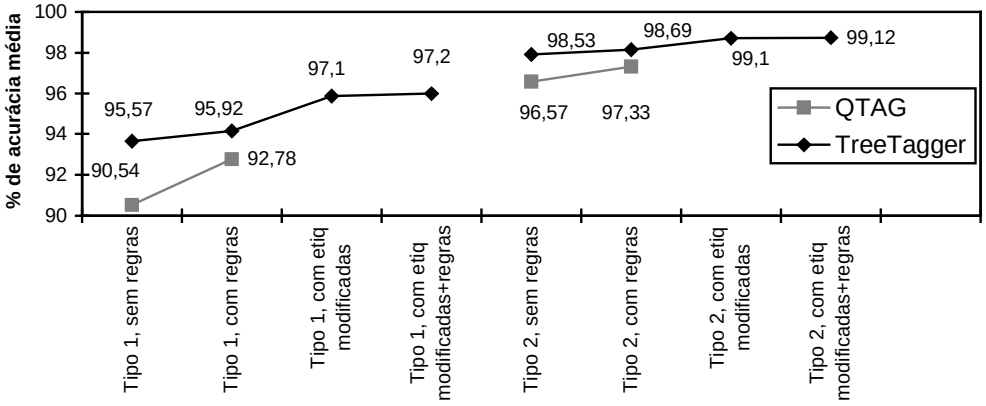

Figura 1. Resultados da etiquetagem com o QTAG e TreeTagger usando as soluções propostas.

do conjunto de etiquetas inicial do corpus. Valores próximos de 99% só foram atingidos em experimentos do Tipo 2: com poucas etiquetas (18) e muitas regras (679), o melhor resultado foi 98,69% e com muitas etiquetas (161) e menos regras (200), o melhor resultado foi 99,12%. Ressalta-se que as regras construídas até o momento cobrem os casos de ambigüidade mais freqüentes e que mais regras ainda podem ser construídas para melhorar a acurácia na etiquetagem. O TreeTagger mostrou melhor acurácia. Uma limitação do trabalho é o corpus de apenas 80.078 palavras. Nos trabalhos futuros, pretende-se expandir esse corpus com o uso do etiquetador já desenvolvido.

Referências

Bick, E. (1996). "Automatic Parsing of Portuguese", In: Anais do II Encontro para Processamento Computacional do Português Escrito e Falado. Curitiba, 1996. p. 91-100.

Harb, M. P. A. A., Brito, S. R., Silva, A. S., Favero, E. L., Tavares, O. L., Francês, C. R. L. (2003). "AmAm: ambiente de aprendizagem multiparadigmático", In: Simpósio Brasileiro de Informática na Educação. Rio de Janeiro: NCE-IM-UFRJ.

Kinoshita, J., Salvador, L. N., Menezes, C. E. D. (2006). "CoGrOO: a Brazilian-Portuguese Grammar Checker based on the CETENFOLHA Corpus", In: The fifth international conference on Language Resources and Evaluation, LREC 2006. Genova, Italy.

Linguateca (2007). "Linguateca, centro de recursos -- distribuído -- para o processamento computacional da língua portuguesa", Disponível em: http://www.linguateca.pt/. Acesso em: 12/02/2007.

Mason, O. (2006). "QTAG", Disponível em: http://www.english.bham.ac.uk/staff/omason/software/qtag.html. Acesso em: 12/06/2006.

Schmid, H. (1994). "Probabilistic part-of-speech tagging using decision trees", In: Proceedings of the Conference on New Methods in Language Processing, p. 44-49, Manchester, UK.

Witten, I.H and Frank, E. (2005). "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann, 2nd Edition, San Francisco.