FreeLing e TreeTagger:
um estudo comparativo no âmbito do Português

Pablo Gamallo, Marcos Garcia
Centro Singular de Investigação em Tecnologias da Informação (CITIUS)
Universidade de Santiago de Compostela
{pablo.gamallo@usc.es;marcos.garcia.gonzalez}@usc.es

Resumo: O presente trabalho tem como objectivo comparar duas ferramentas de anotação morfossintáctica para o Português, treinadas com o mesmo *corpus* e o mesmo léxico. Nomeadamente, comparamos os módulos de PoS-tagging para Português de FreeLing com o sistema TreeTagger para Português. Os resultados da avaliação mostram que o etiquetador morfossintáctico de FreeLing tem um nível de desempenho superior (por volta de 2 pontos de precisão em cada um dos testes realizados) ao de TreeTagger.

1 Introdução

O Processamento da Linguagem Natural engloba um vasto conjunto de tarefas que vão desde tratamentos superficiais do texto (segmentação de orações, tokenização, etc.) até análises profundas do ponto de vista sintáctico e semântico. Tarefas como a extracção e a recuperação de informação, a tradução automática, etc., podem precisar deste tipo de tratamentos, que incluam análises morfossintácticas precisas (incluindo categoria, lema, género, número, etc.) bem como *parsers* robustos que permitam analisar grandes quantidades de *corpora*.

O presente trabalho centra-se na análise morfossintáctica; mais concretamente, o objectivo principal é comparar a precisão de anotação para o Português de dois sistemas: FreeLing 3.1 e TreeTagger 3.2, treinados e avaliados com os mesmos recursos linguísticos.

A primeira das ferramentas, desenvolvida pelo Grupo TALP da Universitat Politècnica de Catalunya (Carreras *et al.* 2004; Padró e Stanilovsky, 2012), foi adaptada para Português Europeu (PE) com desempenhos em PoS-tagging próximos do estado-da-arte, variando entre 94% e 98% de precisão, em função dos *corpora* de treino e de teste (Garcia e Gamallo, 2010). O módulo utilizado para a análise morfossintáctica de FreeLing foi o HMM, com base em trigramas (Brants, 2000).

A segunda das ferramentas, TreeTagger (Schmid, 1994) foi desenvolvida no Institute for Computational Linguistics da University of Stuttgart. O método de PoS-tagging baseia-se em árvores de decisão e o seu desempenho em anteriores experimentos atinge o 96% de precisão

para o Inglês e o Alemão. Para o Português, também foi desenvolvido um *splitter* que separa as contracções e os pronomes clíticos, para além de um reconhecedor de entidades mencionadas que identifica nomes próprios multi-palavra.

2 Recursos para o treino

O *corpus* utilizado para treinar as duas ferramentas é uma adaptação do Bosque_CP 8.0[1], criado a partir de CETEMPúblico[2]. e que contém uns 138.000 tokens revisados manualmente por linguistas. Este *corpus*, compilado a partir de artigos do jornal Público, pertence à variedade PE.

Além do *corpus*, também se utilizou um dicionário de PE, que contém mais de 1.257.000 formas, e que foi extraído do léxico LABEL-LEX (SW) (Eleuterio *et al*., 2003).

Tanto o *corpus* como o léxico requereram um trabalho de adaptação das etiquetas da fonte original a um *tagset* comum (Garcia e Gamallo, 2010). O *tagset* empregado tem 255 etiquetas e baseia-se nas recomendações do Grupo EAGLES (Leach e Wilson, 1996).

3 Avaliação

Para conhecer o desempenho dos sistemas PoS-tagging treinados, foram realizadas avaliações com dois corpus de teste diferentes. Foi medida a precisão dos sistemas, que se calcula dividindo o número de tokens etiquetados correctamente pelo etiquetador entre o número total de tokens do *corpus*. Os *corpora* empregados para realizar a avaliação dos sistemas foram os seguintes:

- *Bosque_CF*
- *Miscelâneo*

O primeiro é uma adaptação do Bosque_CF 8.0, criado a partir de CETENFolha[3], que contém uns 81.000 tokens revisados manualmente por linguistas. Este corpus, compilado a partir de artigos do jornal a Folha de São Paulo, pertence à variedade do Português do Brasil (PB). O segundo *corpus* de teste, *Miscelâneo*, é um pequeno corpus de 600 tokens anotado manualmente especificamente para esta avaliação. Foi compilado a partir de textos da Wikipédia e de excertos de obras literárias portuguesas. Todos os textos deste corpus têm como variedade o PE.

Antes de realizar a avaliação é preciso ter em conta que, em Português, existem problemas de tokenização, nomeadamente existem muitos factores que podem provocar desalinhamento

1 Bosque. Uma floresta integralmente revista por linguistas: http://www.linguateca.pt/Floresta/corpus.html#bosque.
2 http://www.linguateca.pt/CETEMPublico/
3 http://www.linguateca.pt/CETENFolha/

entre o texto anotado automaticamente e o *corpus* de teste (ou *gold-standard*). Os factores que podem causar erros ou discrepâncias na tokenização derivam de problemas de detecção de entidades mencionas, e expressões multi-palavra., *splits* incorrectos de contracções, símbolos de pontuação mal reconhecidos, etc.

