Conversión de Texto en Habla Multidominio

Francesc Alías, Xavier Sevillano, Pere Barnola, Lluís Formiga, Ignasi Iriondo, Joan Claudi Socoró

Departamento de Comunicaciones y Teoría de la Señal
Enginyeria i Arquitectura La Salle. Universitat Ramon Llull
Pg. Bonanova 8. 08022 Barcelona
{falias,xavis,tm05122,llformiga,iriondo,jclaudi}@salleURL.edu

Resumen

Este trabajo presenta nuevas aportaciones relacionadas con la definición de la conversión de texto en habla (CTH) denominada síntesis multidomino. Esta propuesta intenta conseguir una calidad sintética próxima a la de los sistemas de CTH de dominio limitado con la versatilidad de la síntesis de propósito general. La arquitectura multidominio implica disponer de un corpus de voz estructurado, así como de un bloque de clasificación de textos adaptado al trabajo con pequeños corpus de textos. En esta comunicación, se analiza el comportamiento de dos métodos de clasificación: uno basado en Análisis en Componentes Independientes y otro basado en Redes Relacionales Asociativas, para documentos formados por muy pocas frases. Asimismo, se describe el corpus de voz multidominio que se ha grabado, junto a los tests subjetivos preliminares que justifican la viabilidad de la propuesta.

1. Introducción

En la actualidad, la técnica predominante en el ámbito de los sistemas de síntesis de voz concatenativa es la basada en *corpus* o *selección de unidades* [1, 2, 3]. Estos sistemas de conversión texto en habla (CTH) son capaces de generar voz sintética de una buena naturalidad e inteligibilidad. Sin embargo, no suele ser posible mantener esta calidad a lo largo de toda la síntesis [4]. Un primer paso hacia la mejora de estos sistemas ha consistido en aplicar el proceso de selección de unidades a dominios restringidos, logrando una síntesis de alta calidad dentro de los mismos [4].

Como contribución hacia la mejora de la calidad del habla sintética en sistemas CTH basados en selección de unidades, se ha presentado un nuevo sistema de CTH basado en un corpus multidominio [5]. Este enfoque pretende obtener un habla sintética de alta calidad propia de los sistemas de dominio limitado sin renunciar a la versatilidad de un sistema de propósito general. De este modo, el espacio de búsqueda de unidades se reduce, generando habla de gran calidad dentro del dominio objetivo. La arquitectura de conversión texto-habla multidominio requiere el uso de un corpus de voz estructurado y de un clasificador que categorice los textos a sintetizar en el dominio adecuado.

En esta comunicación se analiza la viabilidad de dos clasificadores de texto presentados en trabajos anteriores [5, 6] en el ámbito de los sistemas de CTH multidominio. Las características esenciales de este módulo de clasificación de textos son *i)* durante la fase de entrenamiento, debe ser capaz de estructurar el corpus de voz de forma jerárquica, lo que permitirá la coexistencia de varios dominios en un mismo corpus, y *ii)* durante la fase de explotación, debe ser capaz de clasificar textos cortos, situación muy habitual en CTH. Además, en este trabajo se

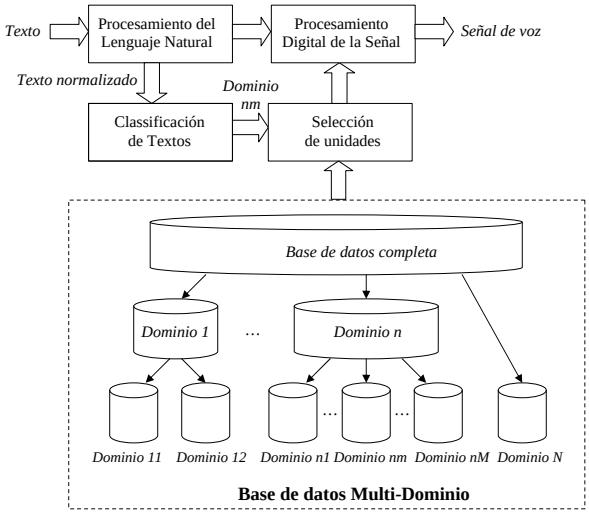

Figura 1: Diagrama de bloques del sistema de conversión texto-habla multidominio con un corpus jerárquico y el módulo de clasificación de textos.

presenta un ensayo preliminar de síntesis de voz multidominio, a fin de evaluar la viabilidad de la propuesta.

