Etiquetación automática en corpus textuales cubanos. Primeros resultados.

Leonel Ruiz Miyares

Centro de Lingüística Aplicada. Ministerio de Ciencia, Tecnología y Medio Ambiente.
Apartado Postal 4067. Vista Alegre. Santiago de Cuba. Cuba. 90400
E-mail: leonel@lingapli.ciges.inf.cu

Abstract

In the last years Cuban linguists have been working on the research on the textual corpora to study the vocabulary of school children and adults. The main problem that they came up against was the absence of the necessary tools to develop their research.
Since 1997 the Group of Computational Linguistics from the Center of Applied Linguistics of Santiago de Cuba has been working on the development of the tagger to help the linguists in their lexicogology and lexicographic researches.
This paper describes the first Cuban probabilistic tagger developed on the basis of Hidden Markov Model (HMM) for the analysis of any textual corpora.
The paper shows the results and the performance of the tagger after its application to textual corpora of students from secondary level.

Resumen

En los últimos años los lingüistas cubanos han trabajado en el estudio del vocabulario de adultos y escolares. El principal problema que tuvieron estos especialistas durante sus investigaciones fue la ausencia de las necesarias herranientas informáticas para desarrollar sus estudios.
Desde 1997 el Grupo de Lingüística Computacional del Centro de Lingüística Aplicada de Santiago de Cuba viene trabajando en el desarrollo de un etiquetador gramatical automático para ayudar a los lingüistas en sus estudios lexicológicos y lexicográficos.
El presente trabajo describe las características del primer etiquetador gramatical cubano, desarrollado sobre la base de los Modelos Ocultos de Markov para el análisis de cualquier corpus textual.
La ponencia presenta los resultados y la efectividad del etiquetador después de su aplicación a un corpus textual de estudiantes de secundaria básica.

Keywords: Tagger, lexicon, probabilistic model.

1. Introducción

El Centro de Lingüística Aplicada de Santiago de Cuba comenzó a aplicar las técnicas de computación en la última etapa del *Estudio del Vocabulario del Escolar Cubano* (nivel primario) llevado a cabo entre los años 1973 y 1975.

En 1990 dicho Centro comenzó una nueva investigación nacional acerca del vocabulario activo funcional del escolar cubano de 2do. a 6to. grados (7-11 años, respectivamente), con el fin de, sobre la base de sus resultados, poner al alcance de nuestros niños un diccionario escolar**.**

Para emprender el análisis del *Léxico Activo Funcional del Escolar Cubano* con vistas a preparar el *Diccionario Escolar*, el Grupo de Lingüística Computacional del Centro de Lingüística Aplicada confeccionó el *Sistema de Computación VEC* (Vocabulario del Escolar Cubano).

Todas las palabras de las composiciones que se procesaron en las computadoras fueron codificadas manualmente por los lingüistas y cada una de ellas recibió la categoría gramatical correspondiente a la función que realizaba en el contexto analizado, según los códigos creados para ese fin, luego las composiciones ya codificadas se introdujeron en las computadoras.

El software analizó 7000 composiciones, de ellas 5873 fueron escritas y 1127 orales. El corpus ascendió a 692 814 palabras, donde 461 299 pertenecen al corpus escrito y 231 515 al corpus oral.

A pesar de todo el apoyo brindado por el *Sistema de Computación VEC,* a los investigadores del Grupo de Lingüística Computacional del Centro de Lingüística Aplicada les quedaba la insatisfacción de la codificación manual de los textos a procesar por este sistema.

A este gran problema se sumaron, además, los errores humanos durante la codificación e introducción de las categorías gramaticales de las composiciones y la falta de este tipo de estudios en Cuba por la carestía de softwares apropiados para desarrollar éstos.

Por todo lo anterior, se decidió emprender los pasos necesarios para crear en Cuba el primer etiquetador gramatical automático de corpus textuales.

2 El sistema computacional

El punto de partida para crear el etiquetador es la investigación *Léxico Activo Funcional del Escolar Cubano*. Ese trabajo aportó toda la información necesaria para desarrollar las herramientas del futuro tagger.

Sobre la base del material disponible, se decidió que el etiquetador fuera supervisado, probabilístico, utilizara bigramas y aplicara los modelos ocultos de Markov para la desambiguación.

2.1. Las etiquetas

El sistema que describimos aplicará el mismo conjunto de etiquetas utilizado en la investigación del léxico del escolar cubano, puesto que consideramos que es bastante completo y abarca todo el espectro de accidentes gramaticales existentes en el idioma español.

