Clasificación de Textos Adaptada para

Francesc Alías, Xavier Gonzalvo, Xavier Sevillano,
Joan Claudi Socoró, José Antonio Montero y David García
Departamento de Comunicaciones y Teoría de la Señal
Enginyeria i Arquitectura La Salle, Universidad Ramon Llull
Pg. Bonanova 8, 08022 Barcelona
{falias, gonzalvo, xavis, jclaudi, montero, dgarcia}@salle.url.edu

Resumen: En este trabajo se presenta un sistema de clasificación de textos adaptado a las necesidades que plantea la conversión de texto en habla multidominio. Este sistema, que es una evolución de una propuesta anterior basada en la representación de los textos mediante un grafo de nodos ponderados, ha sido desarrollado para mejorar la eficiencia de clasificación de textos cortos, así como para minimizar el coste computacional de la misma. Para ello, se trabaja sobre el espacio de comparación definido por el texto a clasificar en lugar de utilizar el construido a partir de los documentos de entrenamiento. Los experimentos de clasificación desarrollados sobre un corpus de textos publicitarios muestran la consecución de los objetivos planteados.
Palabras clave: Clasificación de textos, textos cortos, coste computacional, conversión de texto en habla

Abstract: This paper introduces a text classification system tuned to cope with the requirements of multi-domain text-to-speech synthesis. This method, based on a previous system which represents texts by means of a weighted graph, has been developed to improve the classification efficiency for small texts and to minimize its computational cost. To that effect, the comparison space is built from the input text instead of being built from the training documents. Classification experiments conducted on an advertising text corpus show the achievement of the posed goals.
Keywords: Text classification, small texts, computational cost, text-to-speech synthesis

1. *Introducción*

El propósito final de la conversión de texto en habla (CTH) es la generación de habla sintética completamente natural a partir de un texto de entrada cualquiera. Históricamente, se han seguido dos estrategias para lograr este objetivo (Yi y Glass, 1998): *i)* la que prima la flexibilidad de la conversión ante la calidad de la síntesis, dando lugar a los sistemas de conversión de texto en habla de propósito general (CTH-PG); y *ii)* la que antepone la naturalidad de la síntesis a la generalidad de la CTH, conocida como conversión de texto en habla de dominio restringido (CTH-DR). Siguiendo esta segunda estrategia, se presentó la conversión de texto en habla multidominio (CTH-MD) (Alías, Iriondo, y Barnola, 2003; Alías et al., 2003), que persigue conseguir una calidad sintética equivalente a la de los sistemas de CTH-DR, aumentando su flexibilidad al considerar distintos *dominios* (estilos de locución, emociones, temáticas, etc.) para la síntesis.

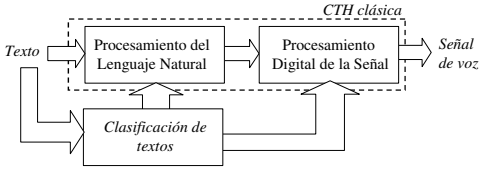


Figura 1: Diagrama de bloques de la arquitectura de un CTH-MD.

En este contexto, es necesario que el sistema de CTH-MD conozca, durante el proceso de conversión de texto en habla, qué dominio o dominios son los más *adecuados* para poder sintetizar el texto de entrada con la mayor naturalidad posible. Para ello, el sistema de CTH-MD incorpora un módulo de clasifica-

* Trabajo financiado en parte por el proyecto IntegraTV-4all (FIT-350301-2004-2) del MCyT.

ción de textos a la arquitectura clásica de los sistemas de CTH (ver Figura 1).

En este trabajo se presenta una evolución del sistema de clasificación de textos (CT) presentado en trabajos anteriores (Alías, Iriondo, y Barnola, 2003; Alías et al., 2003), con el objetivo de optimizar dos de los elementos clave de la CT en el contexto de la CTH-MD: la eficiencia de clasificación, principalmente para textos extremadamente cortos (p.ej. 1 frase por documento), y el coste computacional del proceso de clasificación.

2. *Clasificación automática de dominios*

La clasificación de dominios para CTH-MD se encuentra, de algún modo, entre el enfoque clásico de la clasificación temática de documentos y el problema de la clasificación estilística (no temática) de textos (p.ej. determinación de la autoría o el género de un texto) (Sebastiani, 2002; Sebastiani, 2005). Por un lado, la información temática es importante para organizar los textos que forman el corpus, pero por otro lado, fijarse únicamente en esta información no parece suficiente para determinar la mejor manera de pronunciarlos (prosodia, pausas, etc.), por lo que resulta interesante que el modelado de los textos incorpore información que tome en consideración la secuencialidad inherente del habla (p.ej. coarticulación entre sonidos).