Este artículo está estructurado como sigue: en la sección 2 se describe la arquitectura y los requisitos necesarios para la implementación práctica de un sistema de conversión texto-habla multidominio. En la sección 3 se presentan dos métodos de clasificación de textos susceptibles de ser empleados en este tipo de sistemas. En la sección 4 se describe el corpus de voz multidominio que se ha grabado para poder evaluar el sistema. Por último, en la sección 5 se comparan ambos clasificadores de texto, evaluando su utilidad en el ámbito de conversión texto-habla multidominio y se presentan unas pruebas iniciales de este novedoso sistema de síntesis de voz.

2. Conversión texto-habla multidominio

La aplicación principal de la síntesis de voz por selección de unidades es la construcción de sistemas de conversión texto-habla de propósito general (CTH-PG), capaces de sintetizar *cualquier* texto de entrada [1, 2, 3]. A pesar de que la calidad del habla sintética generada por los sistemas CTH-PG suele ser muy buena, en ciertas situaciones ésta puede empeorar notablemente [4]. Por ello, y a fin de mejorar este aspecto, el proceso de selección de unidades ha sido aplicado a ámbitos restringidos, dando lugar a los sistemas de conversión texto-habla de dominio

limitado (CTH-DL). Estos sistemas son capaces de sintetizar habla de muy alta calidad, aunque restringida a esos dominios (véase la revisión presentada en [7]).

Por otra parte, el corpus de voz de los sistemas de propósito general suele estar diseñado para asegurar que la voz grabada no exhiba ningún estilo particular, es decir, que tenga un estilo de pronunciación *neutro* [8]. Dado que la CTH refleja claramente el estilo y la cobertura de la voz grabada [8, 9], la calidad del habla sintética decae cuando el dominio objetivo del texto de entrada no se ajusta al estilo del corpus de voz de propósito general diseñado [4, 10]. En otros trabajos [10, 11] se han presentado aproximaciones para mejorar la calidad de los sistemas CTH-PG, incorporando cierta adaptación al dominio en favor de la naturalidad del habla sintética e indicando mediante un lenguaje de marcas qué estilo se pretende sintetizar [12].

2.1. Definición y características de la propuesta

Teniendo en cuenta todas estas consideraciones, se diseñó un primer sistema de conversión texto-habla multidominio (CTH-MD) [5] (véase la figura 1) a fin de generar habla sintética de alta calidad (como en los sistemas CTH-DL) en varios dominios. Esta arquitectura permite que coexistan diferentes *dominios* en un mismo corpus de voz: diversas emociones (alegría, tristeza, etc.), varios estilos (periodístico, literario, etc.) e incluso diversas temáticas (política, deportes, ..., cuentos, poesía, etc.).

Así pues, un sistema CTH-MD pretende sintetizar directamente el texto de entrada con las unidades del dominio más adecuado, sin ningún tipo de marcas añadidas al mismo, simplemente a partir del texto normalizado por el bloque de procesamiento natural del lenguaje de cualquier sistema de CTH. Inicialmente, se trata de clasificar el texto de entrada (desde una frase a un párrafo) en uno de los dominios de más bajo nivel. Si la fiabilidad de la clasificación [13] es inferior a un umbral (no pertenece a ninguno de ellos), se pasa al nivel superior, y así sucesivamente. Si finalmente no existe ningún dominio apropiado, la señal de voz se genera a partir del dominio de propósito general ($Dominio_N$, en la figura 1).

Nótese que, aunque el vocabulario de cada dominio será especializado, deberá estar diseñado con una cobertura prosódica y fonética suficiente para permitir la síntesis de alta calidad en ese dominio [14]. Si esto no fuera así, caso de disponer de pequeños corpus particularizados por dominio, no tendría sentido realizar la preselección de dominio para la búsqueda de las unidades del corpus, sino que sería necesario considerar también las unidades de propósito general, ponderando la similitud de *estilos* entre las unidades objetivo y las candidatas [12]. Sin embargo, la buena cobertura del dominio permite, por un lado reducir el tamaño del espacio de búsqueda (reducción del coste computacional equivalente a una poda de alto nivel), y por otro, sintetizar con las unidades más adecuadas (potencialmente, reduce la distorsión de la señal sintética).

3. Clasificación de textos para CTH-MD

Como se ha comentado, la arquitectura del sistema CTH-MD requiere la incorporación de un módulo de clasificación de textos (CT). En la presente aproximación, el clasificador se entrena con los textos que conforman el corpus de voz multidominio, para extraer las características de cada dominio que permitan, a posteriori, la clasificación de los textos a sintetizar en el dominio más adecuado. En este contexto, la técnica de CT que se aplica en el sistema de CTH-MD deberá cumplir los siguientes requisitos:

1. Extraer las características de los dominios en base a un número reducido de textos de entrenamiento en comparación con las aplicaciones clásicas de clasificación de textos.
2. Estructurar los textos del corpus multidominio de forma jerárquica para los distintos dominios presentes en el corpus de voz.
3. Ser capaz de clasificar textos breves (en el límite, 1 frase), ya que serán los que deberá categorizar, mayoritariamente, en una aplicación real de CTH-MD.

