ETIQUETADO INICIAL DE CORPUS EN ESPAÑOL

Alfonso Pereda Gálvez*

RESUMEN

Uno de los grandes problemas que tiene todo usuario de computadoras, es la falta de herramientas que le permitan hacer búsquedas de información en distintas fuentes de datos de Documentos. Cuán útil sería contar con herramientas que permitan la clasificación, búsqueda de ideas, y contenidos en esta gran base de datos de documentos. En el Procesamiento Natural del Lenguaje, NLP (Natural Language Processing), uno de los temas fundamentales es el Etiquetado de Palabras (Part of Speech Tagging: POS Tagging), encargado de asignar una categoría sintáctica a las palabras de un texto de un lenguaje natural. El valor agregado de la categoría sintáctica en los documentos (textos) permite realizar búsquedas de información inteligentes. Se presenta aquí la metodología general para implementar un etiquetador inicial de palabras en un Lenguaje Natural como el Español.

1. ANTECEDENTES.

El Procesamiento de Lenguaje Natural (NLP) es el área de la computación que contribuye con un conjunto de métodos y técnicas para el procesamiento de textos. Tiene aplicación en una gran variedad de campos como: extracción del conocimiento léxico, extracción de información, indización de la información, recuperación de la información (Information Retrieval), construcción de diccionarios de términos (Márquez, 1999]. Adicionalmente, provee de "benchmarks" para la evaluación de estudios teóricos y modelos del lenguaje utilizado en corpus específicos.

Uno de los temas del NLP es el etiquetado automático de palabras de un corpus, el cual analizaremos. Un corpus es un conjunto de textos de un lenguaje natural que describe un contexto del mundo y es empleado para observar características del lenguaje. Un conjunto de corpus se denomina corpora.

El etiquetado de palabras es la tarea de asignar apropiadamente la categoría sintáctica a las palabras de un texto; conocido en Inglés como Part of Speech Tagging (POS Tagging). Parte de la problemática del etiquetado es la ambigüedad en la asignación de la categoría sintáctica apropiada a ciertas palabras del corpus, que pueden tener más de una etiqueta debido al contexto en el que se encuentren. El POS Tagging

es de interés para muchas aplicaciones, como por ejemplo: reconocimiento y generación de palabras (Heeman and Allen, 1997), acceso a bases de datos textuales (Kupiec, 1993), análisis sintáctico parcial y general (Karlsson et al., 1995), inferencia gramatical, extracción de información, recuperación de información, etc. El etiquetado de palabras se considera un preproceso importante en la recuperación de información (Krovetz, 1997).

Uno de los paradigmas prevalecientes en el etiquetado automático de palabras, es la construcción de etiquetadores aplicando técnicas de entrenamiento. Existen dos métodos de entrenamiento: Supervisado y No Supervisado.

El etiquetado automátìco con entrenamiento supervisado, implica utilizar un corpus etiquetado, en donde la categoría sintáctica de cada palabra es conocida a priori. Por el contrario, el método de etiquetado con entrenamiento no supervisado no requiere de un corpus etiquetado para el aprendizaje, el conocimiento que adquiere este sistema esta basado en otras fuentes del lenguaje, como por ejemplo los diccionarios o textos bilingües [Manning and Shütze, 1999].

Hay numerosos trabajos recientes que tienen relación con el tema.

2. FUNDAMENTO DEL ETIQUETADO DE PALABRAS (POS TAGGING)

El etiquetado de palabras se refiere a la asignación de categorías sintácticas a cada palabra de un texto o corpus, denominado en inglés Part Of Speech Tagging (POS Tagging).

El POS Tagging requiere de un conjunto de etiquetas pre-definidas (tagset) y un algoritmo de asignación de etiquetas.

Existen dos principales técnicas para el etiquetado de palabras, una basada en reglas y la otra en técnicas probabilísticas. De estos dos grupos de técnicas existen algunos métodos que emplean elementos de ambos.

El primer paso en cualquier proceso de etiquetado es buscar la palabra a ser etiquetada en un Léxicon o Diccionario. Si la palabra no puede encontrarse, el etiquetador (tagger) tiene que tener algún mecanismo de retroalimentación (fallback), como un componente morfológico o algunos métodos heurísticos para resolver. La

tarea difícil es tratar con las ambigüedades: sólo en los casos triviales será la asignación exacta de una etiqueta por cada palabra.