Con este objetivo, los textos se modelan mediante un *grafo* de nodos interconectados (con tantos nodos como palabras aparecen en el texto), entrelazados mediante conexiones ponderadas (coocurrencias de palabras). Esta estructura se denomina Red Relacional Asociativa (RRA) y fue inicialmente utilizada para la representación visual de documentos (Rennison, 1994) (ver (Alías, Iriondo, y Barnola, 2003) para más detalles). Gracias a su tipología, la RRA no sólo considera los términos (junto a los signos de puntuación) que aparecen en el texto (*bolsa de palabras* (Sebastiani, 2002)), sino que contempla todas sus relaciones, modelando así la continuidad y la estructura de los textos.

2.1. Parametrización y modelado del texto

Como se ha comentado, la parametrización del texto debe contemplar tanto la temática como la secuencialidad del mensaje. Así pues se utilizan, por un lado, parámetros temáticos (frecuencia del término -TF- y frecuencia inversa del término en los documentos -IDF- (Sebastiani, 2002)), y por otro, parámetros que denominamos estructurales (frecuencia de coocurrencia de las palabras -COF- y número de palabras consecutivas coincidentes entre los textos comparados (Alías et al., 2003)). Además, en este trabajo se introduce un nuevo parámetro temático denominado *inverse word frequency* (IWF), definido según la ecuación:

$$iwf_i = log\left(\frac{M}{tf_i}\right), \forall\, tf_i > 0 \qquad (1)$$

donde M es el número de palabras del texto y tf_i el número de veces que el término i aparece en ese texto. IWF se puede interpretar como una aproximación de IDF, ya que pondera cada término según su peso dentro de *cada* documento, en lugar de considerar su distribución a lo largo de *toda* la colección.

Una vez parametrizado el texto, se utiliza el modelo de espacio vectorial (MEV) (Salton, 1989) para representar los textos como vectores en un espacio multidimensional común. Las componentes de estos vectores (pesos) se obtendrán de la concatenación de los parámetros temáticos (TFIDF ó IWF) y estructurales (COF) descritos (para más detalles, ver figuras en la sección 4).

2.2. Entrenamiento

El proceso de entrenamiento parte de una colección de documentos $d_k \in \mathcal{D}^e$ agrupados en un conjunto $\mathcal{C}$ de categorías (dominios). En primer lugar, para disponer de un espacio común de representación y comparación para todos los textos, se genera una RRA *global* ($\mathbb{R}^N$) que contempla todas las palabras de cada uno de los dominios, así como sus relaciones (ver Figura 2(a)). Seguidamente, se modelan los textos de cada uno de los dominios a partir de esta red global, obteniendo una RRA *Full*[1] para cada dominio o RRA F D_n ($\mathbb{R}^N$) (ver Figura 2(b)). El proceso de entrenamiento finaliza cuando cada RRA F D_n se representa mediante un vector patrón ($\vec{p_n} \in \mathbb{R}^N$, $n = 1 \ldots |\mathcal{C}|$), obtenido por los $d_k \in |D_n^e|$ documentos de entrenamiento que le corresponden, representados según el MEV definido por la RRA global (ver Tabla 1).

[1] Sus componentes seguirán el orden dictado por el MEV completo o *Full space*, de ahí su nombre.

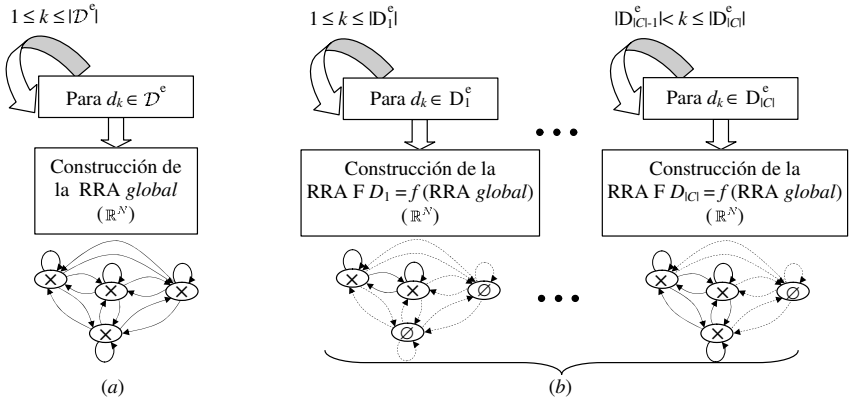


Figura 2: Proceso de generación de (a) la RRA *global* y (b) las RRA F D_n de dominio, desde D_1 hasta $D_{|\mathcal{C}|}$, referenciadas a la RRA *global* de dimensión $\mathbb{R}^N$ y construidas a partir de los documentos de entrenamiento de cada dominio $\mathcal{D}^e = \{D^e_1, \ldots, D^e_{|\mathcal{C}|}\}$. En los grafos, "×" indica nodo ocupado, "⊘" nodo vacío y las conexiones discontinuas denotan coocurrencias inexistentes.

