Métodos de Aprendizaje Automático aplicados a la Predicción de Palabras para Portugués de Brasil *

Machine Learning Approaches applied to Brazilian Portuguese Word Prediction

Daniel Cruz Cavalieri
Teodiano Freire Bastos Filho
Universidad Federal del Espíritu Santo
Av. Fernando Ferrari, s/n.
Campus Universitario, Vitória
Espírito Santo, Brasil
{daniel,tfbastos}@ele.ufes.br

Sira Elena Palazuelos Cagigas
Javier Macías Guarasa
José L. Martín Sánchez
Universidad de Alcalá
A2, Km. 33,6. Campus Universitario
Alcalá de Henares, Madrid, España
{sira,macias,jlmartin}@depeca.uah.es

Resumen: Las personas con discapacidades físicas pueden tener serios problemas para utilizar el teclado de los ordenadores para escribir. Por esta razón suelen utilizar herramientas específicas que incluyen sistemas de ayuda a la escritura como la predicción de palabras para reducir el número de pulsaciones necesarias para escribir el texto. La predicción de palabras se puede basar en información estadística, gramatical, específica del tema y/o del usuario, etc. En este trabajo se trata de incrementar la calidad de la predicción de palabras en portugués de Brasil mejorando la predicción de la categoría de la palabra predicha. Para ello se proponen los siguientes métodos: redes neuronales artificiales, máquinas de soporte vectorial, modelos logísticos regularizados y un clasificador de Bayes. Al incorporarlos a la predicción de palabras se obtienen ahorros en el número de pulsaciones necesarias para escribir un texto entre 32,55 % y 34,58 %.
Palabras clave: Aprendizaje automático, predicción de palabras, procesamiento del lenguaje natural.

Abstract: People with physical disabilities may have serious problems to use computer keyboards to write. For this reason, they may use specific tools that include systems to assist the writing process, such us word prediction, in order to reduce the number of keystrokes needed to write the text. Word prediction may be based on different sources of information: statistical, grammatical, specific of the subject or/and the user, etc. In this paper we increase the quality of the word prediction in Brazilian Portuguese by improving the prediction of the part of speech (POS) of the predicted word. We propose the following methods to predict the POS: artificial neural networks, support vector machines, regularized logistic models and a naïve Bayes classifier. When included in the word prediction system, they save between 32.55 % and 34,58 % of the keystrokes needed to write the text.
Keywords: Machine learning, word prediction, natural language processing.

1. Introducción

Las personas con discapacidad pueden experimentar diversas dificultades a la hora de comunicarse o escribir textos (Cavalieri et al., 2008) dependiendo de su tipo y grado de discapacidad y es común que utilicen ordenadores como ayudas técnicas que faciliten esos procesos.

Los ordenadores son herramientas muy versátiles, que se pueden adaptar a las características de las personas con necesidades diferentes, permitiendo, por ejemplo, conectar distinto tipo de pulsadores para personas que no pueden utilizar el teclado convencional o el ratón, proporcionando salida de voz para las personas que no se pueden comunicar, etc.

* Los autores agradecen a "CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior", así como a "FACITEC - Fundação de Apoio à Ciência e Tecnologia do Município de Vitória" el soporte financiero que ha permitido desarrollar este trabajo.

Las aplicaciones que se controlan con pulsadores normalmente utilizan técnicas de barrido: en cada menú se van resaltando secuencialmente las opciones disponibles y el usuario presiona el pulsador cuando se resalta la opción que desea. Como ejemplo, en la Figura 1 se muestra una aplicación que se controla con un pulsador: un teclado virtual.

←	Esp	↵	⊗	Supr	Signos	⇐	⇓	⇒	⇑	Tecla
e	a	i	⊗	r	t	b	Tab	May	Shift	Teclado
o	s	d	⊗	u	q	v	Ctrl	Alt	Foco	Telecomunicación
n	l	m	⊗	g	y	z	Inicio	Fin	Menu	Teleférico
c	p	f	⊗	h	j	k	ReP	AvP	Salir	Telefonista
´	,	.	⊗	x	ñ	w	F1-12	Teléfono		Teletexto

Figura 1: Pantalla de un teclado virtual controlado por pulsador.

