LISETTE GARCIA MOYA[1]
AURORA PONS PORRATA[1]
LEONEL RUIZ MIYARES[2]
YAMILA COBOS CASTILLO[3]

[1]**Centro de Estudios de Reconocimiento de Patrones y Minería de Datos**
Santiago de Cuba, Cuba
{lisette, aurora}@csd.uo.edu.cu

[2]**Centro de Lingüística Aplicada, CITMA**
Santiago de Cuba, Cuba
leonel@lingapli.ciges.inf.cu

[3]**Diseño de Aplicaciones, Tecnologías y Sistemas**
Ciudad de La Habana, Cuba
yamila.cobos@datys.co.cu

Una propuesta de etiquetador morfosintáctico para el español

Resumen

En este trabajo se presenta un etiquetador morfosintáctico para el español hablado en Cuba, que combina los modelos ocultos de Markov basado en bigramas con diversas heurísticas y diccionarios para asignar la categoría gramatical de cada palabra según el contexto en que aparezca. Se apoya, además, en un analizador morfológico para obtener todas las posibles interpretaciones morfológicas de una palabra, lo cual permite acotar el conjunto de las posibles categorías gramaticales y obtener no sólo la categoría correcta, sino también toda su información morfológica. El etiquetador propuesto alcanza una precisión de 96.6% en un corpus jurídico.

1 Introducción

En la actualidad, la mayor parte de la información disponible se presenta de forma no estructurada, por ejemplo, las noticias digitales, informes médicos, artículos científicos, correos electrónicos, páginas Web, información sobre secuencias genéticas, etc. Las nuevas informaciones son generadas a tal velocidad, que es imposible su análisis manual y su exploración efectiva, por lo que se hace necesario desarrollar técnicas de Minería de Textos para ayudar a los usuarios a procesarla. Estas técnicas requieren un conjunto de herramientas básicas del Procesamiento de Lenguaje Natural, entre las que se encuentran los etiquetadores morfosintácticos.

Muchas palabras del lenguaje natural son ambiguas desde el punto de vista gramatical y semántico. La *desambiguación gramatical* consiste en: dado un documento textual asignar la categoría gramatical correcta a cada uno de sus términos atendiendo al contexto en que ellos ocurren dentro del documento. Las herramientas encargadas de realizar este proceso, pudiendo aportar, además, la información morfológica de la categoría gramatical, son conocidas como *etiquetadores morfosintácticos*.

Los etiquetadores morfosintácticos se pueden clasificar en dependencia de la forma de modelar el conocimiento en *deductivos basados en el conocimiento*, *inductivos basados en técnicas de aprendizaje automático* e *híbridos*. En los primeros, también denominados *basados en reglas*, el conocimiento es modelado por lingüistas empleando formalismos gramaticales (ej. *EngCG* de Samuelsson & Voutilainen, 1997). Los métodos inductivos consideran que el conocimiento lingüístico se puede inferir a partir de la experiencia a menudo recogida en corpus textuales, por lo que en ellos se construye un modelo computacional utilizando métodos estocásticos o probabilísticos a partir de ejemplos. Éstos se pueden clasificar en supervisados o no supervisados en dependencia de si los ejemplos utilizados contienen información lingüística o no. Se han desarrollado muchas técnicas inductivas para resolver el problema de la desambiguación gramatical, entre ellas podemos mencionar el modelo *n*-gramas (Bahl et al., 1983), el aprendizaje basado en ejemplos que se basa en el principio de la similitud (Daelemans et al., 1996), el aprendizaje basado en reglas de transformación (Brill, 1992), los modelos de máxima entropía (Ratnaparkhi, 1996), los árboles de decisión (Schmid, 1994) y los modelos de Markov (Molina, 2004). Por último, los modelos híbridos combinan las reglas de contexto con los métodos probabilísticos (Márquez & Padró, 1997).

Las aproximaciones basadas en los modelos ocultos de Markov que combinan adecuadamente técnicas de suavizado con el tratamiento de las palabras desconocidas han logrado resultados similares a los obtenidos por otros enfoques empleados en la actualidad para la desambiguación gramatical (Brants, 2000; Dandapat et al., 2004).