En este trabajo revisamos dos alternativas presentadas en trabajos anteriores para llevar a cabo la tarea de CT en un sistema CTH-MD: una técnica basada en el Análisis en Componentes Independientes (*Independent Component Analysis* o ICA) [6, 13] y otra fundamentada en las Redes Relacionales Asociativas (RRA) [5].

ICA es una técnica estadística de propósito general fundamentada en un modelo generativo de variables latentes [15]. En el ámbito de la clasificación de textos, la aplicación de ICA se basa en la asunción de un modelo generativo de documentos como combinación de *temáticas semánticas*. Es decir, un documento se debe a la interacción de un conjunto de variables ocultas independientes que lo generan. Por tanto, la clasificación se rige por consideraciones puramente semánticas.

Por su parte, el clasificador basado en RRA modela los textos como un *grafo* de nodos interconectados (con tantos nodos como palabras aparecen en el texto), enlazados entre sí mediante conexiones ponderadas [16]. Una característica particular de este clasificador es que no sólo considera las palabras que aparecen en el texto (aproximación *bag-of-words* [17]), sino que considera las relaciones entre los términos que lo forman, modelando así la continuidad y el estilo de los mismos. Por lo tanto, la clasificación incluye el análisis estructural del texto.

4. Corpus para CTH-MD

Con el objetivo de desarrollar el trabajo presentado en esta comunicación, se grabó un corpus de voz para su uso en conversión texto-habla multidominio. La grabación fue fruto del proyecto MCYT PROFIT FIT-150500-2002-410 en el que también participaba el Departamento de Comunicación Audiovisual y Publicidad de la Universidad Autónoma de Barcelona.

El corpus, grabado por una locutora profesional, está formado por 2590 frases extraídas de una base de datos publicitaria, dividido en tres dominios temáticos: educación (EDU: 916 frases), tecnología (TEC: 833 frases) y cosmética (COS: 841 frases). Además, las frases de cada uno de las tres dominios temáticos se han grabado, respectivamente, con tres emociones distintas [18]: *i)* alegre, que corresponde al estereotipo extrovertido/alegre/fascinado, *ii)* estable, que corresponde al estereotipo estable/inteligente/sensitivo y maduro (coloquialmente conocido como *neutro*) y *iii)* sensual. En este caso, la locutora ha escogido el estereotipo más adecuado al contenido de cada dominio, permitiendo clasificar las emociones a través de las temáticas de los textos correspondientes.

Con el objetivo de evaluar la habilidad de los clasificadores de texto mencionados en la sección 3, se agrupan aleatoriamente las frases grabadas a fin de generar pseudo-documentos susceptibles de ser clasificados. El resultado de la generación de pseudo-documentos se presentan en la tabla 1. Por ejemplo, para el caso de 5 frases por documento, se obtienen 166 documentos de tecnología, con 3 frases sobrantes que son descartadas.

	Pseudo-documentos por dominio		
Número de frases	**EDU**	**TEC**	**COS**
1	916	833	841
2	458	416	420
3	305	277	280
5	183	166	168
6	152	138	140
7	130	119	120
10	91	83	84
15	61	55	56
20	45	41	42
25	36	33	33

Tabla 1: Número de pseudo-documentos por dominio temático en función del número de frases consideradas en cada documento.

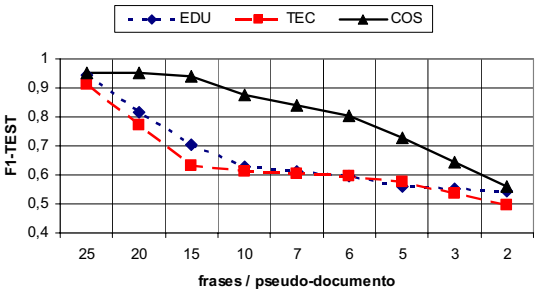

Figura 2: Medidas de clasificación $F1_M$-TEST del clasificador basado en ICA para diferentes longitudes de pseudo-documentos sobre el corpus original.

5. Experimentos

En esta sección se presentan un conjunto de experimentos desarrollados entorno al sistema CTH-MD. El primer experimento está orientado a evaluar los clasificadores de texto presentados en la sección 3. Y en segundo lugar, se presenta un experimento preliminar de síntesis de habla multidominio a fin de validar el sistema desde un punto de vista subjetivo.