2.3. Explotación

Una vez generadas la RRA F D_n y obtenidos sus vectores patrón $\vec{p}_n \in \mathbb{R}^N$, se procede a explotar el sistema de clasificación de textos. Para ello, primero se modela el texto a clasificar $t_k \notin \mathcal{D}^e$ según el MEV definido por la RRA global, obteniendo el vector $\vec{t}_k \in \mathbb{R}^N$ (ver Tabla 1). A continuación, se compara el vector $\vec{t}_k$ con cada uno de los vectores $\vec{p}_n$, utilizando la distancia del coseno. Finalmente, el texto de entrada se asigna al dominio respecto al que presente una menor distancia.

3. Red Relacional Asociativa Reducida

Uno de los factores críticos del modelo RRA F es la complejidad computacional del mismo, ya que para clasificar cualquier texto es necesario recorrer antes toda la RRA global para representarlo de forma coherente con los datos de entrenamiento. Además, por el hecho de no eliminar palabras (no se aplica lista de parada), no reducir la flexión de las mismas (no se extrae el radical), e incluir sus coocurrencias, el espacio multidimensional definido por la RRA global tendrá un tamaño considerable en comparación con el texto a clasificar, lo que puede reducir la separabilidad entre los distintos dominios.

Con el objetivo de minimizar el coste computacional y mejorar la eficiencia de clasificación de la CT basada en RRA F, cuestiones clave en el contexto de la CTH-MD, en este trabajo se propone una nueva estrategia de clasificación basada en una Red Relacional Asociativa Reducida (RRA R).

3.1. Definición

La idea fundamental de la CT basada en RRA R consiste en sustituir el espacio de comparación de los textos definido por la RRA global por un espacio vectorial definido a partir del texto de entrada t_k. Por lo tanto, la RRA generada a partir de t_k (en adelante, RRA R) es la que marcará el orden de comparación de los datos. En este caso, su contenido estará representado a partir del vector $\vec{t}'_k$ dentro del espacio vectorial $\mathbb{R}^{L^k}$ definido por la RRA R (siendo L^k el número de parámetros considerado -palabras y coocurrencias- del texto t_k, con $L^k \ll N$, típicamente). Así pues, durante el proceso de clasificación, los dominios D_n deberán representarse según la RRA R ($\mathbb{R}^{L^k}$) (ver Tabla 2). Gracias a este enfoque, la complejidad computacional que implica representar t_k en el espacio definido por red global se sustituye por el coste de representar cada dominio D_n en el espacio RRA R, coste que será, generalmente, mucho menor. No obstante, la estrategia RRA R no es más que una aproximación de la RRA F, cuya justificación teórica se pasa a describir a continuación.

3.2. Justificación

La reducción de dimensionalidad del MEV utilizado para representar los textos que implica la estrategia RRA R respecto a la RRA

Tabla 1: Ejemplo ilustrativo de la representación de los vectores patrón de dominio $\vec{p}_n$ (RRA F) y del texto a clasificar t_k según la RRA global, dados tres dominios D_1, D_2 y D_3 distintos. Los símbolos $\{\omega_A^n, \omega_B^n, \ldots, \omega_Z^n\}$ representan los pesos correspondientes a los términos de los textos.

Datos	**Contenido**	**Representación global**
MEV	$\{A,B,C,D,E,F,G,H,I,J\}$	$(\omega_A,\omega_B,\omega_C,\omega_D,\omega_E,\omega_F,\omega_G,\omega_H,\omega_I,\omega_J)$
D_1	$\{A,B,C,D,E,F,G\}$	$\vec{p}_1=(\omega_A^1,\omega_B^1,\omega_C^1,\omega_D^1,\omega_E^1,\omega_F^1,\omega_G^1,0,0,0)$
D_2	$\{A,B,C,E,H,I,J\}$	$\vec{p}_2=(\omega_A^2,\omega_B^2,\omega_C^2,0,\omega_E^2,0,0,\omega_H^2,\omega_I^2,\omega_J^2)$
D_3	$\{A,B,H,D\}$	$\vec{p}_3=(\omega_A^3,\omega_B^3,0,\omega_D^3,0,0,0,\omega_H^3,0,0)$
t_k	$\{C,A,Z\}$	$\vec{t}_k=(\omega_A^k,0,\omega_C^k,0,0,0,0,0,0,0)$