En este trabajo se propone un etiquetador morfosintáctico para el idioma español hablado en Cuba que combina los modelos ocultos de Markov basados en bigramas con un analizador morfológico, heurísticas y diccionarios. El

analizador morfológico está basado en la morfología de dos niveles de Kimmo Koskenniemi (Koskenniemi, 1983) y brinda todas las posibles interpretaciones morfológicas de una palabra. Este artículo se centra en el etiquetador morfosintáctico.

2 El Etiquetador Morfosintáctico

La arquitectura del etiquetador propuesto se puede observar en la figura 1.

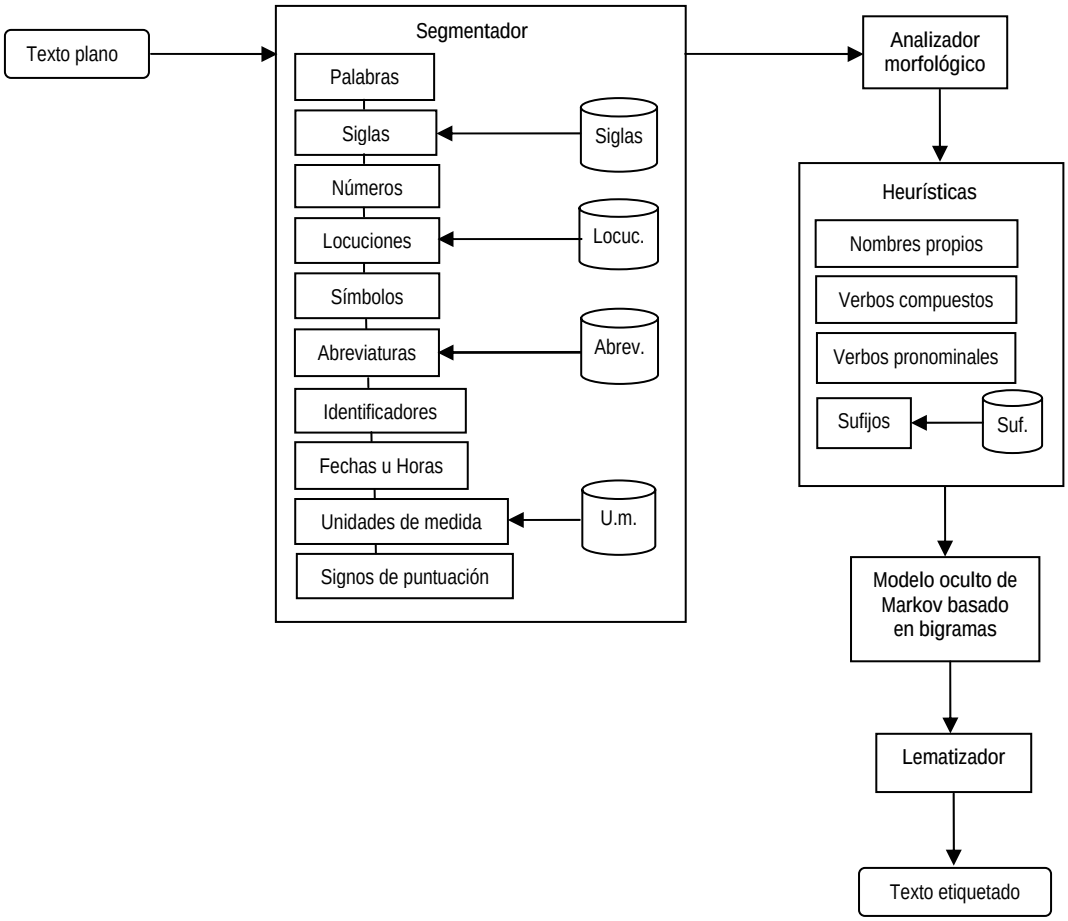

Fig. 1. Arquitectura del etiquetador morfosintáctico.