Un fragmento del conjunto de etiquetas se refleja en la Fig. 1 y éste está compuesto por 36 categorías gramaticales dentro de las cuales existen:

- 9 etiquetas para los sustantivos
- 5 etiquetas para los adjetivos
- 7 etiquetas para los pronombres
- 7 etiquetas para los verbos
- 8 etiquetas para el resto de las categorías gramaticales (artículo, adverbio, preposición, conjunción, interjección, contracción, lexías complejas y fechas y las siglas)

010 Artículo
020 Sustantivo Propio
021 Sustantivo Común Masculino Singular
022 Sustantivo Común Femenino Singular
023 Sustantivo Común Masculino Plural
024 Sustantivo Común Femenino Plural
025 Sustantivo Diminutivo
026 Sustantivo Aumentativo
027 Sustantivo Propio (Nombres Geográficos)
028 Sustantivo Propio (Juegos Infantiles)
030 Adjetivo Calificativo Antepuesto
031 Adjetivo Calificativo Pospuesto
032 Adjetivo Determinativo
033 Adjetivo Diminutivo
034 Adjetivo Aumentativo
040 Pronombre Personal
041 Pronombre Demostrativo
042 Pronombre Posesivo
043 Pronombre Indefinido
044 Pronombre Relativo
045 Pronombre Interrogativo y Exclamativo
046 Variante Pronominal

Fig. 1. Fragmento del conjunto de etiquetas

2.2. El lexicón

El cuerpo principal del lexicón lo conforma todas las voces del vocabulario del escolar, con su respectiva clasificación gramatical, además de poseer cada palabra la frecuencia de su uso en el corpus del escolar cubano, la cual se utilizará en el proceso de desambiguación para aquellas palabras que lo requieran.

En la actualidad el lexicón posee 51467 palabras, cantidad importante que robustece al etiquetador durante el procesamiento no sólo de corpus escolares, sino también de otros corpus textuales.

Al lexicón se le han añadido también otras palabras fuera del léxico escolar, como por ejemplo todas las palabras del *Diccionario Ortográfico del Español* desarrollado por el Centro de Lingüística Aplicada de Santiago de Cuba y el Instituto para los Circuitos Electrónicos de Génova, Italia, además de las palabras resultantes de la aplicación del etiquetador en su primera versión a otras investigaciones.

Por último señalaremos que existe un campo en la base de datos del lexicón donde se reflejan las características semánticas de las palabras complejas en su significado. No todas las palabras tienen información en ese campo, sólo aquellas como *carata* (botánico), *richter* (sismología), *mira* (militar), lb-12 (gas ecológico), etc. que pueden resultar dudosas para los lingüistas. Este aspecto es de gran utilidad para los lexicólogos en sus estudios de corpus textuales.

2.3. Entrenamiento del etiquetador

Como se ha explicado, todo el corpus del léxico del escolar cubano fue codificado manualmente; esta tarea fue bastante ardua, consumió mucho tiempo pero se trabajaba para el futuro sin saberlo.

La codificación manual del corpus fue revisada minuciosamente por el equipo de lingüistas que llevó adelante aquel estudio. Todo ese material fue decisivo en la construcción del etiquetador supervisado.

2.3.1 Matriz de probabilidades de transiciones

A partir de la información obtenida, se realizaron los respectivos cálculos estadísticos de los bigramas y de esta forma se confeccionó la matriz de probabilidades de transiciones, según la fórmula:

$$P(t_i/t_{i-1}) \approx \frac{f(t_{i-1}, t_i)}{f(t_{i-1})}$$

donde $f(t_{i-1}, t_i)$ es la frecuencia de ocurrencia de la pareja de etiquetas t_{i-1}, t_i y $f(t_{i-1})$ es la frecuencia de ocurrencia de la etiqueta t_{i-1}.

Esta matriz juega un importante papel durante la desambiguación de las palabras que posean más de una etiqueta.

2.3.2 Matriz de probabilidades de observación

En los modelos ocultos de Markov surge la matriz de probabilidades de observación la cual es aquella que calcula la probabilidad de ocurrencia de una palabra dada una etiqueta. Esta matriz se representa según la fórmula:

$$P(w_i/t_i) = \frac{f(w_i, t_i)}{f(t_i)}$$

donde $f(w_i, t_i)$ es la frecuencia de la palabra w_i con la etiqueta t_i y $f(t_i)$ es la cantidad de palabras con la etiqueta t_i.