F puede justificarse desde un punto de vista algebraico. Por un lado, la RRA F puede representarse mediante un espacio vectorial $\mathbb{R}^N$, espacio al que pertenecen todos los vectores del conjunto de documentos de entrenamiento $\mathcal{D}^e$. En él se definen tanto los vectores patrón $\vec{p}_n$ de cada dominio, como los vectores $\vec{t}_k$ de los textos a clasificar —con M^k componentes activas y $(N-M^k)$ nulas, donde $M^k \ll N$ (ver Tabla 1). Por otro lado, la RRA R también puede representarse algebraicamente en el espacio vectorial $\mathbb{R}^L$, con $L^k \ll N$ y $L^k \geq M^k$, debido a las palabras de t_k no representadas en la RRA global.

Por otro lado, dentro del espacio vectorial definido por la RRA global, también se puede definir un subespacio vectorial $V \subset \mathbb{R}^N$, a partir de una base $B = \{\vec{b_1}, \vec{b_2}, \ldots, \vec{b}_{M^k}\}$ de M^k vectores ortogonales definidos por las componentes activas del vector $\vec{t}_k$. Esta base estará formada por los vectores de la base canónica de $\mathbb{R}^N$ que ocupan las posiciones no nulas del vector $\vec{t}_k$ (ver Tabla 3). Utilizando esta base, se puede representar el vector patrón $\vec{p}_n$ sobre el subespacio V como la mejor aproximación ($\hat{\vec{p}}_n$), según el criterio de los mínimos cuadrados, mediante la ecuación (2). Cualquier otra proyección de los vectores $\vec{p}_n$ dentro del subespacio V, presentará un error de aproximación mayor.

$$\hat{\vec{p}}_n = \frac{<\vec{p_n}, \vec{b_1}>}{<\vec{b_1}, \vec{b_1}>}\vec{b_1} + \cdots + \frac{<\vec{p_n}, \vec{b}_{M^k}>}{<\vec{b}_{M^k}, \vec{b}_{M^k}>}\vec{b}_{M^k} \quad (2)$$

donde $\{\vec{b_1}, \vec{b_2}, \ldots, \vec{b}_{M^k}\}$ es una base ortogonal del subespacio V definido a partir de $\vec{t}_k$.

Si se compara la representación de los datos en el espacio vectorial $\mathbb{R}^{L^k}$ definido por la RRA R —que genera un MEV′— (ver Tabla 2) con la que se obtiene de su proyección en el subespacio vectorial $V \subset \mathbb{R}^N$ definido por la RRA global (ver Tabla 3), se demuestra que utilizar la estrategia RRA R equivale a haber aproximado, con un error cuadrático mínimo, los vectores patrón de los dominios sobre el subespacio vectorial V. Añadir que, además del cambio de orden de las componentes, cuestión que no afecta al cálculo de las distancias, sólo existe una pequeña diferencia de $(L^k - M^k)$ posiciones nulas de los vectores patrón debidas a las palabras presentes en el texto a clasificar que no han sido observadas durante la fase de entrenamiento. Sin embargo, estas posiciones nulas, por un lado, no afectan al resultado del producto escalar de los vectores $\vec{t}\,'_k$ y $\vec{p}\,'_n$, y por otro, afectarán por igual a todas las comparativas en el contexto de la distancia cosenoidal utilizada (a través de la norma del vector $\vec{t}\,'_k$).

Tabla 2: Representación de los datos de la Tabla 1 en el MEV′ definido sobre $\mathbb{R}^{L^k=3}$ definido por la RRA R del texto a clasificar t_k.

Datos	**Contenido**	**Repr. reducida**
MEV′	$\{C,A,Z\}$	$(\omega_C,\omega_A,\omega_Z)$
t_k	$\{C,A,Z\}$	$\vec{t}\,'_k=(\omega_C^k,\omega_A^k,\omega_Z^k)$
D_1	$\{A,B,C,D,E,F,G\}$	$\vec{p}\,'_1=(\omega_C^1,\omega_A^1,0)$
D_2	$\{A,B,C,E,H,I,J\}$	$\vec{p}\,'_2=(\omega_C^2,\omega_A^2,0)$
D_3	$\{A,B,H,D\}$	$\vec{p}\,'_3=(0,\omega_A^3,0)$

De todos modos, la aproximación RRA R implica perder parte de la información contenida en la representación global de los vectores patrón de las RRA F D_n, cuestión que

Tabla 3: Representación de los datos de la Tabla 1 en el subespacio vectorial V generado a partir de la base ortogonal $B = \{\vec{b_1}, \vec{b_2}\}$ definida por las $M^k = 2$ componentes activas de $\vec{t}_k$, representado según el MEV global.

