Programación Matemática y Software (2014) 6 (1): 14-22. ISSN: 2007-3283

Implementación del clasificador naive Bayes para la acentuación automática de palabras ambiguas del espa|ol

Automatic accent detection using naive Bayes clasifier for Spanish language

Yesenia N. González-Meneses [1]*, Blanca Estela Pedroza-Méndez [1], Francisco López-Briones, [1]Carlos Pérez-Corona, [1] J. Federico Ramírez-Cruz [1]

[1] Instituto Tecnológico de Apizaco.
Av. Instituto Tecnológico s/n. Apizaco, Tlaxcala, México
* Correo-e: yeseniaglez@hotmail.com

PALABRAS CLAVE:

ambigüedad en la acentuación, clasificador naive Bayes, etiquetado de texto

RESUMEN

En este artículo se analiza uno de los problemas más representativos en el tratamiento del lenguaje español, que es el de la ambigüedad en la acentuación gráfica de las palabras. En la escritura del español se utiliza el acento gráfico o tilde, el cual determina la pronunciación o interpretación correcta de las palabras. Algunos vocablos de construcción similar pueden llevar tilde o no, o la llevan en diferente sílaba, lo cual permite que tomen diferentes sentidos en relación con su contexto, para lo cual se utiliza la llamada tilde diacrítica. La asignación correcta de la tilde diacrítica en este proyecto es abordada como un problema de clasificación, donde con base en el contexto se determina si las palabras ambiguas llevan esta marca o no. Para ello se entrenó un modelo con el clasificador naive Bayes.

KEYWORDS:

ambiguity in accentuation, naive Bayes classifier, text labeling

ABSTRACT

This paper analyzes one of the most representative problems in the treatment of Spanish language, which is the ambiguity that exists in the graphic accentuation of words. In written Spanish the diacritic mark representing acute accent is widely used, and helps determine the right pronunciation or interpretation of words. Similarly constructed words can be distinguished by the presence or not of the accent mark, or by its placement in a different syllable, which allows them to take different meanings depending on the context. In this project the correct allocation of the diacritical accent is treated as a classification problem, where the context determines whether ambiguous words should be graphically accented or not. To this end, we trained and tested a model with the naive Bayes classifier.

1 INTRODUCCIÓN

El procesamiento del lenguaje natural (PLN) es un área de la inteligencia artificial, dependiente directamente de la lingüística computacional. Asimismo, es un componente importante de las interfaces de usuarios y los sistemas inteligentes y uno de los objetivos que persigue es el perfecto análisis y entendimiento de los lenguajes humanos [1].

Los esfuerzos de investigación en este campo han sido dirigidos hacia tareas intermedias que dan sentido a alguna de las múltiples características estructurales inherentes a los lenguajes, sin requerir un entendimiento completo. Una de esas tareas es la asignación de categorías gramaticales o morfosintácticas (sustantivo, adjetivo, verbo, etcétera) a cada una de las palabras de una oración. Este proceso se denomina también etiquetación [2]. El proceso de etiquetación debe eliminar ambigüedades y encontrar cuál es el papel más probable que juega cada palabra dentro de una frase. Dicho proceso debe ser capaz también de asignar una etiqueta a cada una de las palabras que aparecen en un texto y garantizar de alguna manera que esa es la etiqueta correcta.

El problema más difícil que se enfrenta en el procesamiento del lenguaje es la ambigüedad, que es cuando pueden admitirse distintas *interpretaciones* a partir de la representación o cuando existe confusión al tener diversas estructuras y no tener los elementos necesarios para eliminar las incorrectas [3]. Este problema se presenta en todos los niveles del lenguaje, sin excepción [4], desde el nivel morfológico (palabras), hasta el discurso (o pragmática).

1.1 Descripción del problema

En el proceso de escribir textos en lenguaje español, muchas veces cometemos errores ortográficos, debido a que es muy común olvidar cómo utilizar las reglas del idioma que regulan esta tarea. Aunque éstas se nos enseñan desde la infancia, se van olvidando ya que no se pone el suficiente empeño en aplicarlas; una de las causas es que tenemos la inteligencia suficiente para entender los textos sin importar que éstos no estén escritos correctamente. Otro de los problemas es que el idioma español en sí es muy ambiguo en la escritura de las palabras, por ejemplo algunas tienen idéntica pronunciación pero su escritura y su significado son diferentes (palabras homófonas), ejemplo: *tuvo* y *tubo*, *huno* y *uno*. Otro caso es el de la *polisemia*, que es cuando una palabra tiene diferentes significados, por ejemplo *banco*, que puede tener significado de institución de crédito o de asiento sin respaldo, etc. En este caso lo que permite darle el sentido correcto a la palabra es su contexto.