El *segmentador* se encarga de separar un texto plano en unidades lingüísticas tratables e identificar sintácticamente las oraciones. Es capaz de reconocer palabras, signos de puntuación, símbolos e identificadores. Consideramos un identificador a cualquier secuencia de caracteres que no constituye una palabra del lenguaje como las direcciones de correo, URLs, LB-12, expresiones como: 2+5*4=22, entre otras. Además, apoyándose en un conjunto de diccionarios, es capaz de reconocer siglas, unidades de medida, abreviaturas y locuciones; y emplea heurísticas para identificar fechas, horas y números en cualquier sistema numérico. Hay que destacar que los diccionarios empleados están enfocados al español que se habla en Cuba, pues se han incluido palabras que no son utilizadas en otras partes del mundo o su uso es muy poco frecuente y, sin embargo, son comúnmente utilizadas en Cuba.

El analizador morfológico brinda todas las posibles interpretaciones morfológicas de una palabra dada y su información asociada, incluyendo al lema como una información morfológica más. Se define como *lema* a la forma canónica de una palabra. Una misma palabra podría tener más de un lema, por ejemplo, el lema de la palabra “camino” es “camino” si ésta es un sustantivo o “caminar” si es un verbo.

El empleo del analizador morfológico permite restringir el conjunto de las posibles categorías gramaticales de una palabra dada. Existen notables diferencias entre el español que se habla en los diferentes países de habla hispana. El analizador morfológico empleado está enfocado a las características propias del español hablado en Cuba.

Mediante el empleo de un conjunto de heurísticas, el etiquetador propuesto reconoce los nombres propios, los verbos compuestos (ej. he comido) y los verbos pronominales (ej. me abstengo). Como resultado de la

segmentación, el análisis morfológico y la aplicación de las heurísticas se obtiene el conjunto de posibles categorías gramaticales de cada palabra.

La desambiguación de cada unidad lingüística se realiza empleando el modelo oculto de Markov basado en bigramas a partir del conjunto de posibles categorías gramaticales obtenido anteriormente. La etiqueta correcta junto con la información aportada por el analizador morfológico le permiten al lematizador determinar el lema de cada palabra. En el caso particular de las siglas, el lematizador considera su significado como lema, y en las abreviaturas y unidades de medida el lema lo constituye su expansión.

Finalmente se obtiene el texto etiquetado. Cada unidad lingüística del texto está acompañada de su etiqueta con la información morfológica (que incluye género, número, persona, modo, tiempo, entre otras) y su lema.

El conjunto de las etiquetas empleado (ver tabla 1) está compuesto por 40 rasgos gramaticales desglosados en:

- 2 etiquetas para los sustantivos
- 6 etiquetas para los pronombres
- 7 etiquetas para los verbos
- 5 etiquetas para los numerales
- 6 etiquetas para las locuciones
- 14 etiquetas para el resto de las categorías

Se incluyeron en el listado varias etiquetas para los distintos tipos de numerales con el objetivo de permitir la posterior integración del etiquetador con una herramienta para reconocer nombres de entidades.

Artículo	Locución adverbial
Sustantivo propio	Preposición
Sustantivo común	Locución preposicional
Locución nominal	Conjunción
Adjetivo	Locución conjuntiva
Pronombre personal	Interjección
Pronombre demostrativo	Contracción
Pronombre posesivo	Locución latina
Pronombre indefinido	Sigla
Pronombre relativo	Número
Pronombre interrogativo y exclamativo	Unidad de medida
Verbo en forma personal	Fecha u Hora
Verbo en infinitivo	Identificador
Verbo en gerundio	Símbolo
Verbo en participio	Signo de puntuación
Verbo en forma personal con enclítico	Numeral múltiplo
Verbo en infinitivo con enclítico	Numeral cardinal
Verbo en gerundio con enclítico	Numeral ordinal
Locución verbal	Numeral colectivo
Adverbio	Numeral fraccionario

Tabla 1. Conjunto de etiquetas utilizadas.

