Implementación de un Modelo para la Acentuación Automática de Palabras Ambiguas del Español, utilizando Etiquetado de Texto.

Raymundo Montiel, Blanca E. Pedroza, María Guadalupe Medina, Carlos Pérez

Instituto Tecnológico de Apizaco
División de Estudios de Posgrado e Investigación
Av. Instituto Tecnológico s/n, C.P. 90300, Apizaco, Tlaxcala, México
{mlirary,thelismedina}@hotmail.com

Resumen. El proceso de acentuar palabras a la hora de escribir texto en español, es un problema de ambigüación, debido a que existen muchas palabras que pueden estar acentuadas o no dependiendo del contexto de la frase. El problema de ambigüación está relacionado con el proceso de asignar etiquetas o categorías gramaticales a las palabras dentro de una frase, es decir indicar si es un verbo, sustantivo, etc. En este trabajo se propone un modelo, que ayuda a determinar en forma automática si una palabra con acento diacrítico debe llevar o no acento ortográfico en base a la asignación de etiquetas, mediante la aplicación de métodos híbridos (algoritmos supervisado y no supervisados). Posteriormente esté método es aplicado para el diseño de un programa de computo, que sirva de apoyo para la enseñanza de las reglas de acentuación y para la acentuación automática de palabras ambiguas, el cual puede ser implementado como herramienta en un procesador de palabras.

Palabras claves: Word sense disambiguation, accent restoration, part of speech tagger.

1 Introducción

Una de las tareas más difíciles y que ha suscitado un gran interés dentro del *Procesamiento del lenguaje natural (PLN)*, se produce cuando se asocian varios sentidos o significados a una palabra, este fenómeno lingüístico se le conoce como *polisemia*. A este fenómeno se asocia lo que dentro del procesamiento del lenguaje natural se conoce como desambiguación del sentido de la palabra (DSA), que es una tarea que consiste en identificar el sentido correcto de una palabra en un contexto [4]. El problema de la polisemia está muy relacionado con el problema de la asignación de categorías gramaticales, el cual consiste en decir si una palabra es un verbo un artículo, un sustantivo [3], dependiendo del significado que le corresponda de acuerdo al contexto de la oración.

M.A Cruz-Chávez, J.C Zavala Díaz(Eds):CICos2009, ISBN:978-607-00-1970-8, pp. 123 - 132, 2009.

El proceso de acentuar palabras es también un problema de ambigüación, debido a que existen muchas palabras dentro del español que pueden estar acentuadas o no dependiendo a diferentes situaciones, como lo es el contexto, el tiempo de acción de las oraciones, etc. El acento diacrítico permite distinguir palabras con idéntica forma, es decir, escritas con las mismas letras, pero que pertenecen a categorías gramaticales diferentes.

Aunque los procesadores de textos de las grandes compañías de software, poseen funciones muy sofisticadas, como es la incorporación de un diccionario para corregir su ortografía, lo hacen hasta cierto punto. "Estos correctores no comprenden el significado de las palabras, si no que se limitan a ir comparando cada palabra del texto con las existentes en su diccionario, de modo que si encuentra la palabra en cuestión la dará por válida si está escrita correctamente, en caso contrario la marca para indicar un error [1]". Sin embargo, cuando existen palabras que pueden o no llevar acento ortográfico, el procesador no las marca como errores si el usuario no les coloca el acento, y tampoco le indica que puede estar mal escrita y se queda con la idea de que está escrita correctamente.

La falta de acentos en algunas palabras dentro de las oraciones, es debido a los problemas de ambigüación. Las ambigüedades más comunes en la acentuación de palabras se dan entre las palabras con terminación –o, tal como el caso de *completo y completó*. Se trata de los tiempos presente / pretérito, relacionado con los verbos con terminación –ar. Existen otras ambigüedades puramente semánticas, las cuales incluyen a sustantivos como: secretaria (RAE [8]: Persona encargada de escribir la correspondencia) y secretaría (RAE [8]: Sección de un organismo, institución o empresa).