Otra de las cosas que genera ambigüedad en el idioma español es la acentuación gráfica de las palabras, ya que algunas se escriben igual pero pueden o no llevar acento, dependiendo del contexto de la frase que contiene la palabra. Por ejemplo a la palabra *grafica* se le debe colocar acento en la *a* de la sílaba *gra* si la palabra es un sustantivo, pero si la palabra dentro de la frase se maneja como un verbo, la sílaba tónica es *fi* y de acuerdo con las reglas de acentuación no lleva acento, ya que es una palabra grave que termina en vocal. Por tanto, se puede observar que existe una relación entre la acentuación y las etiquetas morfosintácticas que se le asignan a las palabras.

Este artículo se enfoca en el análisis de las reglas de acentuación y se propone un modelo basado en métodos de aprendizaje automático, aplicando el clasificador naive Bayes para dar solución al problema de la ambigüedad al asignar el acento diacrítico. El clasificador analiza el contexto de la frase con base en las etiquetas morfosintácticas asignadas a las palabras y determina cuando una palabra debe o no llevar acento diacrítico, para lo cual se deben primero corregir las palabras con acento gráfico, esto es para disminuir el número de errores por omisión de acentos y al mismo tiempo para que las etiquetas generadas sean más precisas. El diccionario utilizado en este proyecto se generó como una de las etapas iniciales, donde se identificaron además otro tipo de ambigüedades en la acentuación gráfica de palabras; para la fase de la etiquetación se utilizó el módulo para este fin del paquete Freeling [5].

La desambiguación del sentido de las palabras (WSD, en inglés) es en esencia una tarea de clasificación: los sentidos de las palabras son las clases, el contexto provee la evidencia y cada una de las palabras es asignada a una o más de las posibles clases basado en la evidencia [6]. El clasificador naive Bayes es uno de los algoritmos que estiman probabilidades a posteriori. Este clasificador asume, para una muestra x, que sus atributos $x1$, $x2$, …, xn presentan una independencia condicional dado el valor de la clase, por lo que la probabilidad condicional puede expresarse como el producto de funciones de probabilidad

condicional de cada atributo por separado [7]. En este sentido los atributos utilizados para la desambiguación son las palabras en contexto a la palabra ambigua, y los valores de cada atributo son las etiquetas morfosintácticas asignadas por Freeling, para que, calculando probabilidades por cada uno de estos valores con respecto a la clase de salida, se pueda definir la clase a la que pertenecen dichas palabras.

2 EL LENGUAJE ESPAÑOL

La ortografía es la rama de la gramática que se ocupa de la escritura correcta [8]. Según el diccionario de la Real Academia Española se define como: "Conjunto de normas que regulan la escritura de una lengua".

Dentro del lenguaje español todas las palabras tienen una sílaba que se pronuncia con mayor intensidad, esto es lo que se conoce como acento prosódico, que es el mayor relieve con que se pronuncia una determinada sílaba dentro de una palabra. Otro tipo de acento que se maneja dentro del español es el gráfico u ortográfico, que es el signo con el cual, en determinados casos, se representa en la escritura el acento prosódico [9].

De acuerdo con las reglas de la gramática del español, los acentos se clasifican de la siguiente manera [10]:

- *Tilde diacrítica o acento diacrítico*. Es la marca que se coloca sobre alguna de las vocales dentro de una palabra para permitir diferenciar entre los significados de ésta.
- *Acento gráfico*. Éste no se utiliza para diferenciar entre los significados sino para saber la pronunciación correcta de una palabra, en el caso contrario la colocación de esta marca es definida por la pronunciación de la palabra.

El error más común cuando escribimos es la omisión tanto del acento gráfico como del acento diacrítico, ya que aunque no es difícil identificar la sílaba tónica, sí lo es recordar las reglas. Actualmente, es muy común el uso de procesadores de texto que ya tienen incluido un diccionario de palabras para ayudar a la acentuación, pero cuando se trata de palabras con ambigüedad en la acentuación, el procesador no indica si deben o no llevar acento.

3 CLASIFICACIÓN

La clasificación es el punto principal en esta investigación, ya que la asignación de la tilde diacrítica a las palabras ambiguas se modela como un problema de clasificación, donde las dos clases para cada palabra son si lleva o no lleva la tilde.

La clasificación es la tarea de aproximar una función objetivo desconocida $\Phi : I \times C \rightarrow \{T, F\}$ por medio de una función $\Theta : I \times C \rightarrow \{T, F\}$ llamada clasificador, donde $C = \{c_1, c_2, ..., c_{|c|}\}$ es un conjunto de clases definido, e I es un conjunto de instancias del problema. Cada instancia ij □ I es representada como una lista $A = \{a_1, a_2, ..., a_{|A|}\}$ de valores característicos, conocidos como atributos. Es decir $ij = \{a_{1j}, a_{2j}, ..., a_{|A|j}\}$. Si $\Phi : I \times C \rightarrow T$ entonces ij es llamado un ejemplo positivo de c_i, mientras que si $\Theta : I \times C \rightarrow F$ es llamado un ejemplo negativo de c_i [11]. En general no se conoce la descripción exacta de las muestras, por lo que el sistema es entrenado a priori para ajustarse a las características propias del problema. A este proceso de adquirir e integrar conocimiento a un sistema de clasificación a partir de ejemplos, se le conoce como *aprendizaje* o *entrenamiento* [7].