5.1. Evaluación de los clasificadores de texto

En este apartado se analiza el comportamiento de los dos clasificadores de texto presentados en este trabajo: el basado en ICA y el basado en RRA en el contexto del entrenamiento y la clasificación de documento de tamaño muy reducido. Todos los experimentos de este apartado han sido realizados empleando la estrategia *10-fold cross-validation*, entrenando el clasificador con el 80 % de los pseudo-documentos del corpus.

5.1.1. Clasificador basado en ICA

En primer lugar evalúa el funcionamiento del clasificador basado en ICA. En la figura 2 se muestra la medida F1 de clasificación *macropromediada* en test ($F1_M$-TEST) [17] para cada una de las categorías del corpus. La asignación de cada documento a la categoría correspondiente se efectúa utilizando la medida relativa de relevancia presentada en [13].

Se puede apreciar que el rendimiento del clasificador ICA decrece alarmantemente cuando la longitud de los pseudo-documentos se reduce, especialmente en los dominios de tecnología y educación, siendo incapaz de clasificar a nivel de frase. Con el objetivo de determinar la causa de este pobre rendimiento, se inspeccionaron las frases del corpus, advirtiendo que gran cantidad de ellas tenían un contenido semántico poco definido, lo que dificulta en gran medida su clasificación temática con ICA. Se optó por eliminar estas frases temáticamente neutras, dando lugar al *corpus reducido* cuyas agrupación en pseudo-documentos se presenta en la tabla 2.

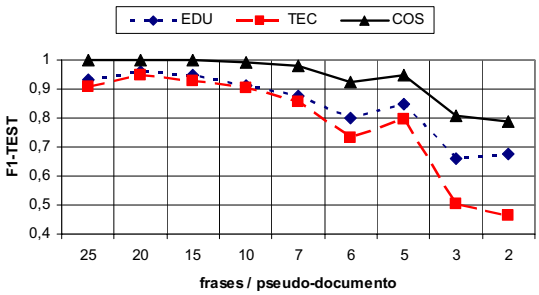

Figura 3: Medidas de clasificación $F1_M$-TEST del clasificador basado en ICA para diferentes longitudes de pseudo-documentos sobre el corpus reducido.

	Pseudo-documentos por dominio		
Número de frases	**EDU**	**TEC**	**COS**
1	527	323	517
2	263	161	258
3	175	107	172
5	105	64	103
6	87	53	86
7	75	46	73
10	52	32	51
15	35	21	34
20	26	16	25
25	21	12	20

Tabla 2: Número de pseudo-documentos por dominio temático en función del número de frases consideradas una vez eliminadas las frases temáticamente neutras.

Los resultados de la clasificación sobre este corpus reducido se muestran en la figura 3. Se puede apreciar como el clasificador ICA es, en este caso, capaz de clasificar pseudo-documentos de longitud menor, llegando a categorizar documentos de hasta 5 frases con una *precisión* y *cobertura* [17] de clasificación cercanas, en promedio, al 90 %. Por otro lado, comentar que se consigue una mejora media del 15 % para todo el experimento, sólo con un pequeño empeoramiento de resultados para el dominio de tecnología al llegar a las 2 y 3 frases por pseudo-documento. Esto es debido al menor número de frases que se han conservado de este dominio en el corpus reducido.

5.1.2. Clasificador basado en RRA

En segundo lugar se analiza el funcionamiento del clasificador basado en RRA bajo las mismas condiciones descritas para el basado en ICA. El grado de pertenencia de los textos a las distintas categorías se evalúa mediante la medida que pondera la distancia del coseno por la longitud del patrón, dado su buen comportamiento en experimentos previos [5, 16].

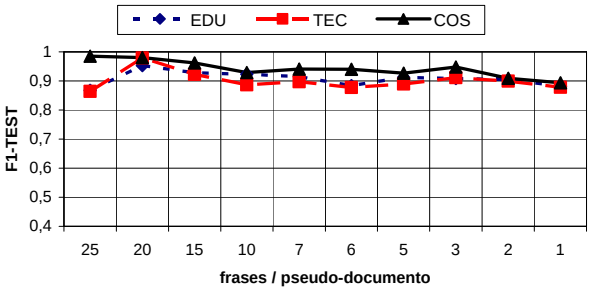

Figura 4: Medidas de clasificación $F1_M$-TEST del clasificador basado en RRA para diferentes longitudes de pseudo-documentos sobre el corpus original.

