

EXPERIMENTOS EN INDEXACIÓN CONCEPTUAL PARA LA CATEGORIZACIÓN DE TEXTO

José María Gómez Hidalgo
Universidad Europea de Madrid
Villaviciosa de Odón, 28670 Madrid, España
jmgomez@uem.es

José Carlos Cortizo Pérez
AINet Solutions
Fuenlabrada, 28943, Madrid, Spain
jccp@ainetsolutions.com

Enrique Puertas Sanz, Manuel de Buenaga Rodríguez
Universidad Europea de Madrid
Villaviciosa de Odón, 28670 Madrid, España
{epuertas,buenaga}@uem.es

RESUMEN

En la Categorización de Texto (CT), una tarea de gran importancia para el acceso a la información en Internet y la World Wide Web, juega un papel fundamental el método de representación de documentos o indexación. La representación de los documentos en CT se basa generalmente en la utilización de raíces de palabras, excluyendo aquellas que aparecen en una lista de palabras frecuentes (modelo de lista de palabras). Este enfoque padece del problema habitual en Recuperación de Información (RI), la ambigüedad del lenguaje natural.

En este artículo exploramos el potencial de la indexación mediante conceptos, utilizando *synsets* de WordNet, frente al modelo tradicional basado en lista de palabras, en el marco de la CT. Hemos realizado una serie de experimentos en los cuáles evaluamos ambos modelos de indexación para la CT sobre la concordancia semántica Semcor. Los resultados permiten afirmar que la indexación mixta, usando lista de palabras y conceptos de WordNet, es significativamente más efectiva que ambos modelos por separado.

PALABRAS CLAVES

Categorización de Texto, Recuperación de Información, métodos de indexación, WordNet

1. INTRODUCCIÓN

La Categorización de Texto consiste en clasificación de documentos en categorías predefinidas (Sebastiani, 2002). Ejemplos de esta tarea son la clasificación de páginas Web en directorios temáticos como Yahoo! o el Open Directory Project (Mladenic, 1998), la clasificación de documentos en portales e intranets corporativas, la detección y filtrado de páginas Web pornográficas (Gómez et al., 2003) o mensajes de correo masivo no solicitado (*spam*) (Gómez, 2002), la clasificación de correo electrónico en carpetas personales, etc. La CT es por tanto una tarea de gran importancia en el acceso a la información en Internet y la World Wide Web.

Aunque es posible construir un sistema de CT creando conjuntos de reglas de forma manual, el enfoque más utilizado hoy en día consiste en usar técnicas de Recuperación de Información (RI) y Aprendizaje Automático (AA) para inducir un modelo de clasificación denominado *clasificador*. Dicho clasificador se obtiene: (1) representando (o *indexando*) un conjunto de documentos manualmente clasificados (la *colección de entrenamiento*) como vectores de pesos de términos – como en el Modelo del Espacio Vectorial de Salton (Salton, 1989), y (2) entrenando una función de clasificación que puede ser un conjunto de reglas, un árbol de

decisión, etc. Este enfoque es bastante efectivo para los sistemas de CT orientados a categorías temáticas, produciendo clasificadores precisos si se dispone de suficientes datos para el entrenamiento.

Una decisión importante en este modelo basado en aprendizaje es la definición de las unidades de representación o términos. Usualmente, los términos se definen como palabras filtradas mediante una lista de parada (*stoplist*) y desprovistas de sus afijos (extracción de raíces o *stemming*). Esta representación se suele llamar en la bibliografía modelo de lista de palabras (*bag of words*). Con frecuencia, esta representación permite un aprendizaje preciso, pues las palabras aisladas concentran una gran parte del significado del texto. Sin embargo, la polisemia (múltiples significados de una palabra) y la sinonimia (varias palabras con el mismo significado) limitan la efectividad de esta representación, como se ha comprobado frecuentemente en RI. En la bibliografía sobre CT se han propuesto una serie de definiciones alternativas de términos con éxito variable, incluyendo frases de palabras obtenidas por análisis estadístico o lingüístico (Caropreso et al., 2001; Lewis, 1992), patrones de Extracción de Información (Riloff, 1996) y conjuntos de sinónimos o *synsets* de WordNet (Fukumoto y Suzuki, 2001; Scott, 1998).