3.1 Clasificador naive Bayes

Uno de los métodos supervisados que estiman probabilidades a posteriori es el algoritmo naive Bayes. Este clasificador asume, para una muestra x, que sus atributos $x1, x2, ..., xn$ presentan una independencia condicional dado el valor de la clase, por lo que la probabilidad condicional puede expresarse como el producto de funciones de probabilidad condicional de cada atributo por separado.

$$p(\mathrm{x}|\omega_i) = \prod_{j=1}^{n} p(x_j|\omega_i) \tag{1}$$

Usando el teorema de Bayes, la probabilidad a posteriori se escribe como

$$p(\omega_i|\mathrm{x}) = p(\omega_i) \prod_{j=1}^{n} p(x_j|\omega_i) \tag{2}$$

Finalmente, el algoritmo naive Bayes asigna una muestra x a una de las L clases existentes utilizando la función:

$$\omega^* = \ argmax_{\omega_j} p(\omega_j) \prod_{i=1}^{n} p\,(x_i|\omega_j) \tag{3}$$

3.2 Validación cruzada

Es conocida como método π o *rotación*, genera aleatoriamente una partición en *K* bloques de tamaño *N/K*. El entrenamiento (*training*) se lleva a cabo empleando *K – 1* bloques, mientras que el subconjunto restante es empleado como prueba (*test*). Este procedimiento es repetido *K* veces, eligiendo en cada iteración una parte diferente para prueba. Una extensión de este método es el llamado *stratified cross validation* (validación cruzada estratificada) con el que, para cada partición, las clases se encuentran distribuidas según sus probabilidades a priori en el conjunto original. Por otra parte, para una mejor estimación, el proceso es repetido *P* veces. La figura 1 muestra un ejemplo de validación cruzada con *K = 3* [12].

Para este trabajo se utilizó la validación cruzada estratificada con *K* = 10.

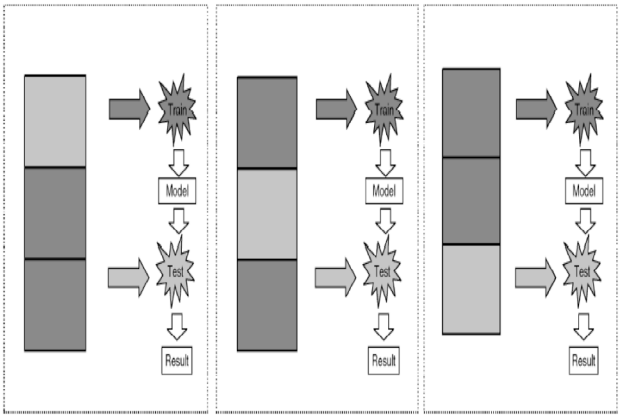

Figura 1. Validación cruzada *K=3*

3.3 Evaluación de la efectividad del clasificador

Las métricas de evaluación más empleadas para medir la efectividad de los clasificadores son la tasa de errores y la tasa de aciertos. Éstas, para un problema de dos clases, pueden obtenerse a partir de una matriz de confusión (tabla 1).

Tabla 1. Matriz de confusión para un problema de dos clases

	Positivos (total)	Negativos (total)
Positivos (clasificador)	Verdaderos positivos (VP)	Falsos positivos (FP)
Negativos (clasificador)	Falsos negativos (FN)	Verdaderos negativos (VN)

Estas tasas pueden calcularse como:

$$Acc = \frac{vp + vn}{vp + fn + vn + fp} \tag{4}$$

$$Err = \frac{fp + fn}{vp + fn + vn + fp} \tag{5}$$

Aunque estas medidas no resultan apropiadas debido a que no consideran distintos tipos de errores, ya que se muestran fuertemente sesgadas a favor de la clase mayoritaria. Por ejemplo, considerando un problema binario cuya clase positiva contiene 1% de objetos sobre el conjunto total; en tal situación, una simple estrategia para asignar todas las muestras a la clase negativa ofrecería una tasa de aciertos de 99%, sin embargo, tal clasificador carecería de valor alguno [7], lo cual ha motivado la búsqueda de medidas alternativas. Algunos ejemplos son los siguientes:

- *Tasa de verdaderos positivos (sensibilidad)*. Es el porcentaje de ejemplos positivos que son correctamente clasificados.