Base del Subespacio Vectorial V		
$\vec{b_1}=(1,0,0,0,0,0,0,0,0,0)$		
$\vec{b_2}=(0,0,1,0,0,0,0,0,0,0)$		
Datos	**Contenido**	**Componentes V**
D_1	$\{A,B,C,D,E,F,G\}$	$\hat{\vec{p}}_1=(\omega_A^1,\omega_C^1)$
D_2	$\{A,B,C,E,H,I,J\}$	$\hat{\vec{p}}_2=(\omega_A^2,\omega_C^2)$
D_3	$\{A,B,H,D\}$	$\hat{\vec{p}}_3=(\omega_A^3,0)$
t_k	$\{C,A,Z\}$	$\hat{\vec{t}}_k=(\omega_A^k,\omega_C^k)$

afecta al cálculo de la distancia del coseno a través de la norma de los mismos. En los experimentos que se presentan seguidamente, se analiza el impacto de esta aproximación en términos de la eficiencia de la clasificación, así como de su coste computacional.

4. Experimentos

Los experimentos se han desarrollado sobre un corpus de textos publicitarios, dividido en tres dominios: educación (527 frases), tecnología (323 frases) y cosmética (517 frases) —con estilos de locución alegre, neutro y sensual, respectivamente—, del que se han eliminado las frases ambiguas (Alías et al., 2004). Los CT estudiados se entrenan con el 80 % de los datos de cada dominio, asegurando la robustez estadística de los resultados mediante un 10-*fold random-subsampling*.

Con el objetivo de evaluar la habilidad de las estrategias de clasificación de texto propuestas, las frases del corpus se agrupan aleatoriamente para generar pseudo-documentos (documentos, en adelante) susceptibles de ser clasificados. De este modo, se pueden comparar los métodos de CT considerados a medida que se reduce el número de frases por documento, pasando desde una situación más cercana al problema de la CT clásica (con muchas frases por documento), hasta el caso extremo de disponer sólo de una frase por documento, situación bastante habitual en CTH.

4.1. Método de referencia

En este primer experimento se analizan distintos métodos de CT sobre el barrido de frases por documento estudiado, con el objetivo de buscar un algoritmo de referencia que permita validar el funcionamiento de las propuestas de CT basadas en RRA.

4.1.1. *Support Vector Machines*

En el contexto de la clasificación temática de grandes colecciones de documentos es donde el algoritmo basado en *Support Vector Machines* (SVM) ha demostrado un funcionamiento óptimo (Joachims, 1998; Sebastiani, 2002). Sin embargo, como se indica en (Sassano, 2003), a medida que el volumen de datos de entrenamiento se reduce respecto al tamaño del MEV empleado para representarlos, el método SVM pierde eficiencia, llegando a dejar de funcionar correctamente en el caso de trabajar con un MEV de dimensión mayor que el número de ejemplos. Este comportamiento de SVM se debe a la relación existente entre el número de datos de entrenamiento y el tamaño del MEV a parametrizar, que para *kernels* lineales, necesita disponer de $O(N)$ ejemplos para modelar correctamente un espacio $\mathbb{R}^N$ (Shawe-Taylor y Cristianini, 2004).

Debido a la estrategia de clasificación utilizada, donde no se trata con un CT únicamente temático y se entrena al CT sólo con los textos del corpus de voz, la diferencia entre el tamaño del espacio vectorial generado por los datos de entrenamiento y el número de ejemplos se acentúa al no eliminar la variabilidad de los términos (extracción del lema) ni las palabras vacías del texto (lista de parada) —es más, esta relación empeora al considerar las coocurrencias de las palabras en el modelado de los textos. Por ello, SVM no presenta resultados satisfactorios para este problema, como se ha indicado en trabajos previos (Alías et al., 2003; Alías et al., 2004).