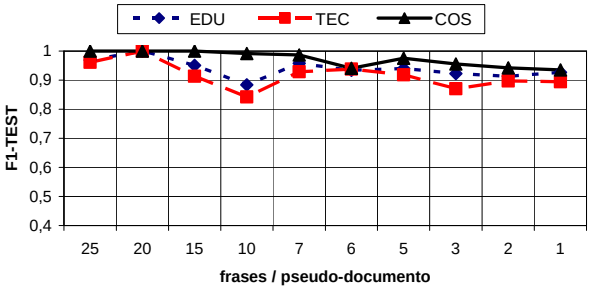

Figura 5: Medidas de clasificación $F1_M$-TEST del clasificador basado en RRA para diferentes longitudes de pseudo-documentos sobre el corpus reducido.

En la figura 4 se presentan los resultados de clasificación sobre el corpus original, empleando también la medida $F1_M$. Se puede apreciar que el clasificador basado en RRA obtiene muy buenos resultados incluso para pseudo-documentos cortos (respuesta muy estable), llegando a clasificar a nivel de una frase con una precisión y cobertura promedio superiores al 90 %. Esto es debido a que la parametrización empleada tiene en cuenta la estructura formal del texto, y en el caso de este corpus publicitario, se repiten numerosos patrones de frase.

Al aplicar el clasificador RRA sobre el corpus reducido (véase figura 5) los resultados, en general, se mantienen, con variaciones medias de un 3 %. Comentar que en los dominios de educación y tecnología, para 10 frases por pseudo-documento, los resultados incluso empeoran (del orden de un 3 %).

5.1.3. Conclusiones del experimento

Así pues, después de analizar el experimento en su conjunto se puede concluir que, del mismo modo que el clasificador basado en RRA es capaz de obtener buenas tasas de clasificación con muy pocas frases por documento, la reducción del número de frases, aunque semánticamente poco significativas, reduce sus prestaciones. Este concepto también queda reflejado al comparar los resultados de los clasificadores entre sí. Al trabajar con el corpus original el algoritmo basado en RRA mejora los resultados del basado en ICA en un 40 %, de promedio. En cambio, al pasar al corpus reducido la mejora media se reduce considerablemente (un 7 %), llegando a ser inferior en algunos casos con un número importante de frases por pseudo-documento.

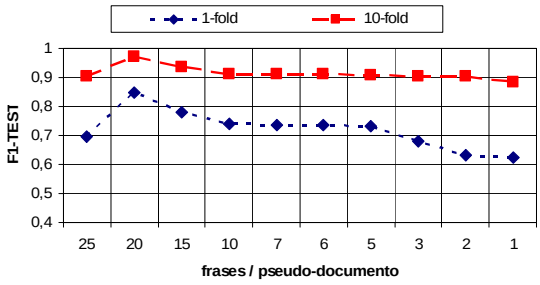

Figura 6: Medidas promedio entre los tres dominios de clasificación $F1_M$-TEST del clasificador basado en RRA para diferentes longitudes de pseudo-documentos con y sin *10-fold cross-validation*.

5.1.4. K-fold cross validation

Finalmente, se presenta un pequeño experimento en el que se muestra la necesidad de aplicar la estrategia de *10-fold cross validation* para un análisis robusto de los resultados para cada categoría. En la figura 6 se muestran los resultados promedio (para los tres dominios) obtenidos por el sistema basado en RRA, obteniendo una mejora promedio del 20 % efectuando *cross validation*. El hecho de disponer de documentos formados por pocas frases provoca que las pruebas sean altamente dependientes del contenido del subconjunto de test. Por este motivo, la filosofía de aumentar el conjunto de pruebas mediante un barrido aleatorio de los datos de test permite obtener unos resultados más robustos y significativos (después de promediarlos entre sí). Añadir que, con esta mejora introducida en el análisis del método basado en RRA, se mejoran los resultados presentados en anteriores trabajos [5, 16].

5.2. Síntesis de habla multidominio

Una vez analizado el funcionamiento de los dos métodos de clasificación, se procede a realizar un test preliminar para evaluar la calidad de la conversión texto en habla basada en una arquitectura multidominio. En este experimento se estudia el comportamiento del método de CT basado en RRA para en el contexto más crítico: la clasificación de textos a nivel de frase. Asimismo, también se han realizado distintas pruebas con el clasificador basado en ICA, pero en este caso para síntesis de pequeños párrafos, ya que en este método se necesita que los pseudo-documentos contengan un mayor número de frases ($\geq$5, ver figura 5) para obtener una buena tasa de clasificación.