$$tvp = \frac{vp}{vp+fn} \tag{6}$$

- *Tasa de verdaderos negativos (especificidad)*. Es el porcentaje de ejemplos negativos que son clasificados como positivos.

$$tvn = \frac{vn}{cn + fp} \tag{7}$$

- *Tasa de falsos positivos*. Es el porcentaje de ejemplos negativos que son erróneamente clasificados.

$$tfp = \frac{fp}{vn + fp} \tag{8}$$

- *Tasa de falsos negativos*. Es el porcentaje de ejemplos positivos que son clasificados como negativos.

$$tfn = \frac{fn}{vp+fn} \tag{9}$$

- *Precisión*. Se define como el porcentaje de ejemplos que fueron etiquetados correctamente como positivos, con respecto a todas las muestras que fueron etiquetadas como tal.

$$precisión = \frac{vp}{vp + fp} \tag{10}$$

3.4 Curvas ROC

La curva ROC es una metodología de análisis desarrollada por ingenieros eléctricos y de radar durante la Segunda Guerra Mundial para resolver problemas prácticos en la detección de señales. El espacio de la curva ROC es un gráfico bidimensional que permite visualizar, organizar y seleccionar clasificadores basados en su efectividad, en nuestro caso se utilizará para comparar los diferentes parámetros para determinar con cuáles se obtienen mejores resultados. Mediante esta representación es posible conocer la relación entre los "verdaderos positivos" y los "falsos negativos". La figura 2 muestra un espacio ROC, cuyo eje Y representa la sensibilidad y el eje X la especificidad. En esta misma figura se encuentran cinco clasificadores etiquetados de la A a la E [7].

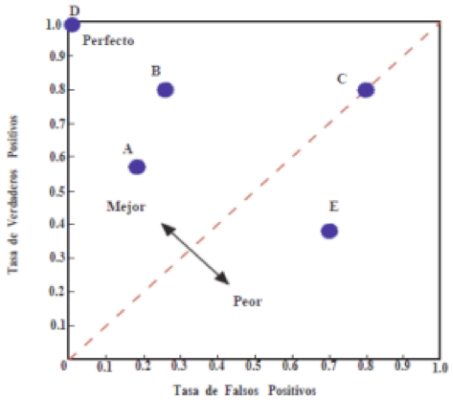

Figura 2. Espacio de una curva ROC

4 METODOLOGÍA Y DESARROLLO

En esta sección se describe la metodología utilizada en la realización del proyecto y se explica cada una de las etapas y los pasos realizados en ellas. Los programas desarrollados en cada una de estas etapas se realizaron en Java, utilizando el entorno de desarrollo integrado NetBeans IDE 6.8 y Excel (Macros). En la figura 3 se muestra el diagrama general del proyecto, del cual se destacan las actividades representadas en los módulos de la figura 4 enumerados de acuerdo con el orden en que se fueron realizando.

Las actividades de la figura 4 son explicadas a continuación:

1. Clasificación de palabras con acento (tilde) de acuerdo con las reglas que determina la Real Academia Española. Al realizar un análisis detallado de las reglas se identificaron los tipos de palabras que contienen tilde por lo que se presenta una clasificación y las diferentes formas en que se pueden solucionar.
2. Diccionario de palabras con acento gráfico. De acuerdo con las reglas expuestas en el capítulo anterior, se pueden ver principalmente dos clasificaciones de palabras con acento, las que tienen acento gráfico y las que tienen acento diacrítico. Por las definiciones dadas a cada uno de estos acentos se puede ver que el acento diacrítico se utiliza para diferenciar entre significados de las palabras, mientras que el acento gráfico no presenta ambigüedad en su significado. Por lo que la generación de un diccionario con su forma correcta es suficiente para la corrección de este tipo de palabras.
3. Palabras con acento diacrítico por tratar con naive Bayes. El principal problema que se abordó en este proyecto es el de la asignación correcta del acento diacrítico, por lo que el primer paso fue identificar las palabras que lo necesitan y determinar la forma en que se llevó a cabo la clasificación de acuerdo con el análisis de frases con palabras ambiguas.