Esta última definición de término es especialmente interesante, de acuerdo con algunos resultados obtenidos para tareas de RI (Gonzalo et al., 1998). En este artículo nos centramos en el uso de *synsets* de WordNet como unidades de indexación, para afrontar los problemas de la sinonimia y polisemia que presentan las tareas de clasificación de texto. En la literatura sobre el tema, se han evaluado una serie de enfoques con resultados variables (Fukumoto y Suzuki, 2001; Junker y Abecker 1997; Liu y Chua, 2001; Petridis, 2001; Scott, 1998). Sin embargo, hasta la fecha, no hay un estudio en profundidad utilizando un amplio rango de representaciones del texto, métodos de selección de atributos, y algoritmos de aprendizaje.

En este artículo realizamos dicho estudio, presentando experimentos destinados a evaluar la hipótesis de que la indexación conceptual usando *synsets* de WordNet es una representación del texto más efectiva que la basada en lista de palabras. En nuestros experimentos, hemos evaluado una serie de representaciones de texto como entrada para un rango representativo de algoritmos de AA.

2. ENFOQUES DE REPRESENTACIÓN DE TEXTO EN LA CT

En CT, los documentos son generalmente representados como vectores de pesos de términos, como en el Modelo del Espacio Vectorial para RI (Salton, 1989). En este modelo, denominado *lista de palabras* en la literatura, los términos se definen como raíces de palabras, después de haber sido filtradas usando una lista de parada, y aplicando algoritmos de extracción de raíces como el de Porter. Los pesos pueden ser binarios (1 una raíz aparece en el documento, y 0 en otro caso), TF (*Term Frequency*, Frecuencia de Términos, el número de apariciones de una raíz en un documento), o TF.IDF (IDF es la *Inverse Document Frequency*, o Frecuencia Inversa de Documentos, usualmente definido como $\log_2(n/df(t))$, siendo n el número de documentos utilizados para el aprendizaje y $df(t)$ el número de documentos en los cuáles aparece el término t). Esta representación para el peso de un término (o en vocabulario de AA, valor de atributo) permite aplicar los algoritmos de aprendizaje, obteniendo un modelo de categorización denominado clasificador. Dependiendo del aprendizaje seleccionado, puede ser necesario seleccionar un subconjunto de los términos originales (*selección de atributos*) de acuerdo con alguna métrica de calidad, como la Ganancia de Información o χ^2 (véanse otras en (Sebastiani, 2002)).

La definición de término, o en otras palabras, de la unidad de indexación, es crítica en CT. Los términos han de ser buenos desde el punto de vista semántico (es decir, deben capturar lo máximo posible el significado de los textos), y desde el punto de vista de aprendizaje (esto es, debe permitir un aprendizaje eficiente y efectivo). Una serie de definiciones alternativas de término han sido estudiadas en la bibliografía, incluyendo: (1) n -gramas, que son secuencias de caracteres alfanuméricos (Cavnar94), apropiadas para documentos con errores de OCR; (2) frases estadísticas y lingüísticas (Caropreso et al., 2001; Lewis, 1992; Scott, 1998), que son expresiones multi-palabra, construidas tanto por métodos estadísticos, como por Procesamiento del Lenguaje Natural superficial; no se ha probado de forma definitiva que sean superiores a la indexación basada en raíces de palabras, aunque son prometedoras (Caropreso et al., 2001); y (3) patrones de Extracción de Información, usados en conjunción con el algoritmo Relevancy Signatures de Riloff y otros (Riloff, 1996); los patrones, llamados *signatures*, son pares del tipo (*palabra, nodo semántico*) en los cuáles las palabras actúan como desencadenantes de nodos semánticos, y se definen para el dominio de forma específica. Esta última aproximación garantiza CT de alta precisión aunque la cobertura puede verse

afectada. Se han estudiado otros enfoques, pero ninguno ha resultado claramente más efectivo que la representación como lista de palabras, sobre un rango de algoritmos de Aprendizaje Automático y una variedad de dominios de aplicación.