En el presente artículo se describe el modelo utilizado para desarrollar una herramienta informática de desambiguación semántica para el español, diseñada para la correcta acentuación de palabras de textos escritos. Dicha herramienta se basa en la etiquetación de las oraciones. Para el modelo se lleva a cabo la implementación de un método híbrido, esto es, mediante un algoritmo supervisado y uno no supervisado.

2 Solución

El objetivo de este trabajo es implementar un sistema que ayude a determinar si una palabra con acento diacrítico debe o no llevar acento ortográfico, lo cual está determinado por el contexto en el que se esté ocupando, con ayuda de las etiquetas asignadas. El modelo general que se propone para la solución del problema se ilustra en la **Fig. 1**. La primera tarea del modelo es el análisis léxico, la cual consiste en la eliminación de los acentos de las palabras de la frase de entrada, ya que para el cálculo de los parámetros del modelo se requieren palabras sin acentos. Además el analizador léxico, identifica y separa los signos de puntuación de las palabras. Posteriormente se identifica si existen palabras ambiguas dentro de la oración que se tienen de entrada al modelo y al mismo tiempo indica la posición de estas. Esto se realiza a través una comparación de cada unas de las palabras de las oraciones, con las palabras del diccionario con acento diacrítico, que previamente se construyó.

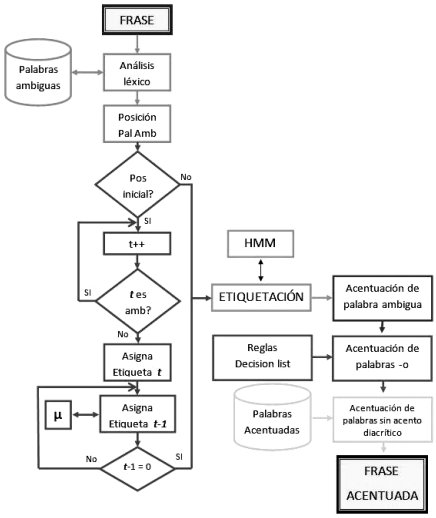

Fig. 1. Arquitectura del modelo de solución

Dentro del método de solución, se utilizan los Modelos Ocultos de Markov (HMM) los cuales se definen, siguiendo la notación presentada por Rabiner [9], como la una 5-tupla $\boldsymbol{\mu = (Q, V, \pi, A, B)}$, donde:

1. ***Q*** es el conjunto de estados del modelo. Aunque los estados permanecen ocultos, para la mayoría de las aplicaciones prácticas se conocen *a priori*. Para nuestro caso de la etiquetación de palabras, cada etiqueta sería un estado. Generalmente los estados están conectados de tal manera que cualquiera de ellos se puede alcanzar desde cualquier otro en un solo paso. Los estados se etiquetan como {1, 2,…, N}, y el estado actual en el instante de tiempo ***t*** se denota como q_t . Para el caso de la etiquetación de palabras, no hablaremos de los instantes de tiempo, sino de las posiciones de cada palabra dentro de la frase.
2. ***V*** es el conjunto de los distintos sucesos que se pueden observar en cada uno de los estados. Cada uno de los símbolos individuales que un estado puede emitir se denota como $\{v_1, v_2, \ldots, v_M\}$. En el caso de etiquetación de palabras, *M* es el tamaño del diccionario y cada $v_k, 1 \leq k \leq M$, es una palabra distinta.
3. $\pi = \{\pi_i\}$, es la distribución de probabilidad de estado inicial. Por tanto,

$$\pi_i = P(q_1 = i), \qquad \pi_i \geq 0, \qquad 1 \leq i \leq N, \qquad \sum_{i=1}^{N} \pi_i = 1 \tag{1}$$

4. $\boldsymbol{A} = \{a_{ij}\}$ es la distribución de probabilidad de las transiciones entre estados, es decir,

$$a_{ij} = P(q_t = j | q_{t-1} = i) = P(j|i), \qquad 1 \leq i, j \leq N, \qquad 1 \leq t$$
$$\sum_{j=1}^{N} a_{ij} = 1, \qquad \forall\ i. \tag{2}$$

5. $\boldsymbol{B}=\{b_j(v_k)\}$ es la distribución de probabilidades de los sucesos observables, es decir,

$$b_j(v_k) = P(o_t = v_k | q_t = j) = P(v_k|j),$$
$$b_j(v_k) \geq 0, \qquad 1 \leq j \leq N, \qquad 1 \leq k \leq M, \qquad 1 \leq t \leq T$$
$$\sum_{k=1}^{M} b_j(v_k) = 1, \qquad \forall i. \tag{3}$$

Este es un método estocástico, el cual necesita de parámetros previamente definidos, los cuales se calculan, mediante métodos de entrenamiento supervisado y uno no supervisado.