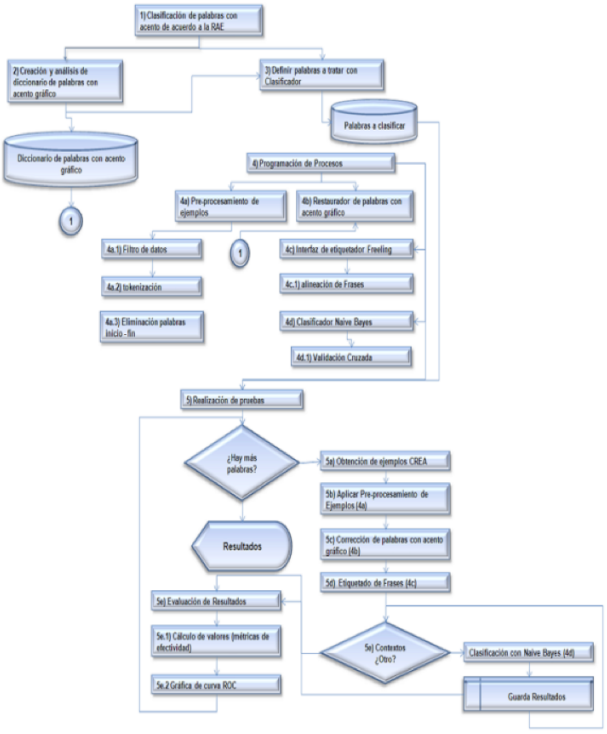

Figura 3. Diagrama general del proyecto

4. Obtención de ejemplos. En este módulo se obtuvieron ejemplos para cada una de las formas que puede tomar cada palabra ambigua, los ejemplos se extrajeron del banco de datos del Corpus de Referencia del Español Actual (CREA) (disponible en línea en http://corpus.rae.es/creanet.html).
5. Preprocesamiento de ejemplos. Partiendo del planteamiento del problema, donde se dice que la omisión de acentos es uno de los principales errores en la escritura y el problema a corregir en este proyecto, se eliminan todos los acentos contenidos en los ejemplos, para posteriormente colocarlos a las palabras que les corresponda.
6. Corrección de palabras con acento gráfico. El diccionario obtenido del módulo dos se aplicará en esta parte, que es la de restauración de acentos a las palabras de esta clase.
7. Etiquetación de ejemplos con Freeling. Las posibles combinaciones de palabras para formar frases dentro del lenguaje son un número infinito dado que la cantidad de palabras es muy grande, sin embrago siguen una misma estructura definida por la gramática del idioma con base en categorías gramaticales (etiquetas), por este motivo se realiza una etiquetación para obtener características de las palabras y clasificar sobre esa información.
8. Implementación del clasificador naive Bayes. Este módulo es el más importante de todo el proyecto, es en él donde se le asigna el sentido correcto a la palabra ambigua con base en la información contenida en las etiquetas que regresa Freeling, los resultados que regrese el clasificador son evaluados por medio de la validación cruzada, que revisa principalmente la capacidad de generalización del modelo entrenado.
9. Realización de pruebas con diferentes contextos y obtención de resultados. Se realizaron diferentes pruebas tomando en cuenta contextos variados, tomando como máxima referencia tres etiquetas hacia adelante de la palabra ambigua, tres etiquetas hacia atrás y la etiqueta de la palabra ambigua.
10. Análisis de resultados por medio de la curva ROC. Los resultados regresados por el clasificador pueden ser vistos como una matriz de confusión, de la cual se pueden obtener los valores necesarios para analizarlos por medio de este método y así determinar el mejor contexto asociado a cada palabra para desambiguarla.

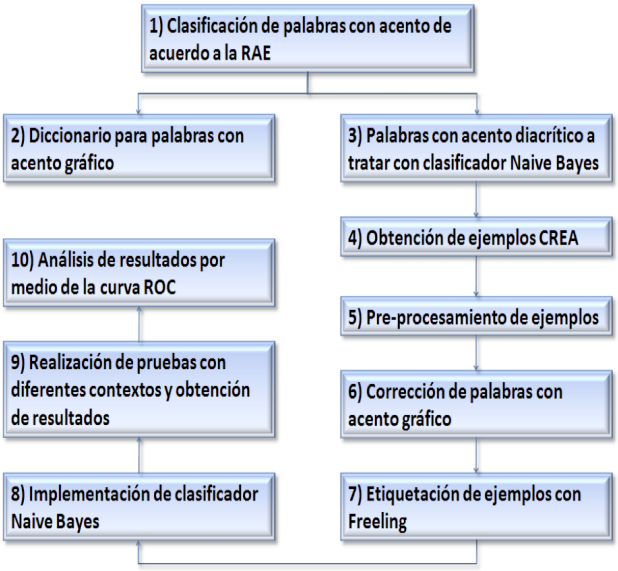

Figura 4. Actividades realizadas durante el proyecto

5 PRUEBAS Y RESULTADOS

Para la mayoría de palabras se realizaron cuatro clases de pruebas: tomando en cuenta que las clases están equilibradas, es decir, clasificando de tal manera que se tenga una probabilidad de 50% ser de una clase o de otra (tabla 2), esto para ver el comportamiento del clasificador y tomando en cuenta la proporción de acuerdo con las consultas realizadas en el CREA (tabla 3). Estas dos formas a su vez fueron divididas en dos: tomando en cuenta la palabra ambigua y sin tomarla en cuenta, dado que en estos ejemplos la palabra ambigua es etiquetada de diferentes maneras dependiendo del contexto, pero inclinándose hacia una de las dos clases. En los casos como *mi*, *te*, *tu*, *cuan* y *quien* sólo se realizaron pruebas sin tomar en cuenta la palabra ambigua, ya que toma la misma proporción que la clase.