2.1 El Modelo Oculto de Markov

Como mencionamos anteriormente, el etiquetador propuesto emplea el modelo oculto de Markov basado en bigramas. Los estados del modelo representan las etiquetas y las salidas, las palabras. Como está basado en bigramas, este modelo asume que para determinar la etiqueta correcta de una unidad lingüística dada sólo se necesita conocer dicha unidad lingüística y la etiqueta asignada a la unidad lingüística anterior, es decir,

$$\arg\max_{T_1,\ldots,T_n} \left\{ P(w_k \mid c_i) \cdot P(c_i \mid c_j) \right\} \qquad (1)$$

donde w_k es la unidad lingüística que se desea desambiguar, $\{T_1,\ldots,T_n\}$ es el conjunto de las posibles etiquetas de dicha unidad lingüística, c_j es la etiqueta asignada a la unidad lingüística anterior y $c_i \in \{T_1,\ldots,T_n\}$.

En el caso de que w_k se encuentre al inicio de una oración, la probabilidad de transición $P(c_i \mid c_j)$ no puede considerarse, por lo que se utiliza la probabilidad de inicio, es decir,

$$\arg\max_{T_1,\ldots,T_n} \left\{ P(w_k \mid c_i) \cdot \pi_i \right\} \qquad (2)$$

donde π_i es la probabilidad de que una unidad lingüística con etiqueta c_i se encuentre al inicio de una oración.

2.2 *Tratamiento de las Palabras Desconocidas*

Durante el procesamiento de un texto se encuentran palabras que no fueron vistas durante la etapa de entrenamiento, las cuales se denominan *palabras desconocidas*. Para determinar las posibles categorías de una palabra desconocida aplicamos la heurística de los sufijos, debido a que éstos son útiles para predecir las posibles categorías gramaticales que se le pueden asignar a una palabra determinada. Para ello, se cuenta con un diccionario de sufijos y sus posibles categorías gramaticales. Por ejemplo, el sufijo –ería es un indicador de que la palabra puede ser un sustantivo (ej.: cafetería, plomería, zapatería) o un verbo en forma personal (ej. atendería, cedería, temería).

Una vez que a la palabra desconocida se le ha determinado su posible conjunto de categorías gramaticales, se requiere de alguna probabilidad de observación para poder aplicar la ecuación (1) ó (2) según sea el caso. Los métodos que tratan de contrarrestar el efecto de los sucesos no vistos se denominan *métodos de suavizado*.

Un método de suavizado básico consiste en asignar una cierta probabilidad al espacio de sucesos no vistos mediante la aplicación de la *Ley de Laplace*, también conocida como *Añadir Uno* (*Adding One*, en inglés) (Jeffreys, 1948). En este método se incrementa la frecuencia de todos los sucesos en una unidad y la probabilidad de observación se define como:

$$P^{suavizada}(w_k \mid c_i) = \frac{f(w_k, c_i)+1}{f(c_i)+|V|}$$

donde V es el diccionario formado por las palabras que aparecen en el corpus de entrenamiento, $f(w_k, c_i)$ es la cantidad de veces que la palabra w_k está etiquetada con c_i y $f(c_i)$ es la cantidad de unidades lingüísticas etiquetadas con c_i en el corpus de entrenamiento.

La probabilidad de observación para las palabras desconocidas sería entonces:

$$P^{suavizada}(w_k \mid c_i) = \frac{1}{f(c_i)+|V|}$$

Si la heurística de los sufijos no brinda ninguna posible categoría gramatical para la palabra desconocida, se aplica el modelo oculto de Markov considerando como posibles categorías aquellas que constituyen clases abiertas (sustantivos, adjetivos, verbos, etc.).

En el caso de las palabras desconocidas consideramos como lema la propia palabra.

3 Resultados experimentales

Con el objetivo de evaluar la calidad del etiquetador propuesto contamos con un corpus formado por documentos jurídicos que contienen 231305 unidades lingüísticas. Dicho corpus fue manualmente etiquetado por lingüistas.

En la evaluación se aplicó la técnica de validación cruzada. El corpus se dividió en 3 subconjuntos. De los 3 subconjuntos, uno es empleado en la prueba y los otros dos en el entrenamiento. El proceso de validación cruzada se repite tres veces considerando en cada caso un subconjunto diferente para la prueba.