5.2.1. Definición del experimento

Las frases utilizadas en el experimento de síntesis son un subconjunto del conjunto de frases de test que componen una de las diez pruebas del barrido por *10-fold crossvalidation*. La elección de esta prueba *representativa* se ha basado en el cálculo de la distancia cuadrática entre la precisión y la cobertura (tanto de entrenamiento como de test) de cada dominio respecto al promedio de esa magnitud sobre todas las pruebas.

A continuación se suman todas las distancias cuadráticas y se escoge aquella configuración que presenta una distancia menor respecto a la media. De este modo, se selecciona una prueba representativa, evitando los problemas que se derivan de la elección aleatoria del conjunto de test. En este caso la prueba escogida presenta unos valores de $F1_M$-TEST = $\{0{,}9339; 0{,}9; 0{,}9494\}$ para 106 frases de educación (alegría),

65 de tecnología (neutro) y 104 de cosmética (sensual), respectivamente.

El proceso seguido para la síntesis es el siguiente: se toma la prosodia real de la frase de test a sintetizar junto a su trascripción fonética y se fija como objetivo del proceso de síntesis. A continuación se aplica el algoritmo de selección, que para este experimento ha sido simplificado al trabajar con una función de coste binaria (búsqueda de la cadena de unidades más larga posible) para la selección de las unidades de síntesis. Finalmente, estas unidades son modificadas mediante PSOLA para obtener la prosodia deseada. Este proceso se realiza para cada una de las frases, generando su versión sintética mediante el corpus del dominio que le corresponde según el CT, así como a partir de las unidades *neutras* (dominio tecnológico). Finalmente se presenta al usuario las parejas de resultados obtenidos para que elija aquella que mejor le transmite el estilo deseado.

5.2.2. Análisis de los resultados

Como resultado objetivo de la prueba se ha constatado cualitativamente que, tal y como se analizó más exhaustivamente en un trabajo previo donde sólo se disponía de un corpus de texto [5], la longitud media de la cadena de unidades (*Average Segment Length*) que conforman las frases sintetizadas con el dominio apropiado supera claramente a las sintetizadas mediante las unidades de estilo neutro. De este modo, el número medio de concatenaciones se ve reducido de forma considerable, mejorando la calidad de la síntesis.

Por otro lado, a nivel subjetivo, se percibe con claridad que las señales sintéticas presentan el estilo del corpus desde el que han sido generadas, lo que permite aprovechar las características propias de la voz, algo que es especialmente notorio en el estilo sensual. Por lo tanto se demuestra que la correspondencia del estilo de las unidades de síntesis con el estilo esperado de la frase a sintetizar es fundamental y ratifica la viabilidad de la propuesta, aunque de manera informal. De todos modos, el experimento ha permitido constatar la necesidad de optimizar algunos aspectos tanto del etiquetado del corpus como del proceso de síntesis para obtener unos resultados de calidad suficiente que permitan el desarrollo de un test subjetivo más exhaustivo (y no sólo evaluado por expertos, que son capaces de obviar pequeños problemas de síntesis para sólo fijarse en los aspectos que se necesitan analizar).

6. Discusión

Existen distintas posibilidades a la hora de entrenar el clasificador de textos. En este caso se ha decidido entrenar el CT sólo con los textos que conforman el corpus de voz. Otra posible aproximación consistiría en el entrenamiento del CT con grandes cantidades de documentos correspondientes a los dominios de interés, escogiendo a posteriori el subconjunto de textos a grabar. A pesar de ello, se ha decidido trabajar con la primera opción ya que, por un lado, no es sencillo recopilar grandes cantidades de documentos adecuados a cada dominio, y por otro, aun con pocos documentos se han logrado resultados satisfactorios.

En este contexto es necesario argumentar la necesidad de adaptar el CT al entorno de la CTH-MD, debido a que se dispone de un corpus con un número de documentos y frases por texto muy reducidos. En cambio, la aplicación clásica de la CT trabaja con colecciones mayores de documentos y de frases por documento, como la colección *Reuters-21758* o la *OHSUMED* [17]. Es en este contexto donde el algoritmo basado en *Support Vector Machines* (SVM) ha demostrado su buen funcionamiento para este tipo de tarea [17].