En las tablas 2 y 3 se presenta un ejemplo de la forma en que se fueron realizando las pruebas, los valores mostrados son explicados a continuación:

- *Proporción*. Distribución de los datos en pruebas con el clasificador.
- *Contexto*. Las palabras circundantes a la palabra ambigua (desde -3 amb +3; hasta -3 +3).

- *Acc (exactitud)*. Porcentaje de ejemplos clasificados correctamente, definido por la ecuación (2.4)
- *VN (verdaderos negativos)*. Ejemplos clasificados correctamente como ejemplos sin acento.
- *FP (falsos positivos)*. Ejemplos clasificados incorrectamente como ejemplos sin acento.
- *VP (verdaderos positivos)*. Ejemplos clasificados correctamente como ejemplos con acento.
- *FN (falsos negativos)*. Ejemplos clasificados incorrectamente como ejemplos con acento.

Tabla 2. Pruebas proporción 50-50

		Proporcion 50 - 50			
contexto	Acc	VN	FP	VP	FN
amb +3	0,7557	33,9	10,1	32,6	11,4
-1 amb +3	0,8966	38,3	5,7	40,6	3,4
-2 amb +3	**0,9216**	**40,3**	**3,7**	**40,8**	**3,2**
-3 amb +3	0,9045	39,2	4,8	40,4	3,6
amb +2	0,7557	34,3	9,7	32,2	11,8
-1 amb +2	0,9148	40,4	3,6	40,1	3,9
-2 amb +2	0,9136	40,0	4,0	40,6	3,6
amb +1	0,6727	31,0	13,0	28,2	15,8
-2 amb +1	0,9136	40,0	4,0	40,4	3,6
-3 amb +1	0,8909	37,8	6,2	40,6	3,4
-1 amb	0,8977	39,4	4,6	39,6	4,4
-3 amb	0,8784	37,6	6,4	39,7	4,3
+3	0,7557	33,6	10,4	32,9	11,1
-1 +3	0,8886	38,7	5,3	39,5	4,5
-2 +3	0,9057	39,6	4,4	40,1	3,9
+2	0,7511	34,2	9,8	31,9	12,1
-2 +2	0,9170	40,6	3,4	40,1	3,9
-3 +2	0,8955	38,8	5,2	40,0	4,0
+1	0,6773	29,2	14,8	30,4	13,6
-1 +1	0,8989	39,4	4,6	39,7	4,3
-3 +1	0,9045	39,4	4,6	40,2	3,8
-1	0,8795	39,2	4,8	38,2	5,8
-2	0,8909	39,1	4,9	39,3	4,7

En estos ejemplos están marcados los mejores resultados de acuerdo con las proporciones que se tomaron en cuenta, el valor de referencia es la exactitud. En la tabla 2 para la proporción 50-50 la exactitud llega a 92.16% (contexto -2 amb +3), mientras que en la tabla 2 (proporción 97-03) la exactitud supera el valor mayor de la proporción (97%) con un valor de 98.48% (contexto -2 +2). Los siguientes son los valores que aparecerán como columnas, además de las anteriores, en las tablas de las pruebas por cada una de las palabras.

Estos valores son las métricas utilizadas para el análisis de resultados:

- *tvp (tasa de verdaderos positivos)*. Porcentaje de ejemplos positivos que son correctamente clasificados, definido por la ecuación (6)

Tabla 3. Pruebas proporción CREA

		Proporcion 97 - 03			
contexto	Acc	VN	FP	VP	FN
amb +3	0,9630	44,2	0,8	0,1	0,9
-1 amb +3	0,9739	44,4	0,6	0,4	0,6
-2 amb +3	0,9783	44,7	0,3	0,3	0,7
-3 amb +3	0,9761	44,7	0,3	0,2	0,8
amb +2	0,9761	44,6	0,4	0,3	0,7
-1 amb +2	0,9826	44,8	0,2	0,4	0,6
-2 amb +2	0,9804	44,7	0,3	0,4	0,6
amb +1	0,9717	44,7	0,3	0,0	1,0
-2 amb +1	0,9826	44,9	0,1	0,3	0,7
-3 amb +1	0,9783	44,7	0,3	0,3	0,7
-1 amb	0,9804	44,8	0,2	0,3	0,7
-3 amb	0,9717	44,7	0,3	0,0	1,0
+3	0,9587	44,1	0,9	0,0	1,0
-1 +3	0,9826	44,8	0,2	0,4	0,6
-2 +3	0,9804	44,7	0,3	0,4	0,6
+2	0,9674	44,5	0,5	0,0	1,0
-2 +2	**0,9848**	**44,9**	**0,1**	**0,4**	**0,6**
-3 +2	0,9804	44,7	0,1	0,4	0,6
+1	0,9783	45,0	0,0	0,0	1,0
-1 +1	0,9783	44,6	0,4	0,4	0,6
-3 +1	0,9783	44,8	0,2	0,2	0,8
-1	0,9826	44,9	0,1	0,3	0,7
-2	0,9804	44,8	0,2	0,3	0,7