Para desambiguar las palabras homónimas el etiquetador escoge el mayor resultado de las multiplicaciones de la matriz de transiciones y la matriz de probabilidades de observación para cada caso específico:

$$\max \prod_i^n P(w_i/t_i) P(t_i/t_{i-1})$$

2.4. Funcionamiento

Las partes que posee el etiquetador son las siguientes:

1. El segmentador
2. El analizador-desambiguador
3. Procesamiento de palabras desconocidas

2.4.1 El segmentador

Los textos primeramente son analizados en el segmentador, cuya función es la siguiente:

- dividir en palabras las oraciones del texto introducido directamente desde el teclado o desde un fichero previamente editado (siempre en formato ASCII)
- reconocer las cifras dentro del texto (Ej. 9759, 5,23, 98.5%, $5000, etc.)
- reconocer los signos de puntuación y signos especiales dentro del texto (. , ; : " ' ( ) ¿ ? ¡ ! - # $ & % /)

2.4.2 El analizador-desambiguador

Las palabras del segmentador pasan al analizador-desambiguador y éste por su parte realiza los siguientes pasos:

1. Busca la palabra en el lexicón
2. Le asigna la etiqueta a la palabra
3. Si a la palabra se le puede asignar más de una etiqueta, a través de la técnica de los modelos ocultos de Markov se halla la etiqueta más probable.
4. Si la oración comienza con una palabra ambigua, la etiqueta de mayor frecuencia en el corpus de entrenamiento es la que se asignará a la palabra.

En esta parte del etiquetador también se reconocen:

1. Unidades léxicas complejas:

- nombres propios (José Martí, Ernesto Che Guevara, etc.)
- nombres geográficos (Las Tunas, Cartagena de Indias, etc.)
- juegos infantiles (El Canguro Saltador, El Coge Coge, etc.)
- lexías complejas (aula magna, astro rey, ciencia ficción, etc.)
- frases adverbiales (a diario, con anterioridad, día a día, etc.)
- frases prepositivas (a cambio de, delante de, respecto a, etc.)
- frases conjuntivas (desde que, por eso, sin embargo, etc.)
- verbos modo indicativo tiempos compuestos (ha ido, han salvado, había puesto, etc.)
- verbos modo subjuntivo tiempos compuestos (haya recibido, hubiera besado, hayan prestado, etc.)
- perífrasis verbales (estuviera bailando, fue cumpliendo, etc.)

2. Siglas:

- JADT, FAO, CITMA, MINSAP, ONU, UNESCO, U.R.S.S., URSS, etc.

3. Abreviaturas:

- p.m., a.m., am, dr., excmo, ilmo, etc.

2.4.3 Procesamiento de palabras desconocidas

En el caso de que la palabra a etiquetar no aparezca en el lexicón ésta se considera desconocida.

El procedimiento para asignarle la etiqueta correspondiente es como sigue:

1. Se buscan en la base de datos de sufijos las últimas nueve letras de la palabra (o menos, en dependencia de su longitud)

2. En el caso de que el sufijo buscado tenga asociada sólo una etiqueta, entonces a la palabra desconocida se le asigna esa categoría gramatical.
3. Si el sufijo buscado tiene asociada más de una etiqueta, se analiza cuál es la etiqueta más probable de ocurrir conjuntamente con la etiqueta anterior.
4. Si la terminación de la palabra no aparece en la base de datos de los sufijos, entonces la palabra desconocida puede tener cualquiera de las 36 etiquetas, exceptuando a aquellas de inventario cerrado, es decir, aquellas etiquetas cuyo uso es de alta frecuencia y se sobreentiende que sus respectivos vocablos están en el lexicón (contracciones, pronombres, preposiciones, conjunciones, etc.). En este caso se le asigna a la palabra desconocida la etiqueta que tenga mayor probabilidad de aparición conjuntamente con la anterior.

3. Resultados de la aplicación del etiquetador a un corpus textual de secundaria básica

Para comprobar la efectividad del sistema computacional que analizamos aplicamos éste a un corpus escrito de estudiantes de secundaria básica (7mo., 8vo. y 9no. grados, 12-14 años, respectivamente) donde se obtuvieron resultados muy interesantes que a continuación expondremos.

El total de textos analizados fue de 210, los cuales incluyeron 26747 palabras.

La muestra por provincias es la siguiente:

- Santiago de Cuba: 110 textos, 55 hembras y 55 varones de 7mo. grado
- Holguín: 30 textos, 15 hembras y 15 varones de 7mo. grado
- Ciego de Ávila: 30 textos, 15 hembras y 15 varones de 8vo. grado
- Ciudad de La Habana: 40 textos, 20 hembras y 20 varones de 9no. grado

En las tablas se puede observar la alta efectividad del etiquetador, el cual alcanza el 98.24%.

Un aspecto a resaltar en esta investigación es el tiempo de procesamiento de la información.

Según el estudio del léxico del escolar cubano, donde las composiciones se codificaron de forma manual, el tiempo de introducción promedio por operador fue de 12 composiciones por jornada laboral. Si aún se utilizara el *Sistema de Computación VEC* se necesitarían 17.5 días para procesar las 210 composiciones empleadas para probar la eficacia del etiquetador. Sin embargo, el tagger que aquí describimos sólo necesita 29 minutos para analizar las 26747 palabras pertenecientes a los 210 textos.