Si el usuario utiliza una aplicación de este tipo para escribir, para elegir cada letra del texto debe esperar a que se barran todas las anteriores, y este proceso es muy lento y puede resultar muy cansado. Esta lentitud es especialmente relevante si lo está utilizando para comunicarse y desea mantener una conversación (Garay-Vitoria y González-Abascal, 1997).

La predicción de palabras es una técnica comúnmente incluida en las herramientas utilizadas por personas con discapacidad para escribir con el objetivo de reducir la cantidad de pulsaciones necesarias para introducir el texto. En la columna derecha de la Figura 1 podemos ver las palabras que se predicen cuando el usuario ha escrito "Te".

Este trabajo explora técnicas de predicción de palabras para portugués de Brasil, con los objetivos futuros de introducirlas en el sistema PREDWIN (Palazuelos, 2001), ya desarrollado para el idioma español, y de utilizarlas para la predicción en otros idiomas.

La estructura del artículo es la siguiente: en primer lugar se explica brevemente el funcionamiento del sistema de predicción de palabras. A continuación se describen los métodos de predicción de categorías que se evalúan en este artículo. El siguiente apartado trata sobre la evaluación automática del sistema, describiendo el procedimiento, los resultados y su discusión, y por último las conclusiones.

2. *Predicción de palabras*

La Figura 2 presenta la arquitectura general del sistema utilizado para la evaluación de los métodos de predicción de palabras. Este sistema fue desarrollado para su uso en español (Palazuelos, 2001) y adaptado para la predicción de palabras en portugués de Brasil en este trabajo.

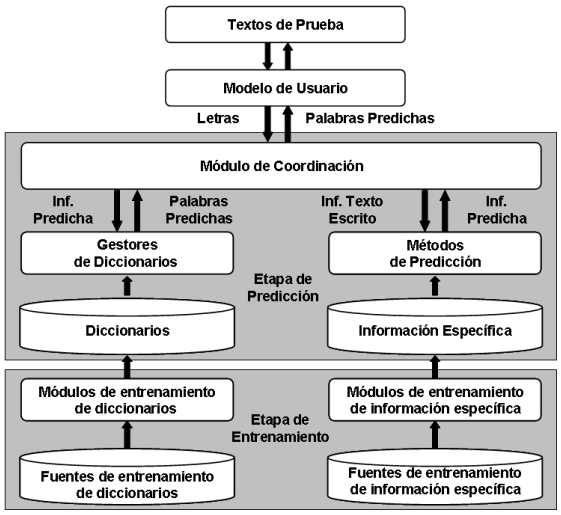

Figura 2: Arquitectura general del sistema de predicción utilizado. Adaptado de (Palazuelos, 2001).

A continuación se describe brevemente cada uno de los bloques:

- **Textos de prueba:** Son textos, extraídos principalmente de Internet, cuyo contenido es fundamentalmente periodístico y literario.
- **Modelo de usuario:** Es el algoritmo que utiliza el sistema de evaluación automática para simular que hay un usuario intentando escribir el texto de prueba. Elige las letras del texto una a una y se las envía al sistema de predicción de palabras. Por cada letra el sistema de predicción devuelve una lista de palabras predichas. Si la palabra deseada está en esa lista, el modelo de usuario la elige. Si no, continúa con la siguiente letra. Este proceso se repite hasta el texto de prueba se acaba.
- **Módulo de coordinación:** Parte encargada de procesar las letras provenientes del modelo de usuario y organizar el flujo de trabajo del sistema de predicción de palabras de forma que se utilicen todos los diccionarios y métodos de predicción en el orden adecuado y se genere la lista de palabras óptima con los algoritmos y diccionarios disponibles.

Este flujo de trabajo ha sido determinado por experimentación. Una vez generada la lista de palabras predichas, la envía al modelo de usuario.