Tabla 4. Ressultados para sustantivo/verbo "palabras con terminación -o"

	proporción	contexto	Acc	VN	FP	VP	FN	tvp	tvn	tfp	AUC
1	50 - 50	-1 amb	0,9444	46,8	2,7	46,7	2,8	0,9434	0,9455	0,0545	0,9444
2	50 - 50	-2	0,9384	46,1	3,4	46,8	2,7	0,9455	0,9313	0,0687	0,9384
3	91 - 09	-2 amb +1	0,9680	45,0	0,5	3,4	1,1	0,7556	0,9890	0,0110	0,8723
4	91 - 09	-2 +1	0,9680	45,1	0,4	3,3	1,2	0,7333	0,9912	0,0088	0,8623

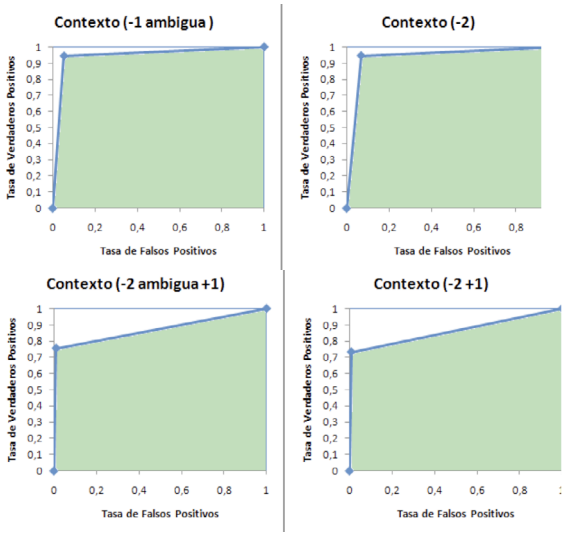

Figura 5. Área bajo la curva ROC para palabras con terminación -o

- *tvn (tasa de verdaderos negativos)*. Porcentaje de ejemplos negativos que son clasificados como positivos, definido por la ecuación (7)
- *tfp (tasa de falsos positivos)*. Porcentaje de ejemplos negativos que son erróneamente clasificados, definido por la ecuación (8)

- AUC (área bajo la curva). Área bajo la curva ROC, donde los valores que son graficados en la curva ROC son el tvp y el tfp, y el área marcada es el valor de AUC.

Dentro de la investigación se hizo un análisis con todos estos parámetros para diferentes clases de palabras, a continuación se muestra un ejemplo para palabras con terminación -o (sustantivo/verbo).

La tabla 4 presenta los mejores resultados para las palabras con terminación –o en las diferentes pruebas. En la figura 4 se pueden ver gráficamente estos resultados.

Esta prueba es la más importante dentro del proyecto, ya que se está demostrando que el trabajar con etiquetas no sólo permite generalizar las palabras en contexto a la palabra ambigua como en las pruebas anteriores, sino que también es posible utilizar etiquetas para generalizar palabras ambiguas, en este caso los verbos que, como se mencionó en el apartado anterior, se están probando diez palabras diferentes como si fueran una sola, esto porque tienen las mismas características.

6 CONCLUSIONES

Los resultados obtenidos de las diferentes palabras con acento diacrítico fueron buenos, con una exactitud que va desde 72.12% (demostrativo *cuan*) hasta 98.94% (monosílabo *se*) cuando se toman en cuenta clases balanceadas. Y tomando en cuenta clases desbalanceadas (proporción CREA) un valor AUC (área bajo la curva ROC) que va desde 67.77% (demostrativo *cuan*) hasta 96.15% (monosílabo *te*).

Los resultados más bajos que se obtuvieron fueron para el interrogativo *cuan*, los cuales se dieron debido a que en el corpus CREA, de donde se obtuvieron los datos para el proyecto, no contenía muchos ejemplos para esta palabra, lo que indica que no es muy común su uso y por lo mismo en algunos de los ejemplos están acentuadas incorrectamente.

Otra de las cosas que se puede concluir es que para los monosílabos un contexto cercano es suficiente para desambiguar, mientras que para los interrogativos es necesario un contexto mayor.