Para la fase del entrenamiento supervisado se utilizó un cuerpo de texto etiquetado, conocido como CONLL, que es una colección de artículos de noticias ofrecidas por la Agencia de Noticias EFE del año 2000. Los parámetros del modelo se pueden estimar por *máxima verosimilitud* [5], a partir de las frecuencias relativas de aparición de los eventos del texto etiquetado. Ver **Fig. 2**.

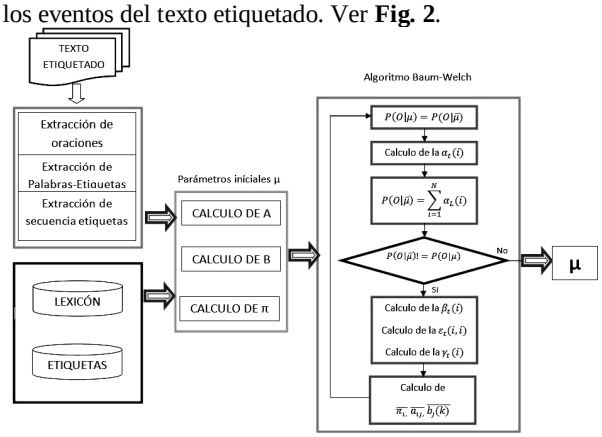

Fig. 2. Entrenamiento supervisado y no supervisado

Las *probabilidades de transición* a_{ij} se obtienen contando las veces que se transita del estado s_i al estado s_j y dividiendo por las veces que se transita por el estado s_i.

$$a_{ij} \approx \hat{P}\left(q_t = s_j \middle| q_{t-1} = s_i\right) = \frac{f(q_{t-1} = s_i, q_t = s_j)}{f(q_{t-1} = s_i)} \quad \textbf{(4)}$$

Las *probabilidades de emisión* se obtienen contando las veces que un símbolo (v_k) ha sido emitido en un estado (s_j) y dividiendo por las veces que se ha transitado por dicho estado:

$$b_j(v_k) = P\left(o_t = s_i, \middle| q_t = s_j\right) = \frac{f(o_{t-1} = v_k, q_t = s_j)}{f(q_t = s_j)} \quad \textbf{(5)}$$

Una vez que tenemos los parámetros iníciales de ***μ,*** se aplica el algoritmo Baum-Welch [6], lo cual incrementa la probabilidad de las transiciones entre los estados y sus símbolos, y en consecuencia se mejora la probabilidad para la secuencia de observaciones dada.

Dentro del modelo de solución (**Fig. 1**) se maneja una fase de etiquetación, la cual se realiza utilizando el algoritmo de Viterbi, con algunas modificaciones, ya que no se consideran todos los estados posibles, es decir, todas las etiquetas del juego de etiquetas utilizado, sino sólo las etiquetas candidatas asignadas a cada una de las palabras ambiguas.

Para ejemplificar el funcionamiento del modelo, consideremos la oración "*El jugo frío esta sobre la mesa"*, en la cual existen varias palabras ambiguas. Además, como existen palabras ambiguas al inicio de la oración, se recorre *t* palabras hasta encontrar una palabra no ambigua, en este caso ***"frío".*** Una vez encontrada la primera palabra no ambigua, se le asigna la etiqueta más probable. Después de esto, se asigna la etiqueta más probable a *t-1* a partir de la etiqueta ya conocida (*t*) (**Fig. 3**), hasta llegar a t=1.

Obtenidas las primeras etiquetas de la oración (si es el caso), el cálculo para obtener las etiquetas de las palabras ambiguas dentro de la oración, se realiza solo sobre las etiquetas asignadas manualmente a estas palabras. En la **Tabla 1** se muestra la oración junto con las etiquetas asignadas.