Os *corpora* de teste foram construídos com uma tokenização baseada no *split* das contracções e na detecção de expressões multi-palavra. e de entidades mencionadas. Estas mesmas tarefas também são levadas a cabo pelos sistemas avaliados, o que provocou o desalinhamento dos dois textos: corpus anotado automaticamente e corpus de teste. Para poder alinhar os dois textos, decidimos aplicar um algoritmo de alinhamento por ancoragem. Utilizamos âncoras para identificar trechos de texto (ou segmentos) paralelos. Uma âncora é um *token* (palavra ou símbolo) comum aos dois textos e que é utilizado para dividi-los em segmentos de longitude semelhante. Quanto mais frequente for a âncora, mais pequenos e precisos são os segmentos alinhados. Uma vez alinhados os textos em segmentos de tamanho semelhante, buscamos os *tokens* avaliáveis por segmento. Um token avaliável é aquele que aparece em dois segmentos alinhados e que, quando ocorre várias vezes no mesmo segmento, não pode ser etiquetado com diferentes *tags*, é dizer, não pode ser ambíguo nesse segmento. A precisão do sistema é calculada sobre os *tokens* avaliáveis encontrados nos textos alinhados. Deste jeito, não se contabilizam os problemas de *splitting* nem de reconhecimento de entidades.

Na Tabela 1 podemos ver os resultados das avaliações realizadas com o corpus de teste *Bosque_CF*.

Sistema	**#tokens avaliáveis**	**Precisão**
FreeLing	66.612	**93,037%**
TreeTagger	69.118	91,390%

Tabela 1: Precisão de FreeLing e TreeTagger no corpus de teste *Bosque_CF*

Os valores de precisão são inferiores aos obtidos em anteriores avaliações de FreeLing e TreeTagger. Isto é provavelmente devido ao facto de o corpus de teste pertence a uma variedade linguística (PB) diferente da do corpus de treino (PE).

Os resultados obtidos a partir do *corpus Miscelâneo* mostram-se na Tabela 2. Aqui, os resultados dos etiquetadores encontram-se próximos do estado-da-arte, situado sobre o 97%, em função dos *corpora* de treino e teste, do *tagset*, ou da língua analisada (Megyesi, 2001; Branco e Silva, 2004). Neste experimento, os *corpora* de treino e de teste pertencem a PE.

Sistema	#tokens avaliáveis	Precisão
FreeLing	532	**98,308%**
TreeTagger	529	96,030%

Tabela 2: Precisão de FreeLing e TreeTagger no *corpus* de teste *Miscelâneo*

Os dois testes mostram que o sistema FreeLing tem um desempenho superior ao de TreeTagger, cujos valores de precisão se situam 2 pontos por baixo dos do primeiro etiquetador.

4 Conclusões

O presente trabalho avalia o desempenho de dois sistemas de anotação morfossintáctica para o Português: FreeLing e TreeTagger. Os mesmos recursos linguísticos (*corpora* e dicionários) foram utilizados tanto no processo de treino como no de teste de ambos os etiquetadores.

Os resultados indicam que o módulo de PoS-tagging de FreeLing (que utiliza o algoritmo HMM) atinge melhores resultados do que TreeTagger (baseado em árvores de decisão), com diferenças de ≈2%.

FreeLing incorpora, desde a sua versão 2.0, os módulos para Português aqui avaliados.[4] Pela sua parte, a adaptação de TreeTagger para Português também é disponibilizada na página oficial do etiquetador.[5]

Referências

Branco, António e João Silva (2004): "Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese". Lino, Maria Teresa, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva (eds.) *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*: 507-510. ELRA: Paris.

Brants, Thorsten (2000): "TnT - A Statistical Part-of-Speech Tagger". *Proceedings of the 6th Conference on Applied Natural Language Processing, (ANLP 2000)*, ACL.

Carreras, Xavier, Isaac Chao, Lluís Padró e Muntsa Padró (2004): "FreeLing: An Open-Source Suite of Language Analyzers". *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*.

Eleutério, Samuel, Elisabete Ranchhod, Cristina Mota e Paula Carvalho (2003): "Dicionários Electrónicos do Português. Características e Aplicações". *Actas del VIII Simposio Internacional de Comunicación Social*: 636-642. Santiago de Cuba.

4 http://nlp.lsi.upc.edu/freeling/
5 http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Garcia, Marcos e Pablo Gamallo (2010): “Análise morfossintáctica para Português Europeu e Galego: Problemas, soluções e avaliação”. *Linguamática*, 2(2): 59-67.

Leach, Geoffrey e Andrew Wilson (1996): “Recommendations for the Morphosyntactic Annotation of Corpora”. Relatório Técnico. Expert Advisory Group on Language Engineering Standard (EAGLES).

Megyesi, Beáta (2001): “Comparing Data-Driven Learning Algorithms for PoS Tagging of Swedish”. *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing:* 151-158.

Padró, Lluís e Evgeny Stanilovsky (2012) , “Freeling 3.0: Towards wider multilinguality”, *Proceedings of Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey.

Schmid, Helmut (1994): “Probabilistic Part-of-Speech Tagging Using Decision Trees”. *Proceedings of the International Conference on New Methods in Language Processing:* 44-49.