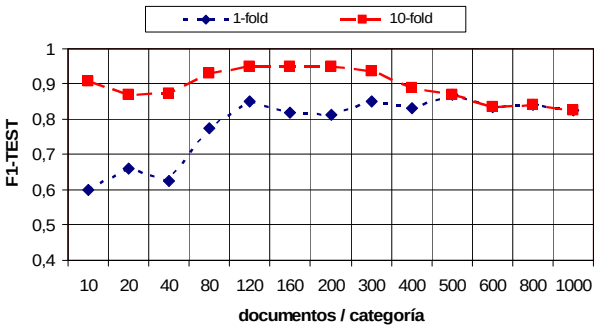

Figura 7: Medidas promedio de clasificación $F1_M$-TEST del CT basado en RRA para distintos tamaños de 5 de las categorías de *Reuters-21758* con y sin *10-fold crossvalidation*.

Es por ello que las primeras pruebas que se realizaron para el diseño de nuestro CT se basaron en el algoritmo SVM^{light} de Joachims [19]. Sin embargo, la aplicación de este método al ámbito de la CTH-MD fue descartado al no presentar unos resultados satisfactorios, debido, fundamentalmente, al insuficiente volumen de datos del que se dispone. En el trabajo de presentado en [20] se presentan bajas tasas de clasificación para un número reducido de documentos por dominio (nótese que este número debe multiplicarse por dos al considerarse siempre el mismo número de ejemplos positivos y negativos para el entrenamiento de SVM).

Una vez diseñado el método de CT basado en RRA, éste ha sido probado sobre 5 de las categorías más pobladas de la colección *Reuters-21758*: *acq*, *earn*, *grain*, *crude* y *trade*; las cuales disponen de 2210, 3776, 574, 566 y 513 documentos, respectivamente, con un número medio de 500 palabras por documento, aproximadamente. Del mismo modo que en el caso del corpus de voz presentado, cuando el número de documentos por categoría es muy reducido resulta esencial aplicar la estrategia de *10-fold crossvalidation* para obtener unos resultados significativos. En la figura 7 se puede comprobar que hasta 500 documentos por categoría, la aplicación del promediado de resultados de los 10 distintos conjuntos de test (escogidos aleatoriamente) permite obtener unos resultados bastante satisfactorios. Por lo tanto, el método basado en RRA presenta unas mejores prestaciones que el algoritmo de SVM, tal y como se demuestra en [20], para estas condiciones de trabajo.

7. Conclusiones

Este trabajo ha permitido dar un paso más hacia la consecución de un sistema de conversión de texto en habla multidominio. Su objetivo es la obtención de una calidad de síntesis elevada, como en los sistemas de síntesis de dominio restringido, permitiendo la versatilidad de la síntesis de propósito general. En esta comunicación se han revisado las dos propuestas presentadas en trabajos anteriores para el módulo de clasificación de textos, esencial para este tipo de arquitectura: *i)* la basada en Redes Relacionales Asociativas, que al incorporar análisis estructural del texto, permite clasificaciones a nivel de frase con resultados muy satisfactorios; y *ii)* la basada en el Análisis en Componentes Independientes, que permite la búsqueda de dominios de forma no supervisada más la jerarquización del contenido del corpus [6], así como una buena tasa de clasificación

cuando los documentos contienen un número no muy reducido de frases. Hasta el momento ambas estrategias han seguido caminos paralelos, pero a corto plazo se pretende buscar una solución mixta que aporte las ventajas de ambos métodos en uno.

Por otro lado se ha demostrado la necesidad del diseño de nuevas propuestas para la clasificación de textos (CT) distintas a las clásicas en este campo de investigación, debido a las particularidades del problema al que se aplican. El problema principal a superar radica en el hecho de no disponer de un gran volumen de datos para que el método de CT clasifique durante la fase de síntesis. También se ha demostrado que, por el mismo motivo, para evaluar el comportamiento del clasificador en este contexto es necesario realizar un barrido del conjunto de test, evitando mínimos y máximos locales.

Además, comentar que los primeros experimentos subjetivos realizados avalan la propuesta y permiten continuar trabajando en ella, para disponer de un sistema de síntesis multidominio de alta calidad, el cual deberá ser validado con un conjunto de experimentos subjetivos formales.

En otro orden de cosas, conviene indicar que, hasta el momento, la incorporación de un nuevo dominio al corpus requiere de su grabación y etiquetado correspondientes, cuestión que repercute negativamente tanto en el coste de desarrollo del CTH-MD como en la eficiencia del mismo. Una solución a este problema podría consistir en la definición, si es posible, del conjunto mínimo de dominios a grabar para un ámbito de aplicación determinado. A continuación se debería analizar el resultado de la síntesis del habla para un dominio distinto a los que forman el corpus, obtenida mediante la modificación prosódica de las unidades del dominio más cercano. De este modo, el punto de partida para la síntesis sería mucho más adecuado a la del dominio deseado en comparación con un CTP-PG modificado prosódicamente, cosa que repercutiría positivamente en la calidad de la señal sintética. Así pues, se pretende disponer de un sistema basado en selección de unidades que permita la generación de varios dominios con un compromiso entre el tamaño del corpus y la modificación de señal necesaria.