4.1.2. ICA, bigramas o NN

A la vista de los pobres resultados ofrecidos por SVM, se estudiaron otras alternativas para obtener un CT de referencia. Entre las técnicas existentes se escogieron tres estrategias con enfoques de clasificación distintos. En primer lugar, se analiza el funcionamiento de un CT basado en el análisis en componentes independientes (ICA) de los textos (trabajando con tantas componentes como dominios) —aplicado satisfactoriamente a la búsqueda semisupervisada de dominios y a la jerarquización del corpus en (Alías et al., 2004; Sevillano, Alías, y Socoró, 2004).

En segundo lugar, se implementa un CT probabilístico a nivel de carácter basado en *n-gramas*, concretamente *bigramas*. En este caso, cada una de los dominios (categorías) se representa mediante un modelo de lenguaje probabilístico, obtenido a partir de la distri-

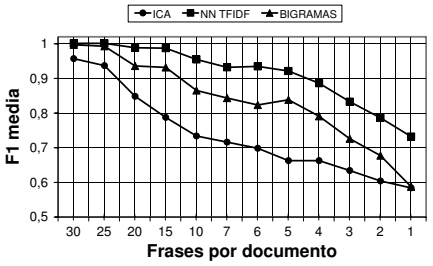
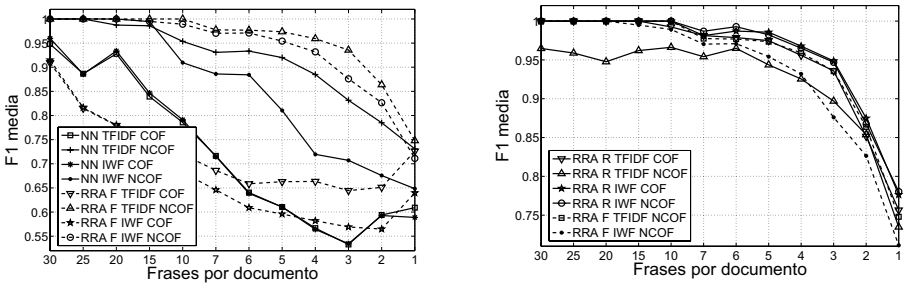


Figura 3: Eficiencia de clasificación de los métodos de referencia a lo largo del barrido de frases/documento estudiado.



(a) RRA F *vs.* NN.

(b) RRA F NCOF *vs.* RRA R.

Figura 4: Eficiencia de clasificación de los métodos de CT estudiados a lo largo del barrido de frases/documento considerado, para distintas parametrizaciones del texto.

bución de las parejas de caracteres presentes en el texto de ese dominio (Cavnar y Trenkle, 1994). En este trabajo, se ha escogido trabajar con bigramas por su buen compromiso entre la información de contexto considerada y el volumen de datos disponible para modelar de forma robusta cada dominio (a mayor n, se obtiene mayor información contextual, pero por el contrario, a menor n, se logra una mayor robustez estadística para el mismo número de datos).

Finalmente, se estudia el funcionamiento del clasificador basado en *Nearest Neighbour* (NN), trabajando con la ponderación TF·IDF, como método básico de CT sobre un MEV. Según esta estrategia, una vez representados todos los documentos en un espacio vectorial común, el texto de entrada se asigna a la categoría asociada al documento más cercano, según la distancia del coseno.

En la Figura 3 se presenta la eficiencia de clasificación (medida F_1 media (Sebastiani, 2002)) de los tres métodos comparados a lo largo del barrido de frases por documento estudiado. Se puede observar cómo NN es el método que presenta un mejor comportamiento global, seguido del clasificador basado en bigramas y, finalmente, se encuentra el CT basado en ICA, que sufre rápidamente la reducción del tamaño de los documentos. Por ello, se escoge NN como método de referencia para validar el funcionamiento de la CT basada en RRA en los siguientes experimentos.

4.2. Análisis de la propuesta

En este experimento se compara la eficiencia de clasificación de los métodos de CT basados en RRA respecto a NN, para cuatro parametrizaciones del texto distintas. Por un lado, se diferencia la representación del texto que incorpora parámetros estructurales, en este caso, las coocurrencias de las palabras (COF), de las que no los incluyen (NCOF). Por otro lado, se analizan las ponderaciones de los términos mediante los parámetros temáticos TFIDF e IWF (combinados dentro de los vectores que definen el contenido de los textos según lo indicado en la Figura 4).