Estas técnicas difieren en que, mientras el etiquetado basado en reglas aplica conocimiento lingüístico (usualmente incluido en los algoritmos del etiquetador), un etiquetador probabilístico determina cuáles son las posibles secuencias, usando un modelo del lenguaje basado en frecuencias de transiciones entre los diferentes etiquetados de las palabras.

3. ALGORITMO DE UN ETIQUETADOR INICIAL

El etiquetado de palabras es el proceso de asignar una etiqueta (tag) u otro marcador de la clase léxico a cada palabra en un corpus.

La entrada en el algoritmo de etiquetado es una cadena de palabras y un conjunto de etiquetas especificadas (tagset) denominado LEXICON. La salida es la mejor etiqueta para cada palabra del texto (corpus).

El proceso de etiquetado inicial es:

1. Leer la siguiente palabra
2. Ver en el Lexicon
3. Si no se encuentra, almacenar la palabra en un archivo de WORDUNTAG y asignarle una etiqueta DUMMY.
4. Por cada posible etiqueta de la palabra:
5. calcular Pw = P(tag\word), la probabilidad de cada palabra tiene la dependencia del tag.
6. calcular Pc = P(tag\t1,t2), la probabilidad del tag dada la lista de tags t1 y t2.
7. calcular P w,c = Pw x Pc, la unión de probabilidades dará como resultado la visión contextual de probabilidad.
8. Repetir el proceso por cada palabra del CORPUS.

4. RECURSOS EMPLEADOS

Lexicon

Es un archivo de texto que contiene el conjunto de palabras y sus correspondientes etiquetas en orden de frecuencia.

WORDUNTAG

Es un archivo de texto con las palabras que no tienen un tag en el lexicon.

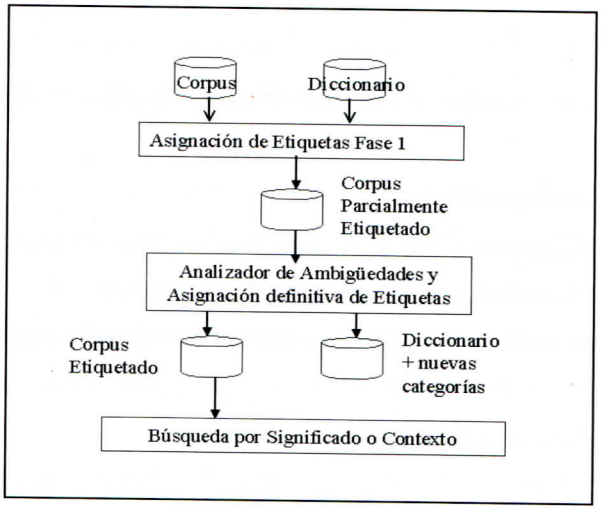
Hábeas

Es el archivo de Texto en un Lenguaje Natural, sin etiquetar.

Frecuencias o Probabilidades

Todas las frecuencias son determinadas en base a las palabras inequívocamente etiquetadas.

5. DIAGRAMA DE UN ETIQUETADOR DE TEXTOS



El etiquetador inicial se ubica en la FASE 1 del Diagrama.

6. METODOLOGÍA PARA LA CONSTRUCCION DEL ETIQUETADOR INICIAL

Empleo de técnicas estadísticas descritas en NLP como UNIGRAMS (uso de probabilidades).

Utilización de recursos y conocimientos del lenguaje natural en el caso del Español, los diccionarios, estructuras gramaticales, etc.

Uso de algoritmos de "bootstrapping" para lograr la convergencia en las técnicas de etiquetado.

Aplicación de la metodología de construcción de sistemas informáticos.

Utilización de técnicas estadísticas para obtener una medida de la precisión y completitud del etiquetado.

7. CONCLUSIONES

- El Procesamiento de Lenguaje Natural (NLP) es el área de la computación que contribuye con un conjunto de métodos y técnicas para el procesamiento de textos.

- El etiquetado de palabras es un tema fundamental de NLP y se refiere a la asignación de categorías sintácticas a cada palabra de un texto o corpus, denominado en inglés Part Of Speech Tagging (POS Tagging).

- Un modelo formal de etiquetado inicial, descrito en el ALGORITMO.

- La definición del estado inicial del etiquetado de textos es empleado en el etiquetado automático de palabras.

- Existen muchos temas que constantemente se estan investigando y desarrollando en el área de ETIQUETADO AUTOMATICO DE PALABRAS.