- **Gestores de diccionarios:** Encargados de seleccionar las palabras de cada diccionario que cumplen las restricciones impuestas por los métodos de predicción (por ejemplo, seleccionan las palabras más frecuentes de las categorías preferidas con los rasgos adecuados).
- **Diccionarios:** El diccionario general contiene las palabras y la información de cada una necesaria para apoyar los métodos de predicción: categorías gramaticales, rasgos y frecuencia. Los diccionarios personales de usuario (diccionario de texto en curso y temáticos) contienen también información sobre las secuencias de palabras (bigramas y trigramas) que han aparecido en los textos de entrenamiento.
- **Módulos de entrenamiento de los diccionarios:** Engloban los procedimientos automáticos y manuales necesarios para generar los diccionarios utilizados por el sistema, a partir de las fuentes de entrenamiento disponibles.
- **Fuentes de entrenamiento de los diccionarios**. En este trabajo, el **diccionario general** se ha generado a partir de un conjunto de 730 textos periodísticos categorizados, pertenecientes al periódico "Folha de São Paulo" del período de 1994 a 1995 del corpus portugués CHAVE (Santos y Rocha, 2004). La información morfosintáctica fue obtenida mediante un categorizador automático llamado PALAVRAS (Bick, 2000). Se han elegido 76 categorías: las 10 clases gramaticales básicas del portugués de Brasil (substantivo, verbo, adjetivo, etc.), añadiendo sus posibles flexiones (género, número, grado, etc.), y algunos símbolos (punto, coma y guión). Esa elección de categorías tiene como objetivo principal facilitar el aprendizaje de la estructura de la frase por los métodos de predicción. Para seleccionar este conjunto de categorías se realizaron experimentos con combinaciones diferentes de categorías hasta que se consiguió el conjunto óptimo para la predicción de palabras. Los demás diccionarios (**diccionario de texto en curso y temáticos**) se generan a partir de textos escritos por los usuarios o seleccionados por ellos sobre temas particulares (deportes, política, etc.).
- **Métodos de predicción:** Son los distintos algoritmos encargados de averiguar, a partir de las informaciones proporcionadas por el módulo de coordinación (frecuencia de las categorías anteriores, categorías gramaticales de las palabras anteriores, etc.), el conjunto de categorías y rasgos que puede tener la palabra siguiente y su probabilidad. Pueden estar basados en redes neuronales, etc.
- **Información específica:** Es la información que necesita cada método de predicción para funcionar (secuencias de categorías gramaticales, frecuencia gramatical, conocimiento lingüístico específico, etc.). Por ejemplo, un analizador necesita las reglas gramaticales.
- **Módulo de entrenamiento de información específica:** Engloba los procedimientos, automáticos o manuales, necesarios para generar la información específica que necesita cada método de predicción utilizado.
- **Fuentes de entrenamiento de la información específica:** Es el conjunto de fuentes de información necesarias para generar la información específica. Por ejemplo, para entrenar los pesos de las redes neuronales se necesitan textos categorizados.

Una explicación más detallada de cada una de las partes del sistema para español puede encontrarse en (Palazuelos, 2001).

3. Métodos de predicción

En este apartado se describen brevemente los métodos de predicción propuestos en este artículo. Todos ellos son métodos de predicción de categorías gramaticales, que en función de la categoría de las palabras anteriores escritas por el usuario predicen la categoría de la palabra siguiente. Para integrarlos en el sistema de predicción de palabras visto anteriormente, el módulo de coordinación debe categorizar las palabras anteriores y proporcionarle a los métodos de predicción la lista

de categorías anteriores. Éstos aplicarán el algoritmo de predicción de categorías y obtendrán el listado de posibles categorías siguientes y la probabilidad de cada una. Por último, el módulo de coordinación proporcionará ese listado de categorías siguientes a los gestores de diccionarios para que éstos obtengan el listado de palabras adecuado.

3.1. Redes Neuronales Artificiales

La red básica elegida es una red neuronal de Elman con tres capas (Elman, 1990), una capa oculta recurrente y con función de transferencia tan-sigmoide y neuronas con función log-sigmoide en su capa de salida. Así, podemos obtener las categorías más probables, como se muestra en la Figura 3.

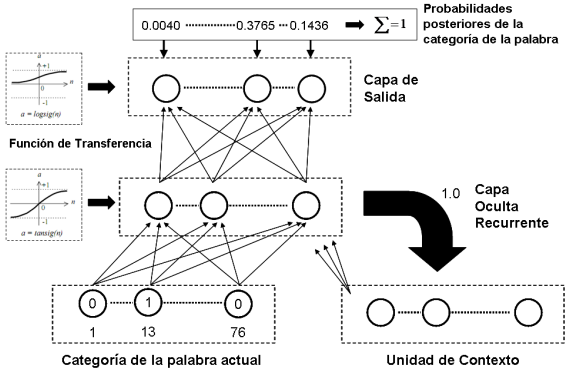

Figura 3: Ejemplo de una red neuronal recurrente de Elman con una función log-sigmoide en la capa de salida, utilizando secuencias de 2 categorías y 76 categorías de palabras.