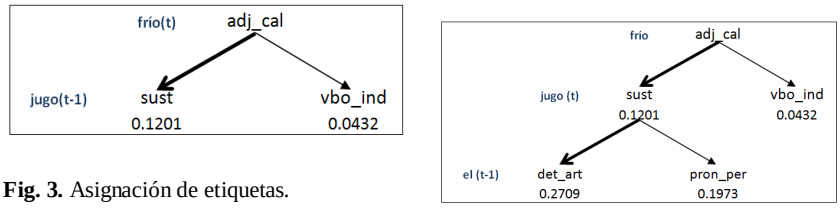

Fig. 3. Asignación de etiquetas.

Tabla 1. Resultado de la etiquetación

El	**jugo**	**frío**	**esta**	**sobre**	**la**	**mesa**
det_art	sust	adj_cal	vbo_aux	prep	det_art	sust

Después de haber obtenido las etiquetas de cada una de las palabras de la oración, se determina si las palabras ambiguas deberán o no llevar acento, en base a nuestro diccionario de palabras ambiguas etiquetadas (Tabla 2).

Tabla 2. Ejemplo de palabras ambiguas con sus etiquetas

PALABRA AMBIGUA	ETIQUETA	PALABRA
EL	**det_art** pron_per	**el** él
ESTA	det_dem **vbo_aux**	esta **está**
JUGO	**sust** vbo_ind	**jugo** jugó
SOBRE	sust **prep** vbo_sub vbo_ind	sobre **sobre** sobre sobré

Por último, se hace una comparación de las palabras de la oración junto con la etiqueta asignada y determina si la palabra debe o no llevar acento. El resultado de nuestro ejemplo se muestra en la **Tabla 3**.

Tabla 3. Resultado de ejemplo

El	**Jugo**	**frío**	**está**	**sobre**	**la**	**mesa**

Uno de los problemas que no se resolvió con la asignación de categorías gramaticales, fue al encontrar una palabra con acento diacrítico etiquetada con la misma categoría gramatical, como se muestra en la **Tabla 4.**

Tabla 4. Palabras con terminación –o

Palabra	**Etiqueta**
abrazo	sust
abrazo	**vbo_ind**
abrazó	**vbo_ind**

Las dos últimas palabras, corresponden a la conjugación en modo indicativo del verbo *Abrazar*, sin embargo, la primera corresponde al tiempo presente en primera persona y la segunda al tiempo pasado en tercera persona.

Para solucionar este problema se crearon un conjunto de reglas utilizando Listas de Decisiones, similares a las utilizadas por Yarowsky [2].

En la **Tabla 5,** se presentan un extracto de las reglas obtenidas, ordenadas de acuerdo a su aparición en el corpus de entrenamiento, este mismo también indica el orden en que se han de ir verificando cada una de las reglas al obtener nuevos ejemplos para identificar el patrón que le corresponda.

Tabla 5. Extracto del conjunto de reglas

%	Regla		Clasificación
19.63	**YO** -o	»	-o
19	**SE** -o	»	-ó
16.6	-o **DE**	»	-o
14.6	-o **det_art**	»	-ó
10.83	**sust** -o	»	-ó
10.6	-o **CON**	»	-o
9.09	**adv** -o	»	-o
8.33	-o **det_ind**	»	-ó

Además de acentuar aquellas palabras que tengan ambigüedad, se trata de acentuar las palabras que siempre llevan acento, es decir aquellas palabras que tienen acento pero no acento diacrítico. Para esto se realizó una base de datos de palabras con acento.

3 Resultados

Para la parte de la evaluación del modelo, se reunió una colección de 63 artículos de distintos contextos (Deportes, Ciencia, Salud, etc.) obtenidos del periódico digital "El Universal[1]", estos fueron evaluados por el modelo a partir de frases completas, la **Tabla 6** muestra un resumen de los resultado obtenidos.