4.2.1. Eficiencia de clasificación

La Figura 4 presenta los resultados de F_1 obtenidos del barrido de frases por documento realizado para todos los métodos y todas las parametrizaciones del texto estudiadas, utilizando la distancia del coseno como medida de similitud. Por un lado, se puede observar como los métodos basados en RRA presentan mejores resultados que NN, y por otro, como los métodos de representación global (RRA F y NN) reducen significativamente su eficiencia de clasificación al incorporar COF en la parametrización del texto, con un funcionamiento óptimo para TFIDF NCOF. Sin embargo, para RRA R la inclusión COF ayuda a mejorar los resultados. Además, RRA R obtiene un rendimiento óptimo para IWF —con resultados muy parejos entre COF y NCOF.

A partir de este análisis, se concluye que la CT basada en RRA R, aunque definida como aproximación de RRA F, presenta un comportamiento similar al obtenido por la CT basada en RRA F TFIDF NCOF, llegando a conseguir los mejores resultados de clasificación a nivel de 1 frase/documento.

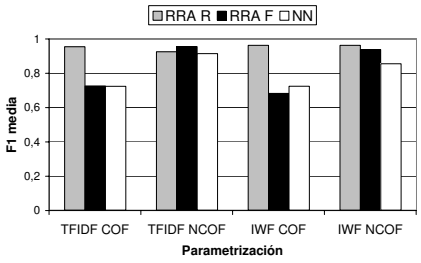


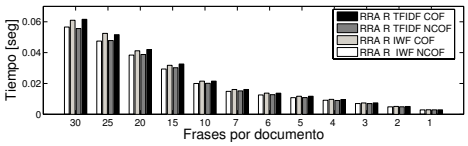
Figura 5: Eficiencia de clasificación media de los métodos dentro del barrido de frases/documento de la Figura 4, para distintas parametrizaciones del texto.

En la Figura 5 se observa que, a nivel global, RRA R presenta una respuesta más robusta respecto a las distintas parametrizaciones consideradas, en contraposición a RRA F y NN. Asimismo, la inclusión de COF provoca un impacto más negativo en RRA F que en NN por el hecho de trabajar con un único vector patrón por dominio, en lugar de usar un vector por documento. Sin embargo, RRA F consigue mejores resultados que NN con NCOF, tanto para IWF como TFIDF.

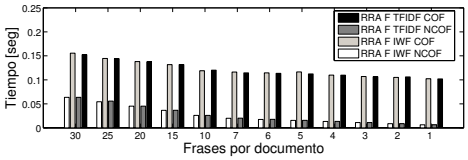
4.2.2. Coste computacional

En este experimento se estudia el coste computacional de la CT basada en RRA R *vs.* RRA F. Este coste se obtiene como el tiempo medio de clasificación sobre 10 ejecuciones de 1-*fold* del *random subsampling* utilizado en las pruebas. El estudio se ha realizado sobre un PC (PIV 1.79GHz 1GB RAM) con sistema operativo Linux y compilador gcc 3.3.5.

Del análisis de la Figura 6, se deduce que la CT basada en RRA R presenta un menor tiempo de clasificación que RRA F a lo largo del barrido de frases/documento. Para RRA F, el estudio muestra una clara diferencia de comportamiento cuando se comparan las configuraciones que utilizan las coocurrencias (COF) con las que no las utilizan (NCOF), debido al aumento de la dimensión del espacio vectorial cuando se trabaja con vectores que incluyen COF. En cambio, para RRA R los resultados son más homogéneos, independientemente de la parametrización utilizada. Concretamente, a medida que disminuye el tamaño de los textos a clasificar, el coste computacional también disminuye de forma lineal (en dos tramos lineales con cambio de pendiente alrededor de 10 frases/doc). Este comportamiento decreciente se debe, fundamentalmente, a que el espacio de comparación entre vectores lo define el texto a clasificar (de menor tamaño que el global), reduciéndose el tiempo de clasificación.



(a) Coste computacional de test de RRA F.



(b) Coste computacional de test de RRA R.

Figura 6: Coste computacional de clasificación dentro del barrido de frases/documento, para distintas parametrizaciones del texto.

Tomando la CT basada en RRA F como referencia, la RRA R presenta unas reducciones relativas del coste computacional que van desde un 12 % para 30 frases/doc, pasando por un 47 % para 4 frases/doc, hasta un 120 % para 1 frase/doc, utilizando sus configuraciones óptimas en términos de tiempo de ejecución (TFIDF NCOF para RRA F y IWF NCOF para RRA R, aunque las variaciones en RRA R son mínimas). Por lo tanto, parece claro que el método de CT basado en RRA R consigue mejorar los resultados respecto al método anterior, en lo que se refiere al coste computacional del proceso de clasificación, siendo éste más o menos *proporcional* al tamaño de los textos a clasificar.

