Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Nuevas fuentes de información para entrenamiento de etiquetadores gramaticales

Propuesta de Tesis presentada para optar al título de
Licenciado en Ciencias de la Computación

Fernando Jorge Rodriguez
ferrod20@gmail.com
LU: 516/00

Director: Dr. José Castaño

Buenos Aires, 2012

1. Introducción

El etiquetado o anotado gramatical, también conocido como Part-of-speech tagging, POS tagging o simplemente POST, es el proceso de asignar una etiqueta gramatical a cada una de las palabras de un texto según su categoría léxica. Por ejemplo tomemos la oración siguiente:

There is no asbestos in our products now.

El resultado de etiquetarla gramaticalmente es:

There/EX is/VBZ no/DT asbestos/NN in/IN our/PRP products/NNS now/RB ./.

donde cada palabra está sucedida por una barra oblicua seguida de la etiqueta gramataical asignada.

Un tagger o etiquetador es un programa que realiza este proceso automáticamente. La mayoría de los taggers actuales utilizan modelos estadísticos. Estos modelos se nutren entrenando el tagger con un texto anotado previamente (corpus de entrenamiento). El rendimiento del tagger (que se mide por las etiquetas asignadas correctamente) es fuertemente dependiente del corpus de entrenamiento utilizado.

El problema reside en que la generación de corpus de entrenamiento es una tarea muy costosa, por lo tanto la cantidad y calidad de los mismos es limitada. La idea central de esta tesis es la de suplir la falta de corpus de entrenamiento generando una nueva fuente de información a partir de una fuente de información existente: un diccionario.

Generalmente un diccionario contiene la definición de una palabra, una explicación de su significado, algunas características como su pronunciación y particularmente su clase gramatical y uno o más ejemplos que muestran su uso. Por lo tanto, extrayendo todos los ejemplos de un diccionario, se puede generar un corpus anotado parcialmente, es decir, un conjunto de oraciones donde alguna/s de las palabras que comprenden cada oración posee/n una etiqueta gramatical.

2. Descripción de la propuesta

El objetivo de este trabajo como se mencionó anteriormente es generar una nueva fuente de información a partir de un diccionario y luego utilizarla como corpus de entrenamiento intentando mejorar el rendimiento de un tagger.

El diccionario elegido para el trabajo es Cobuild: Cobuild es un diccionario de la lengua inglesa basado en la información del corpus Bank of English y el corpus Collins. Todos los ejemplos del diccionario Cobuild muestran patrones gramaticales típicos, vocabulario típico y contextos típicos para cada palabra. En consecuencia, Cobuild presenta una cantidad exhaustiva del vocabulario inglés derivado de observaciones directas del lenguaje.

2.1. Primer etapa

La primer etapa consiste en extraer cuidadosamente los ejemplos (que son oraciones completas conteniendo la palabra definida con información léxica) del diccionario Cobuild y generar a partir de estos un corpus parcialmente anotado. Luego se completará el etiquetado del corpus utilizando un etiquetador automático.

2.2. Segunda etapa

La segunda etapa consiste en entrenar los taggers con este nuevo corpus de entrenamiento, etiquetar y luego medir los resultados.

Los taggers que se utilizarán para el entrenamiento y medición son:

- Trigrams'n'Tags tagger (estocástico)
- Stanford tagger (tagger de máxima entropía).

Los corpus elegidos para realizar la medición son:

- Wall Street Journal
- British National Corpus

Ambos pertenecientes a la lengua inglesa (americana y británica respectivamente).

British National Corpus es una colección de 100 millones de palabras provenientes de textos escritos e inglés hablado creado en los 1990s por un consorcio de editores, universidades (Oxford y Lancaster) y la British Library.

Wall Street Journal es un corpus anotado, parte del Penn Treebank, de algo más de 1 millón de palabras.

Referencias

[1] Jurafsky, D. Martin, J. H., Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition, Second edition, chapter 5, New Jersey: Prentice Hall.

[2] Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999

[3] Thorsten Brants, TnT: a statistical part-of-speech tagger, Proceedings of the sixth conference on Applied natural language processing, p.224-231, April 29-May 04, 2000, Seattle, Washington

[4] Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

[5] Mitchell P. Marcus , Mary Ann Marcinkiewicz , Beatrice Santorini, Building a large annotated corpus of English: the penn treebank, Computational Linguistics, v.19 n.2, June 1993

[6] Reference Guide for the British National Corpus (World Edition) edited by Lou Burnard, October 2000

[7] Stevenson M., A corpus-based approach to deriving lexical mappings, EACL '99 Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, Pages 285-286

[8] Brown K. (Editor) 2005. Encyclopedia of Language and Linguistics 2nd Edition. Oxford: Elsevier.

[9] Sinclair, J. 'The automatic analysis of corpora', in Svartvik, J. (ed.) Directions in Corpus Linguistics (Proceedings of Nobel Symposium 82). Berlin: Mouton de Gruyter. 1992.