El objetivo principal a lograr con esta topología es una red con una memoria a corto plazo, así, el contexto de la categoría de la última palabra vista puede ser parametrizado por el proceso de aprendizaje. Dado que esta red se entrena con la categoría de la palabra siguiente en función de la anterior, en la capa oculta se espera que se aprendan algunas estructuras lingüísticas que relacionan secuencias de pares de categorías del texto (Nakamura et al., 1990).

Como entrada de la red se utiliza el listado de las categorías de las palabras del corpus de texto. Las entradas y la salida se codifican utilizando 76 bits, que es el número de categorías gramaticales utilizadas. Para cada categoría, un solo bit es activado (1), su posición es diferente para cada categoría gramatical y los demás bits están a cero (0), como se muestra en la Figura 4. Esta representación permite a la red identificar perfectamente cada categoría, y, en un futuro, permitirá incluir la información de las distintas categorías que pueden estar asociadas a cada palabra, y su frecuencia.

Se utilizan redes neuronales *feed-forward* entrenadas con los algoritmos *Resilient Backpropagation* (RP), *Scaled Conjugate Gradient* (SCG) y *Gradient Descent Adaptive with Momentum*. Se realizaron pruebas variando el número de neuronas de la capa oculta entre el 15 y el 75 con incrementos de 5 neuronas, y variando el número de categorías anteriores consideradas entre 1 y 3. Esto permitió determinar la mejor estructura para uso en futuras aplicaciones.

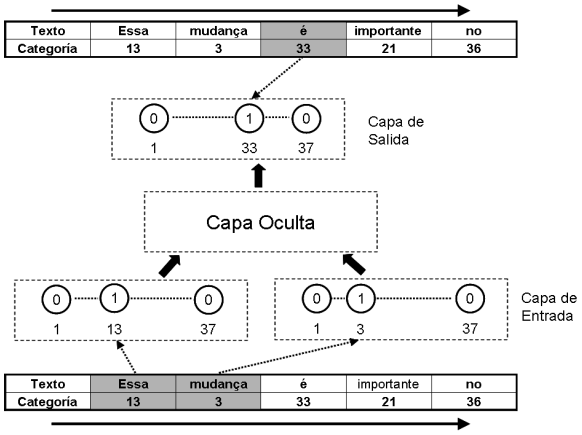

Figura 4: Entrenamiento adoptado para una red neuronal de Elman. Ejemplo utilizando secuencias de 3 categorías. Adaptado de (Nakamura et al., 1990)

3.2. Máquinas de Soporte Vectorial

Las máquinas de soporte vectorial o máquinas de vectores soporte, del inglés *Support Vector Machines* (SVMs) (Burges, 1998; Christianini y Shawe-Taylor, 2000), son un conjunto de algoritmos de aprendizaje estadístico utilizados en numerosas aplicaciones reales (Vert, 2002; Joachims, 1998; Osuna, Freund, y Girosi, 1997).

Las SVMs básicamente aprenden cómo clasificar objetos $\mathbf{x} \in \mathbf{R^n}$ en dos o más clases a partir de un conjunto etiquetado de vectores $\{(\mathbf{x_i}, \mathbf{y_i})\}$, con $i = 1, ..., m$, siendo $\mathbf{x_i}$ el vector de características de entrenamiento e y_i la clase a la que pertenece. El clasificador resultante se basa en la siguiente función de decisión:

$$f(x) = \sum_{i=1}^{m} y_i \alpha_i K(\mathbf{x_i}, \mathbf{x}) + b, \qquad (1)$$

donde $\mathbf{x}$ es cualquier nuevo vector de características a ser clasificado, $K(,)$ es la función *kernel*, $\{\alpha_1, \ldots, \alpha_m\}$ y b son los parámetros de la función de decisión encontrados durante la etapa de entrenamiento, necesarios para la optimización del problema o, en otras palabras, necesarios para garantizar el error mínimo (Vert, 2002). Cualquier vector $\mathbf{x_i}$ que se corresponde con un $\alpha_i > 0$ es un vector de soporte del hiperplano optimizado (Sundarkantham y Shalinie, 2007) que permitirá realizar la separación entre distintas clases.