- Los problemas que se identifican en el proceso de etiquetado inicial son:

- Palabras sin categoría sintáctica (WORDUNTAG).
- Asignación de más de una categoría sintáctica a algunas palabras. (Ambigüedad : VINO / VB_SUST verbo y sustantivo)

BIBLIOGRAFÍA

Aone, C. and Hausman, K. (1996). "Unsupervised Learning of a Rule based Spanish Part of Speech Tagger." In: Proceedings of the International Conference on Computational Lingustics (COLING).

Becker, M. (1998). "Unsupervised Part of Speech Tagging with Extended Templates". In: Proceedings of the European Summer School for Logic, Language and Information (ESSLLI), Student Session,

Brill, E. (1992). "A Simple Rule Based Part of Speech Tagger". In: Proceedings of the Third Conference on Applied Natural Language Processing (ANLP),

Brill, E. (1994). "Some Advances in Rule Based Part of Speech Tagging". In: Proceedings of the Twelfth National Conference on Artificial Intelligence.

Brill, E. (1995). "Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging". In: Proceedings of the 3rd Workshop on Very Large Corpora,

Church, K. W. (1988). "Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text." In: Proceedings of the 2nd Conference on Applied Natural Language Processing (ANLP),

Golding, A.R. (1995) "A Bayesian-hybrid Method for context-sensitive Spelling Correction". In: Proceedings of 3rd Workshop on very large Corpora.

Heeman, P.A. and Allen, J.F. (1997). "Incorporating POS Tagging into Language Modeling". In: Proceedings of Eurospeech Conference. ALLC/ACH Conference,

Karlsson F, Voutilainen A., Heikkila, J. and. Anttila, A. eds.. (1995). Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text. **Mouton de Gruyter.**

Krenn, B. and Samuelsson, C. (1997). **The Linguist's Guide to Statistics.** http://www.coli.uni-sb.de/~christerg.

Krovetz, Robert (1997). "Homonymy and Polysemy in Information Retrieval". In: Proceedings of the 35th Annual Meeting of Associaton for Computational Linguistics. Joint ACL/EACL,, **Madrid, July 1997. pp. 72-79**

Kupiec, J. and Murax (1993). "A Robust Linguistic Approach for Question Answeering Using an Online Enciclopedia". In: Proceedings of SIGIR '93, **pp. 181-190,**

Manning and Shütze, (1999). Foundations of Statistical Natural Language Processing, **Cambridge, MIT Press,**

Márquez, Lluís (1999). Part Of Speech Tagging: A Machine Learning Approach based on Decision Trees. **PhD. Thesis, Dep. Llenguatges I Sistemes Inormátics. Universitat Politécnica de Catalunya, Spain May 1999.**

Nelson, F. W. and Kucera, H. (1979). Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers. **Technical report, Department of Linguistics, Brown University, Providence,**

Piskorski, J. (1999). Finite-State Machine Toolkit. **Technical report, DFKI, Germany,**

Pop, M. (1996). Unsupervised Part-of-Speech Tagging. **Technical report, Johns Hopkins University, Baltimore, 1996.**

Rabiner, LR. (1990). "A tutorial on hidden Markov models and selected applications in speech recognition". Readings in Speech Recognition. **Morgan Kaufmann Publishers, San Mateo, CA,**

Roche, E and Schabes, Y. (1996). "Deterministic Part of Speech Tagging with Finite State Transducers". In: Proceedings of the International Conference on Computational Lingustics (COLING),

Samuelsson, C. (1996). "Handling Sparse Data by Successive Abstraction". In: Proceedings of the 16th International Conference on Computational Linguistics (ICCL),

Schapire, Robert E., Singer, Yoran and Singhal, Amit (1998). "Bootting and Rocchio applied to text filtering". In: Proceedings of SIGIR '98.

Schmid, H. and Kempe, A. (1996). "Tagging von Korpora mit HMM, Entscheidungsbäumen und Neuronalen Netzen". In: H. Feldweg and E. W. Hinrichs, eds., Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschliessung des Deutschen,

Tapanainen, P. (1996). The Constraint Grammar Parser CG-2. **Technical report, Department of General Linguistics, University of Helsinki, 1996.**

Volk, M. and Schneider, G. (1998). "Comparing a statistical and a rule-based tagger for German". In: Proceedings of KONVENS-98.