REFERENCIAS

1. Moreno Sandoval. A. (1998). Lingüstica computacional: introducción a los modelos simbólicos, estadísticos y biológicos. Madrid, Síntesis.
2. Simard, M. (1996). Automatic restoration of accents in french text . Industry Canada. Centre for Information Technology Innovation. Automatic Restoration.
3. Gelbukh. A. Galicia Haro. S. (2007). Investigaciones en análisis sintáctico para el español . Instituto Politécnico Nacional.
4. Traductores. Capítulo 1. Lenguajes. (Consultado junio de 2009). Disponible en: http://tikal.cifn.unam.mx/~jsegura/academic/traductores/Cap1.htm
5. Universitat Politècnica de Catalunya. (consultado noviembre de 2010). "Freeling Home Page". Centro de investigación TALP, Universitat Politècnica de Catalunya. Disponible en: http://nlp.lsi.upc.edu/freeling/
6. Ríos Gaona, M. (2008). "Desambiguación de sentidos de palabras usando sinónimos". ESCOM-IPN.
7. Garcia, V. (2010). "Distribuciones de clases no balanceadas: métricas, análisis de complejidad y algoritmos de aprendizaje". Tesis Doctoral. Departament de llenguatges i Sistemes Informàtics, Universitat Jaume I.
8. Monjas Llorente, M. Á. (Consultado junio 2009). "Cómo acentuar en español". Versión 2.01. 2 de febrero de 1998. Disponible en: http://www.dat.etsit.upm.es/ ~mmonjas/acentos.html
9. Real Academia Española. (2005). Diccionario panhispánico de dudas .
10. Real Academia Española. (1999). Ortografía de la Lengua Española . Edición revisada por las Academias de la lengua Española.
11. Sánchez, C. R. (2008). "Clasificación de entidades nombradas utilizando información global". Tesis de Maestría, INAOE.
12. Refaeilzadeh, P. (2008). Cross-validation . Arizona State University.

Acerca de los autores

Yesenia Nohemí González Meneses es egresada de la Licenciatura en Informática del Instituto Tecnológico de Apizaco. Obtuvo el grado de Maestra en Ciencias con Especialidad en Sistemas Computacionales por la Universidad de las Américas, Puebla. Actualmente es docente investigador del Instituto Tecnológico de Apizaco. Sus áreas de investigación son procesamiento de lenguaje natural e ingeniería de software y bases de datos.

Blanca Estela Pedroza Méndez es egresada de la Universidad Autónoma de Tlaxcala de la Licenciatura en Matemáticas Aplicadas (1993), obtuvo el grado de Maestra en Ciencias de la Computación por la Benemérita Universidad Autónoma de Puebla (1998). Actualmente es coordinadora de la Maestría en Sistemas Computacionales y docente investigador del Instituto Tecnológico de Apizaco. Sus áreas de investigación son procesamiento del lenguaje natural y tutoriales inteligentes.

Francisco López Briones es egresado de la Licenciatura en Informática del Instituto Tecnológico de Apizaco. Obtuvo el grado de Maestro en Sistemas Computacionales por el Instituto Tecnológico de Apizaco (2011). Actualmente es docente de la Universidad Tecnológica de Tlaxcala. Su área de interés es el procesamiento de lenguaje natural.

Carlos Pérez Corona estudió la Licenciatura en Informática en el Instituto Tecnológico de Apizaco (1992). Tiene una especialidad en Simulación y Control de Procesos de Ingeniería Química, en la Facultad de Ciencias Básicas, Ingeniería y Tecnología de la Universidad Autónoma de Tlaxcala (1995) y una Maestría en Inteligencia Artificial, por las instituciones LANIA-Universidad Veracruzana (1999). Actualmente es profesor investigador en la facultad de Ciencias Básicas, Ingeniería y Tecnología de la Universidad Autónoma de Tlaxcala. Es profesor de tiempo parcial de la División de Estudios de Posgrado e Investigación del Instituto Tecnológico de Apizaco. Sus áreas de interés son redes neuronales, redes bayesianas, sistemas multiagentes y redes de computadoras.

José Federico Ramírez-Cruz se graduó de Ingeniero Industrial en Electrónica en el Instituto Tecnológico de Puebla en 1993. Obtuvo el grado de Maestro en Ciencias con la especialidad de Electrónica por el Instituto Nacional de Astrofísica y Óptica en 1994 y de Doctorado en Ciencias, con especialidad en el área de Ciencias Computacionales, por el Instituto Nacional de Astrofísica y Óptica en 2003. Realizó una estancia postdoctoral en la Universidad de Texas, en El Paso, en 2011. Es docente de tiempo completo del Instituto Tecnológico de Apizaco. Sus áreas de interés son algoritmos evolutivos, procesamiento paralelo y aprendizaje automático.