Los *kernels* K son una familia de funciones, lineales o no, que proyectan los datos de entrada en un espacio de *Hilbert* de dimensión superior al espacio de características, donde vectores linealmente no separables en este último espacio pueden ser linealmente separados en el espacio trasformado. A pesar de investigaciones recientes en la búsqueda de nuevos *kernels*, se pueden destacar las siguientes cuatro funciones *kernel* principales:

- lineal: $K(\mathbf{x_i}, \mathbf{x}) = \mathbf{x_i}^T \cdot \mathbf{x}$;
- polinomial: $K(\mathbf{x_i}, \mathbf{x}) = (\gamma \mathbf{x_i}^T \cdot \mathbf{x} + r)^d, \gamma > 0$;
- RBF: $K(\mathbf{x_i}, \mathbf{x}) = exp(-\gamma \|\mathbf{x_i} - \mathbf{x}\|^2), \gamma > 0$;
- sigmoide: $K(\mathbf{x_i}, \mathbf{x}) = tanh(\gamma \mathbf{x_i}^T \cdot \mathbf{x} + r)$.

donde γ, d y r son parámetros del *kernel*.

En este trabajo fueron evaluadas SVMs con las cuatro funciones *kernel* básicas presentadas anteriormente, siendo elegida la que proporcionó mejores resultados. Para eso, nuevamente fueron utilizados como datos de entrada las categorías desde la primera palabra en la frase hasta la última, siendo formadas secuencias de entrenamiento con 2, 3, 4 y 5 categorías (1, 2, 3 y 4 categorías anteriores y la de la palabra actual). Esas secuencias fueron nuevamente codificadas en un vector de 76 *bits*, donde cada *bit* representa una clase gramatical diferente y para cada unidad correspondiente a la categoría, solamente una es activada (1), y las otras estarán a cero (0). Con eso, se forma un vector de características de tamaño igual al numero de categorías anteriores multiplicado por el número máximo de categorías (76). En la Figura 5 se muestra un ejemplo para 4 categorías.

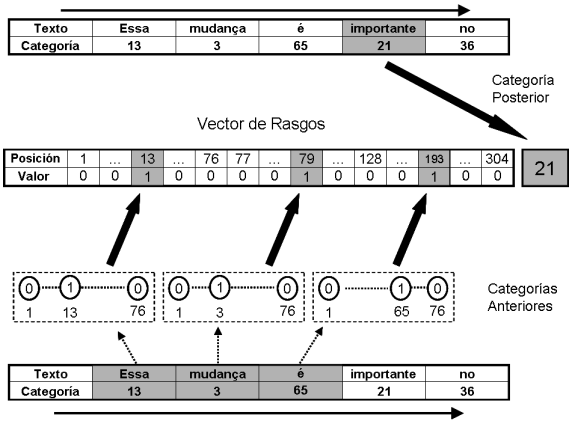

Figura 5: Entrenamiento de las SVMs. Ejemplo con secuencias de 4 categorías (3 anteriores y la actual).

3.3. Modelo de Regresión Logística

La regresión logística es una técnica estadística que tiene como objetivo generar, a partir de un conjunto de observaciones, un modelo que permita la predicción de valores tomados por una variable categórica, frecuentemente binaria, a partir de una serie de variables explicativas continuas y/o binarias. Es un modelo de regresión para variables dependientes o de respuesta con distribución binomial (Figueira, 2006). Es un modelo lineal generalizado, cuya función es:

$$P(y = class|\mathbf{x}, \mathbf{w}) = \frac{1}{1 + e^{-y(\mathbf{w}^T \mathbf{x} + b)}}, \qquad (2)$$

donde $\mathbf{x} \in R^n$, representa el vector de características, $y \in \{1, \ldots, 76\}$ representa la clase de cada vector de características, $\mathbf{w}$ y b son los parámetros del modelo.