Donde:

1. # del artículo
2. # Palabras por artículo
3. # Palabras con acento diacrítico
4. # Palabras con acento diacrítico, sin acento
5. # Palabras con acento diacrítico, sin acentuar
6. # Palabras con acento diacrítico, con acento
7. # Pal con acento diacrítico, acentuadas
8. # Palabras con acento no diacrítico
9. # Palabras con acento no diacrítico, acentuadas

[1] http://www.eluniversal.com.mx/noticias.html

Tabla 6. Resumen de resultados obtenidos

1	2	3	4	5	6	7	8	9
1	164	26	26	26	0	0	12	12
7	166	24	22	22	2	2	14	8
10	214	30	26	26	4	3	22	21
19	376	54	49	49	5	3	26	23
24	1180	215	191	170	22	19	125	101
47	322	69	67	66	2	2	26	23
48	384	87	79	78	8	7	23	20
53	898	187	164	162	23	21	78	74
58	114	78	68	63	10	8	34	30
63	165	34	28	26	6	6	21	21
RESUL	**22883**	**4238**	**3769**	**3611**	**472**	**397**	**1721**	**1526**

Con estos datos nos damos cuenta que:

$$palabras\ ambiguas = \frac{4238\ x\ 100}{22883} = 18.52\ \%$$

$$palabras\ acentuadas\ (acento\ no\ diacrítico) = \frac{1526\ x\ 100}{1721} = \ 88.66\ \%$$

Aparte de estos resultados, los datos mostrados en la **tabla 7** son los resultados obtenidos utilizando las métricas de efectividad de Tasa de recuperación (r), Precisión (p), Error (e), Exactitud (a), y $F_{B.}$

En la primera parte de la **Tabla 7**, se muestran los resultados obtenidos a partir de las palabras acentuadas por nuestro modelo, donde la *precisión* (total de palabras acentuadas entre el número total de palabras con acento diacrítico) tiene como resultado 84.11 %, se ha identificado que uno de los problemas, está en la etiquetación de la palabra "***el***", ya que siempre que el sistema encuentra a esta palabra al inicio de la frase, le asigna la etiqueta 'pronombre personal' y la acentúa , sin embargo, también puede tomar la etiqueta de 'artículo', en cuyo caso no lleva acento.

Tabla 7. Resultados

En base a palabras acentuadas		En base a palabras No acentuadas
r = .7153	**e = 0.0549**	r = .9796
p = .8411	**a = .945**	p = .9580
f_{-B} = .7731		f_{-B} = .9687

Pero, como se menciona anteriormente, este modelo también determina cuando una palabra con acento diacrítico no debe de llevar acento ortográfico. La segunda parte de la **Tabla 7**, muestra los resultados obtenidos a partir de las palabras que no deben de llevar acento ortográfico, en este caso la *precisión* (total de palabras que el modelo

decide no acentuar, entre el número total de palabras con acento diacrítico) tiene como resultado 95.8 %.

Sin embargo, tomando en cuenta el resultado de exactitud (total de decisiones correctas tomadas por el modelo) es del 94.5 %, teniendo un error de 5.49 %.

El modelo propuesto se programó en "Java" y se esta trabajando en la implementación de un programa de computo que sirva como auxiliar en la enseñanza de las reglas de acentuación y para acentuar palabras ambiguas que por lo regular los editores de texto no acentúan. En la figura 4, se muestran algunas de las pantallas del sistema.

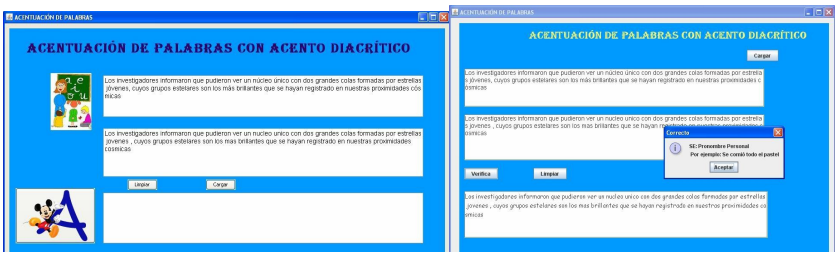

Fig. 4. Pantallas del sistema para la acentuación de texto en español.