8. Agradecimientos

Agradecer la colaboración del Dr. Ángel Rodríguez y la Dra. Patricia Lázaro del Departamento de Comunicación Audiovisual y Publicidad de la Universidad Autónoma de Barcelona en las tareas de definición y grabación del corpus de voz publicitario. Este trabajo ha estado subvencionado en parte por el Ministerio de Ciencia y Tecnología mediante el Programa de Fomento de la Investigación Técnica con el proyecto MCYT PROFIT FIT-150500-2002-410.

9. Referencias

[1] A.W. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *EuroSpeech*, Rodes, Greece, 1997, pp. 601–604.

[2] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS system," in *Joint Meeting of ASA, EAA, and DAGA2*, Berlin, Germany, 1999, pp. 18–24.

[3] E. Eide, A. Aaron, R. Bakis, P. Cohen, R. Donovan, W. Hamza, T. Mathes, M. Picheny, M. Polkosky, M. Smith, and M. Viswanathan, "Recent Improvements to the IBM Trainable Speech Synthesis System," in *Proceedings of ICASSP*, Hong Kong, 2003.

[4] A.W. Black and K. Lenzo, "Limited Domain Synthesis," in *ICSLP*, Beijing, China, 2000.

[5] F. Alías, I. Iriondo, and P. Barnola, "Multi-domain text classification for unit selection Text-to-Speech Synthesis," in *The 15th International Congress of Phonetic Sciences (ICPhS)*, Barcelona, August, pp. 2341–2344.

[6] X. Sevillano, F. Alías, and J.C. Socoró, "ICA-based hierarchical text classification for multi-domain text-to-speech synthesis," in *IEEE International Conference on Speech, Acoustics and Signal Processing (ICASSP)*, Montréal, May 2004, vol. 5, pp. 697–700.

[7] B. Möbius, "Corpus-based speech synthesis: methods and challenges," *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS)*, vol. 6, no. 4, pp. 87–116, 2000.

[8] A. Breen and P. Jackson, "Non-uniform unit selection and the similarity metric within BT's LAUREATE TTS system," in *The 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, NSW, Australia, 1998.

[9] A.W. Black, "Perfect Synthesis for all of the people all of the time," in *IEEE TTS Workshop 2002 (Keynote)*, Santa Monica, USA, 2002.

[10] M. Chu, C. Li, P. Hu, and E. Cahng, " Domain adaption for TTS Systems ," in *ICASSP*, Orlando, USA, 2002.

[11] V. Fischer, J. Botella Ordinas, and S. Kunzmann, "Domain adaptation methods in the IBM trainable Text-To-Speech System ," in *ICSLP*, Jejú, Corea, 2004.

[12] W. Hamza, R. Bakis, E.M. Eide, M.A. Picheny, and Pitrelli J.F., "The IBM Expressive Speech Synthesis System," in *ICSLP*, Jejú, Corea, 2004.

[13] X. Sevillano, F. Alías, and J.C. Socoró, "Reliability in ICA-based text classification," in *Puntonet C.G and Prieto A. (Eds.), Lecture Notes in Computer Science: Proc. of the 5th International Conference on Independent Component Analysis and Blind Signal Separation*, Granada, September 2004, number 3195, pp. 1210–1217.

[14] J.M. Montero, R. Córdoba, J.A. Vallejo, J. Gutiérrez-Arriola, E. Enríquez, and J.M. Pardo, "Restricted-domain female-voice synthesis in Spanish: from database design to ANN prosodic modelling," in *ICSLP*, Beijing, China, 2000, pp. 621 – 624.

[15] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, 2001.

[16] F. Alías, X. Sevillano, P. Barnola, and Socoró J.C., "Arquitectura para conversión texto-habla multidominio," in *Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*, Alcalá de Henares (Spain), September, number 31, pp. 83–90.

[17] F. Sebastiani, "Machine learning in automated text categorisation," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[18] N. Montoya, "El uso de la voz en la publicidad audiovisual dirigida a los niños y su eficacia persuasiva," Tesis Doctoral, Universitat Autònoma de Barcelona, 1999.

[19] T. Joachims, "SVMlight v3.50," 2000, http://ais.gmd.de/~thorsten/svm_light/.

[20] M. Sassano, "Virtual Examples for Text Classification with Support Vector Machines," in *Proceedings of 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, Japón, 2003, pp. 208–215.