Para estimar los valores de $\mathbf{w}$ y b, una solución sería resolver el siguiente problema de optimización (Ng, 2004).

$$min_{\mathbf{w},b} \sum_{i=1}^{m} \log(1 + e^{-y(\mathbf{w}^T \mathbf{x} + b)}) - \alpha R(\mathbf{w}), \qquad (3)$$

donde $R(\mathbf{w})$ es el término de regularización y tiene como objetivo penalizar los parámetros/pesos muy grandes, y $\alpha \geq 0$ es el término para elegir entre ajustar bien la curva que separa los datos y obtener términos regularizados pequeños (Ng, 2004). En este trabajo fue elegido un modelo con término de

regularización $L2$ y la determinación de los parámetros optimizados ($\mathbf{w}$ y b) puede ser encontrada en (Fan et al., 2008) y (Lin, Weng, y Keerthi, 2007).

Los datos de entrada utilizados para entrenamiento y prueba del modelo de regresión logística tienen el mismo formato de los datos utilizados para entrenar las SVMs, mostrados en la Figura 5. Se han generado modelos utilizando 1, 2, 3 y 4 categorías gramaticales anteriores más la de la palabra actual, avanzando desde la primera palabra de la frase hasta la última.

3.4. Clasificador Bayesiano Naïve

El clasificador Bayesiano Naïve es uno de los más eficaces y eficientes algoritmos de clasificación, con un simple clasificador probabilístico basado en la aplicación del teorema de Bayes con una fuerte hipótesis de independencia (Zhang y Su, 2008).

Asumiendo que C_t es la categoría actual y $C_{t-n}, \ldots, C_{t-2}, C_{t-1}$ son las n categorías anteriores, para el clasificador de Bayes la categoría siguiente es C, tal que:

$$g(C) = \text{argmax}_{C_t}\ p(C_t) \prod_{i=1}^{n} p(C_t|C_{t-i}), \quad (4)$$

donde

$$p(C_t|C_{t-i}) = \frac{Frec(C_{t-i}C_t)}{Frec(C_t)},\ i = 1, 2, \ldots, n. \quad (5)$$

$Freq(C_{t-i}C_t)$ es el número de veces que las categorías C_{t-i} y C_t aparecen en el texto de entrenamiento separadas i posiciones, y $Freq(C_t)$ es el número de veces que la categoría C_t aparece. Se han generado modelos con i igual a 1, 2, 3, y 4.

4. Resultados

En este apartado se muestra el rendimiento de los métodos de predicción descritos cuando se integran en la predicción de palabras, su objetivo final.

Para ello, los métodos de predicción de categorías fueron incorporados al sistema general de predicción de palabras y se realizaron experimentos con el sistema global. En los experimentos se evaluaron los parámetros que habitualmente se contabilizan en los sistemas de predicción de palabras:

- el **ahorro de pulsaciones** que se produce con respecto al sistema sin ayuda de predicción, que valora la disminución del esfuerzo físico necesario para escribir el texto, y
- el **porcentaje de palabras** que se predicen antes de que el usuario las acabe de escribir por completo, que puede dar una idea del grado de apoyo que la predicción proporciona a una persona con problemas lingüísticos.

Ambos resultados se proporcionan en valor absoluto y como mejora relativa con respecto a la escritura del texto sin predicción gramatical, y con información de las bandas de fiabilidad. Además, también se añade información del tiempo que tarda cada algoritmo en predecir cada palabra.

En la evaluación de la predicción de palabras hay que considerar que un cambio en cualquier parámetro de la configuración del experimento (el idioma, los textos de prueba, la configuración del sistema de predicción o de la propia interfaz donde esté instalado) puede llevar a variaciones muy significativas en los resultados. En (Palazuelos, 2001) se puede encontrar una exposición y discusión de los factores que afectan a los resultados de la predicción de un sistema determinado. Por esta razón, para poder verificar realmente la eficacia de un determinado método o diccionario, y, para poder hacer comparaciones entre ellos, es necesario realizar los experimentos en series en las que lo único que varíe sea el o los factores concretos a evaluar, ya que, en caso contrario, la influencia de las demás modificaciones puede cambiar totalmente los resultados.