4 Conclusiones

Podemos decir, que el etiquetado de palabras es una técnica que nos puede ayudar a identificar la ambigüedad en la acentuación de palabras de un texto determinado y generar herramientas computacionales que pueden ayudar a la correcta acentuación de textos en español, ya sea, para hacer más fácil la escritura de documentos en español o utilizarlo como un asistente en la enseñanza de las reglas gramaticales.

Aunque el problema de la ambigüedad en la acentuación de las palabras, no es un problema sencillo de resolver, se puede seguir analizando la gramática del español, para encontrar ciertos patrones en el comportamiento de las palabras con ciertas características de acuerdo al contexto en el que se encuentren las oraciones. Además de que se pueden aplicar diversas teorías y algoritmos existentes para el manejo de las gramáticas, así como diversas teorías del área de los sistemas inteligentes, como el aprendizaje automático. Por lo que todavía hay mucho que investigar en este tema, para poder generar sistemas de computo que ayuden al usuario a corregir sus errores ortográficos en la acentuación.

Trabajos Futuros

- Aumento de:
  - Palabras con acento diacrítico.
  - Palabras con acento (no diacrítico).
  - Categorías gramaticales para disminuir ambigüedad en la etiquetación.

- Creación de un módulo dentro del modelo, para el entrenamiento de los casos no resueltos.
- Diseño de Interfaz, para la enseñanza de las reglas gramaticales, en específico, del uso de las palabras con acento diacrítico.
- Implementar algoritmo de desambiguación semántica, para mejorar resultados.
- Realizar un análisis sintáctico previo para reducir ambigüedad en la frase.
- Creación de un sistema global de apoyo para la enseñanza o reafirmación de las reglas ortográficas (acentuación).

Identificación de un Patrón para palabras que generan ambigüedad, como el caso, antes mencionado, de la palabra "el".

5 Referencias

1. Pascual, F. (2002). *Domine Microsoft© Office XP Profesional, edición especial versión 2002*. Ed. Alfaomega. Mexico DF.
2 . Yarowsky, D. (1994). *Decision list for lexical ambiguity resolution: Application to accent restoration in Spanish and French*. In Proceeding of the 32th Annual Meeting of the Association for Computational Linguistics, (ACL´94). 88-95.
3. Perea, J. (2005). *Etiquetado de textos y su aplicación a la traducción*. Unplublished monitired research. University of Granada.
4. Stevenson, M., Y. Wilks. (2000). *Combining independent knowledge sources for word sense disambiguation*. En R. Mitkov (ed), Recent Advances in Natural Language Processing, John Benjamins Publisher.
5. Dempster A., Laird, N., et al. (1977). *Maximum likelihood from incomplete data via the EM algorithm*. Journal of Royal Statistical Society. Series B (Methodological), vol. 39, No. 1, 1-38.
6. Baum, L. (1966). *Statistical inference for probabilistic functions finite state Markov chains*. Ann. Math. Stat. vol. 37, 1554-1563.
7. Yarowsky, D. (1994). *A comparison of corpus-based techniques for restoring accents in Spanish and French text.* To appear in Proceedings, 2nd annual Wordshop on Very Large Text Corpora, Kyoto, Japan.
8. REAL ACADEMIA ESPAÑOLA: Banco de datos (CREA) [en línea]. Corpus de referencia del español actual. http://www.rae.es [29 Abril 2009].
9. Rabiner, L. R. (1989). *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. IEEE, vol. 77, no. 2, pp. 257-286.
10. Wagacha, P., De Pauw, G., et al. (2006). *A Grapheme-Based approach for accent restoration in Gikuyu*. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, pp. 1937 – 1940.
11. Bobiceva, V. (2008). *O altã metodã de restabilire a semnelor diacritice*. En Pistol I., Cristea D. Tufiş D. (eds): Resurse Lingvistice şi Instrumente pentru Prelucrarea Limbii Române, pp. 179-188.
12. De Pauw, G, Wagacha, P. W., de Schryver, G-M (2007). *Automatic Diacritic Restoration for Resource-Scarce Language*. En V. Matousek and P. Mautner (Eds.): TSD 2007, LNAI 4629, pp. 170–179