El ahorro de tiempo y la humanización del trabajo de los lexicólogos y lexicógrafos es sustancial durante la aplicación de este etiquetador.

Resultados de la etiquetación automática:

Palabras nuevas	439	Correctas	362	**82.6 %**
		Incorrectas	77	17.5 %

Total de palabras en el corpus	Errores del etiquetador	Efectividad
26747	472.5	**98.24 %**

4. Conclusiones

1. El software desarrollado ha demostrado su alta efectividad en el procesamiento de corpus textuales.
2. La información semántica en el lexicón resulta un rasgo interesante en este etiquetrador.
3. El tiempo de procesamiento de la información se reduce drásticamente al utilizarse este nuevo sistema.
4. La introducción de este software humaniza el trabajo de los investigadores y operadores.
5. Los lingüistas cubanos cuentan ya con un instrumento útil para sus investigaciones.

Bibliografía

Alphen P. (1992). *HMM-based continuous-speech recognition. Systematic evaluation of various system components***.** Doctoral thesis, University of Amsterdam, The Netherlands.

Buitelaar P. (1997). A Lexicon for Underspecified Semantic Tagging. http://xxx.lanl.gov/ps/cmp-lg/9705011

Charniak E. (1993). *Statistical language learning*. Massachusetts Institute of Technology, Massachustts, United States of America.

Garside R. et al. (1987). *The computational analysis of English*, Longman.

Marconi L. et al. (1999). Características generales del Diccionario Ortográfico del Español en *Actas del VI Simposio Internacional de Comunicación Social*, Santiago de Cuba, 25-28 de enero de 1999. Ediciones Editorial Oriente, Centro de Lingüística Aplicada y el Consiglio Nazonale delle Riceche, páginas 130-135.

Márquez Ll. y Padró Ll. (1998). Etiquetado morfosintáctico de corpus textuales en *Actas del Congreso Anual de la Asociación Española de Lingüística Aplicada* (AESLA'98), Logroño, España. http://www.lsi.upc.es/ ~padro/

Miyares E. (1997). Algunas consideraciones acerca del Diccionario Escolar Ilustrado. *Estudios de Comunicación Social*, Editorial Academia, La Habana, páginas 17-22.

Miyares E. (1998). *Diccionario Escolar Ilustrado*. Editorial Oriente, Santiago de Cuba y Ediciones Libertarias Prodhufi, Madrid, España.

Moreno-Torres I. (1994). Desambiguación morfológica: una aproximación híbrida en *Actas del X Congreso de Lenguajes naturales y lenguajes formales*, Universidad de Sevilla, España, páginas 479-486.

Nijholt A. (1992). Linguistic Engineering: A Survey, *Proceedings of the Second Twente Workshop on Language Technology*, University of Twente, Enschede, The Netherlands, pages 1-22.

Padró Ll. (1997). *A hybrid environment for syntax-semantic tagging*. Tesis de doctorado, Universidad Politécnica de Cataluña, España.

Paulussen H. (1992). *Automatic grammatical tagging: description, comparison and proposal for augmentation*. Universidad de Antwerpen, Wilrijk, Bélgica.

Ruiz L. (1994). Aplicación de la computación al estudio del vocabulario básico del escolar cubano. *Estudios de Comunicación Social*, Editorial Academia, La Habana, páginas 96-105.

Ruiz L. (1997a). Versión avanzada de un sistema computacional aplicado a una investigación lexicológica. *Estudios de Comunicación Social*, Editorial Academia, La Habana, páginas 85-113.

Ruiz L. (1997b). Development of two probabilistic morphological taggers for Spanish corpus. Evaluation. Internal Report, University of Twente, Enschede, The Netherlands.

Ruiz L. (1999). Primeros pasos de la etiquetación automática en Cuba en las *Actas del VI Simposio Internacional de Comunicación Social*, Santiago de Cuba, 25-28 de enero de 1999. Ediciones Editorial Oriente, Centro de Lingüística Aplicada y el Consiglio Nazonale delle Riceche, páginas 710-714.

Ruiz V. y Miyares E. (1984). *El consonantismo en Cuba*, Editorial Ciencias Sociales, La Habana.

Sánchez F. (1987). El etiquetado del Corpus de Referencia del Español Actual (CREA). Seminario Internacional de Industrias de la Lengua, Soria, España.

Sánchez F. y Nieto A. (1995). Development of a Spanish version of the Xerox tagger. http://xxx.lanl.gov/ps/cmp-lg/9505035.

Segura J. (1991). *Modelos de Markov con cuantización dependiente para reconocimiento de voz*. Tesis de doctorado, Universidad de Granada, España.