Esta variabilidad implica también que sea muy difícil comparar nuestros sistemas con otros (que, por ejemplo, pueden ser para otro idioma, incluir más o menos palabras en la lista de palabras predichas, o utilizar textos de prueba más o menos adaptados a los textos de entrenamiento).

A continuación se muestran los resultados de una serie representativa de experimentos, en que se comparan todos los métodos descritos en este artículo. En esta prueba se utilizaron textos periodísticos procedentes de Internet con 70694 palabras, siendo necesarias 449194 pulsaciones para su escritura sin apoyo de predicción de palabras. Ninguno de los textos de prueba solapaba con los uti-

Tabla 1: Resultados obtenidos en la etapa de prueba por los distintos métodos de predicción.

Método de predicción	**# Categorías anteriores**	**# Palabras predichas (± 0,31 %)**	**Mejora relativa (%)**	**# Pulsaciones ahorradas (± 0,14 %)**	**Mejora relativa (%)**	**Tiempo de procesamiento (ms/pal.)**
Unigramas	0	54267 (76,77 %)	Base	140394 (31,25 %)	Base	3,06
RNA	2	54354 (76,89 %)	0,16	152411 (33.93 %)	8,56	37,02
	3	54549 (77,16 %)	0,52	153659 (34,21 %)	9,45	57,70
	4	54535 (77,14 %)	0,49	153197 (34,10 %)	9,12	77,53
SVM	2	54858 (77,60 %)	1,09	153813 (34,24 %)	9,56	188,75
	3	54840 (77,58 %)	1,06	154214 (34,33 %)	9,84	186,13
	4	54841 (77,58 %)	1,06	154277 (34,35 %)	9,89	124,60
	5	54856 (77,60 %)	1,09	154290 (34,35 %)	9,90	134,35
LR	2	54558 (77,18 %)	0,54	154579 (34,41 %)	10,10	121,58
	3	54553 (77,17 %)	0,53	154990 (34,50 %)	10,40	114,92
	4	54522 (77,13 %)	0,47	155203 (34.55 %)	10,55	107,72
	5	54512 (77,11 %)	0,45	155348 (34.58 %)	10,65	108,09
Bayes	2	54474 (77,06 %)	0,38	154041 (34,29 %)	9,72	6,46
	3	54354 (76,89 %)	0,16	153903 (34,26 %)	9,62	6,62
	4	54334 (76,86 %)	0,12	153936 (34.27 %)	9,65	6,45
	5	54308 (76,82 %)	0,08	153870 (34.25 %)	9,60	6,51

lizados en el entrenamiento. Se predicen 5 palabras como máximo, ordenándolas por su probabilidad.

Como base de comparación se utilizó un método de predicción sin información gramatical, basado solamente en la frecuencia absoluta de las palabras (unigramas). En este método el conjunto de palabras predichas está formado por las palabras más frecuentes que empiezan por la parte ya escrita de la palabra en curso (Palazuelos, 2001).

Los resultados finales se presentan en la Tabla 1. En ella se puede observar, con respecto al ahorro de pulsaciones, que los mejores resultados fueron proporcionados por los modelos LR, en particular los entrenados con secuencias de 5 categorías. Así, se puede inferir que la predicción de categorías de palabras en portugués de Brasil puede ser un problema linealmente separable y cuya distribución de probabilidades de categorías está próxima de la distribución normal (Figueira, 2006).

Se puede observar, en general, que la tendencia de todos los métodos (excepto el bayesiano) es mejorar a medida que se aumenta el número de categorías que se consideran. En el caso del modelo bayesiano, el rendimiento empeora cuando la categoría anterior utilizada está más alejada de la actual. Esto prueba que la categoría que aporta más información sobre la posible categoría siguiente es la categoría inmediatamente anterior, y que a medida que las categorías están más alejadas tienen menor capacidad para predecir cual será la categoría siguiente (hay que tener en cuenta que al considerar una categoría alejada en este método no se tienen en cuenta las categorías que están entre la anterior y la actual).

Con respecto al porcentaje de palabras predichas (como ocurre en español), cuando aumenta el número de pulsaciones ahorradas disminuye el número de palabras predichas. La razón es que en los métodos sin apoyo gramatical se predicen muy pronto las palabras más frecuentes, que son muy cortas (artículos, preposiciones). Los métodos con apoyo gramatical predicen mejor las palabras largas, que son más productivas en cuanto a número de pulsaciones, aunque predigan menos palabras en total.