Para medir la calidad se empleó la tradicional medida de *precisión* que se define, en este contexto, como el cociente entre el número de unidades lingüísticas etiquetadas correctamente y el número total de unidades lingüísticas del corpus. Los resultados obtenidos se pueden apreciar en la tabla 2. La segunda y tercera columna de la tabla contienen la cantidad de unidades lingüísticas que conforman los conjuntos de entrenamiento y prueba, respectivamente.

Subconjuntos	**Tamaño Entrenamiento**	**Tamaño Prueba**	**Precisión**
Subconjunto 1	154786	76519	96.5 %
Subconjunto 2	153836	77469	96.6 %
Subconjunto 3	153988	77317	96.6 %

Tabla 2. Resultados de la validación cruzada.

Como puede apreciarse se obtuvieron resultados satisfactorios, acordes con la efectividad de los etiquetadores existentes (Márquez, 1999).

4 Conclusiones

En este trabajo se propuso un etiquetador morfosintáctico capaz de etiquetar las unidades lingüísticas presentes en textos escritos en español, haciendo énfasis en las particularidades del español hablado en Cuba. Además, puede ser entrenado para procesar textos sobre cualquier rama del conocimiento.

El etiquetador combina los modelos ocultos de Markov basados en bigramas con diversas heurísticas y diccionarios. Se apoya, además, en un analizador morfológico para asignar la etiqueta y el lema correcto a cada unidad lingüística de un texto. El empleo del analizador morfológico no sólo permite añadir la información morfológica a la etiqueta, sino que permite acotar el conjunto de las posibles categorías gramaticales a considerar para una determinada unidad lingüística e incluye especificaciones del español hablado en Cuba.

En los experimentos realizados se utilizó un corpus jurídico obteniendo resultados satisfactorios. La precisión obtenida está alrededor del 96.6 % equivalente a la efectividad de los etiquetadores existentes.

Como trabajo futuro pretendemos entrenar al etiquetador morfosintáctico con corpus de otros dominios del conocimiento e integrar esta herramienta con un reconocedor de nombres de entidades para el procesamiento de textos en lenguaje natural.

Referencias bibliográficas

Bahl, L. R. Jelinek, F. & Mercer, R. L. (1983). A Maximum-Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI, pp. 179-190.

Brants, T. (2000). TNT-A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seatle, pp. 224—231.

Brill, E. (1992). A simple rule-based Part of Speech tagger. *Proceedings of the Third Conference on Applied Natural Language Processing of the Association for Computational Linguistics*, Trento, Italy, pp. 152-155.

Daelemans, W., Zavrel, J., Berck, P. & Gillis, S. (1996). MBT: A Memory-Based Part of Speech Tagger-Generator. In *Proceedings of the Fourth Workshop on Very Large Corpora (WVLC'96)*, In: E. Ejerhed and I. Dagan (eds), Copenhagen , pp. 14-27.

Dandapat, S., Sarkar, S. & Basu, A. (2004). A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali. International Conference on Computational Intelligence, pp. 169-172.

Jeffreys, H. (1948).Theory of Probability, Second Edition, Section 3.23, Oxford, Clarendon Press.

Koskenniemi, K. (1983). Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. Tesis doctoral. Universidad de Helsinki, Finlandia.

Márquez, L. (1999). *Part-of-speech Tagging: A Machine Learning Approach based on Decision Trees*. Tesis doctoral. Universidad Politécnica de Cataluña, Barcelona, España.

Márquez, L. & Padró, L. (1997). A flexible POS tagger using an automatically acquired language model. In *Proceedings of ACL-97*, Madrid, pp. 238--245.

Molina, A. (2004). *Desambiguación en procesamiento del lenguaje natural mediante técnicas de aprendizaje automático*. Tesis doctoral. Universidad Politécnica de Valencia, España.

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-of-Speech Tagging. In *Proceedings of the 1st Conference on Empirical Methods in Natural Language Processing*, EMNLP, Pennsylvania.

Samuelsson, C. & Voutilainen, A. (1997). Comparing a Linguistic and a Stochastic Tagger. In *Proceedings of 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, ACL, Madrid, pp. 246-253.

Schmid, H. (1994). Probabilistic Part-of-speech Tagging Using Decision Trees. In *Proceedings of the Conference on New Methods in Language Processing*, Manchester, UK, pp. 44-49.