5. Discusión

A lo largo de los experimentos descritos en este trabajo, se ha podido observar cómo todos los métodos de CT analizados se ven afectados, en mayor o menor medida, por la reducción del número de frases de los documentos a clasificar. Sin embargo, la estrategia de CT que presenta un comportamiento más robusto respecto al tamaño de los textos es la basada en RRA R, que consigue además los mejores resultados en términos de eficiencia de clasificación (seguida de cerca por la con-

figuración óptima de RRA F) cuando se trabaja con muy pocas frases por documento. El comportamiento particular de RRA R respecto a RRA F y NN está ligado al cambio de enfoque que esta estrategia propone para clasificar los textos. En este contexto, toma mayor importancia la mera presencia de la palabra (IWF) que su peso y singularidad a lo largo de la colección (TFIDF). Asimismo, la parametrización estructural del texto, considerada en este caso mediante la inclusión de las coocurrencias, permite mejorar las tasas de clasificación obtenidas en algunos casos, a diferencia del enfoque global (RRA F o NN), donde los resultados siempre empeoran —el aumento del tamaño de los vectores implica una menor separabilidad de los datos, por lo que resulta más fácil cometer errores de clasificación.

6. Conclusiones

En este trabajo se ha presentado un nuevo paso para el desarrollo de un sistema de clasificación de textos (CT) adaptado a las necesidades de la conversión de texto en habla multidominio (CTH-MD), es decir: textos cortos y bajo coste computacional. Para ello, se ha propuesto una evolución del método de CT basado en una Red Relacional Asociativa (RRA) de representación global de los textos (RRA F), a partir de un cambio en el enfoque del proceso de clasificación. El nuevo método de CT, basado en una RRA Reducida, ha demostrado, a lo largo de las pruebas realizadas, una buena eficiencia de clasificación, una mayor robustez frente a la reducción del tamaño del texto de entrada y un menor coste computacional respecto a RRA F —reduciendo la sobrecarga introducida por la CT sobre el proceso de CTH-MD. De todos modos, los moderados resultados conseguidos para pocas frases por documento (p.ej. $F_1 = 0.78$ para 1 frase/doc) indican que todavía existe un margen de mejora para continuar investigando en esta dirección.

Bibliografía

Alías, F., I. Iriondo, y P. Barnola. 2003. Multi-domain text classification for unit selection Text-to-Speech Synthesis. En *The 15th International Congress of Phonetic Sciences (ICPhS)*, páginas 2341–2344, Barcelona.

Alías, F., X. Sevillano, P. Barnola, L. Formiga, I. Iriondo, y Socoró. J. C. 2004. Conversión de Texto en Habla Multidomino. En *III Jornadas en Tecnología del Habla*, páginas 101–106, Valencia.

Alías, F., X. Sevillano, P. Barnola, y J.C. Socoró. 2003. Arquitectura para conversión texto-habla multidominio. *Procesamiento del Lenguaje Natural*, 31:83–90.

Cavnar, W.B. y J.M. Trenkle. 1994. N-Gram-Based Text Categorization. En *Proceedings of 3rd Annual Symposium on Document Analysis and Information Retrieval*, páginas 161–175, Las Vegas, USA.

Joachims, T. 1998. Text categorization with support vector machines: learning with many relevant features. En *Proceedings of ECML-98, 10th European Conference on Machine Learning*, numero 1398, páginas 137–142. Springer Verlag, Heidelberg, DE.

Rennison, E. 1994. Galaxy of News: An Approach to Visualizing and Understanding Expansive News Landscapes. En *ACM Symposium on User Interface Software and Technology*, páginas 3–12.

Salton, G. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley.

Sassano, M. 2003. Virtual Examples for Text Classification with Support Vector Machines. En *Proceedings of 2003 Conference on Empirical Methods in Natural Language Processing*, páginas 208–215, Japón.

Sebastiani, F. 2002. Machine learning in automated text categorisation. *ACM Computing Surveys*, 34(1):1–47.

Sebastiani, F. 2005. Text categorization. En *Text Mining and its Applications*. WIT Press, UK, capítulo 4, páginas 109–129.

Sevillano, X., F. Alías, y J.C. Socoró. 2004. ICA-Based Hierarchical Text Classification for Multi-domain Text-to-Speech Synthesis. En *Proceedings of ICASSP*, volumen 5, páginas 697–700, Montreal.

Shawe-Taylor, J. y N. Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Yi, J. y J. Glass. 1998. Natural-sounding speech synthesis using variable-length units. En *Proceedings of ICSLP*, páginas 1167–1170, Sydney, Australia.