Aunque el tiempo de computación de algunos métodos parezca grande, hay que considerar que las personas con discapacidad manejan en las aplicaciones tiempos de barrido mucho mayores, por lo que estos márgenes no son problemáticos.

5. Conclusiones y Trabajos Futuros

Este trabajo trata del uso de redes neuronales artificiales, máquinas de soporte vectorial, modelos de regularización logística y un clasificador de Bayes en la predicción de categorías para el portugués de Brasil.

Estos métodos de predicción de categorías se han integrado en el sistema de predicción de palabras, consiguiendo una mejora relativa de hasta un 10,65 % de pulsaciones ahorradas con respecto al método de predicción sin información gramatical (unigramas). En general sus resultados mejoran al aumentar

el número de categorías consideradas, aunque esto no ocurre con el clasificador bayesiano, ya que las categorías más alejadas tienen menos información para predecir.

Los resultados aquí presentados abren camino para otros trabajos futuros, como, por ejemplo, un estudio más profundo acerca del número de categorías anteriores óptimo para cada método de predicción, gestión de la ambigüedad de las palabras y la fusión de todos estos algoritmos.

Actualmente se están realizando pruebas de estos métodos para español con resultados preliminares altamente satisfactorios.

Bibliografía

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. tesis, Aarhus University, Aarhus, Denmark, November.

Burges, Christopher J. C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167.

Cavalieri, Daniel C., Teodiano F. Bastos-Filho, Mário Sarcinelli-Filho, y Sira Elena Palazuelos Cagigas. 2008. Redes neuronales artificiales para predicción de categorias de palabras en portugués de brasil. En *V Congreso IBERDISCAP*, Cartagena, Colombia.

Christianini, N. y Shawe-Taylor. 2000. *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.

Elman, Jeffrey L. 1990. Finding structure in time. Cognitive Science 14, University of California, California, San Diego.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, y Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Figueira, Cleonis Viater. 2006. Modelos de regressão logística. Master's thesis, Universidade Federal do Rio Grande do Sul - UFRGS, Instituto de Matemática da UFRGS, Porto Alegre, Brasil, March.

Garay-Vitoria, N. y J. González-Abascal. 1997. Intelligent word prediction to enhance text input rate (a syntactic analysis based word prediction aid for people with severe motor speech disability). En *Annual International Conference on Intelligent User Interfaces*, páginas 241-247.

Joachims, Thorsten. 1998. Text categorization with support vector machines: Learning with many relevant features. páginas 137–142. Springer Verlag.

Lin, Chih-jen, Ruby C. Weng, y S. Sathiya Keerthi. 2007. Trust region newton method for large-scale logistic regression. En *An Interior-Point Method For Large-Scale l1-Regularized Logistic Regression*.

Nakamura, Masami, Katsuteru Maruyama F, Takeshi Kawabata F, y Kiyohiro Shikano Tit. 1990. Neural network approach to word category prediction for english texts. En *Helsinki University*, páginas 213–218.

Ng, Andrew Y. 2004. Feature selection, l1 vs. l2 regularization, and rotational invariance. En *In ICML*.

Osuna, Edgar, Robert Freund, y Federico Girosi. 1997. Training support vector machines: An application to face detection. páginas 130–136.

Palazuelos, Sira Elena. 2001. *Contribution to Word Prediction in Spanish and its Integration in Technical Aids for People with Physical Disabilities*. Ph.D. tesis, Universidad Politécnica de Madrid, Madrid, España.

Santos, Diana y Paulo Rocha. 2004. The key to the first clef in portuguese: Topics, questions and answers in chave. En *5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004*, páginas 821-832, Bath, UK, September 15-17.

Sundarkantham, K. y S. M. Shalinie. 2007. Word preditor using natural language grammar induction technique. *Journal of Theoretical and Applied Information Technology*, 3(3):1–8.

Vert, J. p. 2002. Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. En *Proceedings of the Pacific Symposium on Biocomputing*, páginas 649-660. World Scientific.

Zhang, Harry y Jiang Su. 2008. Naive bayes for optimal ranking. *J. Exp. Theor. Artif. Intell.*, 20(2):79–93.