3. INDEXACIÓN CONCEPTUAL CON SYNSETS DE WORDNET

La popularidad del modelo de lista de palabras se justifica con el hecho que las palabras y sus raíces incluyen gran parte del significado del texto, estando especialmente indicadas para la clasificación basada en temas. Sin embargo, esta representación presenta dos problemas fundamentales: la sinonimia y polisemia de las palabras.

3.1 La Base de Datos Léxica WordNet

Los synsets de WordNet son candidatos idóneos a términos de indexación en las tareas de clasificación de textos, permitiendo abordar dichos problemas. WordNet es una Base de Datos Léxica que acumula información léxica sobre las palabras del idioma inglés (Miller, 1995). WordNet usa conjuntos de sinónimos o synsets como unidades básicas de información y organización. Un synset contiene una serie de sinónimos que definen un concepto, el cuál es uno de los posibles significados de las palabras en el synset. WordNet también almacena información sobre relaciones léxicas y conceptuales entre palabras, y conceptos, incluyendo hiponimia o vínculos ES-UN, meronimia o enlaces TIENE-UN, y otras. Esta clase de información en WordNet la convierte más en una red semántica y ontología que en un diccionario electrónico. En WordNet 1.7.1 se incluyen más de 146.000 palabras y expresiones, y 111.000 synsets para los nombres, verbos, adjetivos y adverbios del inglés.

3.2 Los synsets de WordNet como unidades de indexación para CT

La amplia cobertura de WordNet y su libre disponibilidad ha promovido su utilización para una gran variedad de tareas de clasificación de texto, incluyendo RI y CT¹. Mientras el uso de WordNet para clasificación de texto no ha demostrado ampliamente su efectividad (e.g. (Scott, 1998; Voorhees, 1998)), algunos trabajos en los cuáles se usan synsets de WordNet como términos de indexación para RI y CT son muy esperanzadores (Gonzalo et al., 1998; Fukumoto y Suzuki, 2001; Mihalcea y Moldovan, 2000; Petridis et al., 2001); véase (Stokoe et al. 2003) para una discusión en mayor profundidad.

La idea básica de la indexación conceptual con synsets de WordNet es reconocer los synsets a los cuáles hacen referencia las palabras en los textos, para luego usarlos como términos de la representación de los documentos en el Modelo del Espacio Vectorial. El uso de vectores de pesos de synsets para representar los documentos pueden mejorar la RI, tal y como comenta en (Gonzalo et al., 1998): “(...) Usar synsets de WordNet como espacio de indexación en lugar de palabras (...) combina dos beneficios para la recuperación: uno, que los términos se desambiguan totalmente (esto debería mejorar la precisión); y dos, que los términos equivalentes pueden ser identificados (esto debería mejorar la cobertura).”

Los experimentos centrados en indexación conceptual con synsets de WordNet para CT han tenido resultados variables. Por una parte, debido a la falta de desambiguación se ha perdido efectividad en algunos trabajos. Por otra parte, no está claro que la desambiguación completa sea absolutamente necesaria para obtener una representación de documentos más efectiva que el modelo de lista de palabras. Hay tres trabajos especialmente relevantes:

- Scott (Scott, 1998) ha evaluado una representación de texto en la cuál los synsets de WordNet correspondientes a las palabras, y sus hiperónimos, se usaron como unidades de indexación para el algoritmo de aprendizaje de reglas Ripper, sobre la colección de evaluación Reuters-21578. Los resultados de los experimentos fueron desalentadores, probablemente debido al hecho de que no se realizó desambiguación alguna, y a la incapacidad del Ripper para aprender de forma precisa en un espacio altamente dimensionado.

¹ Véase la amplia bibliografía en la web de WordNet (<http://www.cogsci.princeton.edu/~wn/>).

- Fukumoto y Suzuki (Fukumoto y Suzuki, 2001) han realizado experimentos extrayendo sinónimos e hiperónimos de los sustantivos de WordNet de una forma más sofisticada. Primero, los synsets no se usan como unidades de indexación; en su lugar, se extraen las palabras de los synsets relacionados con las palabras de los documentos. En segundo lugar, la altura hasta la cuál se busca en la jerarquía de WordNet es dependiente del campo semántico (localización, persona, actividad, etc.), y se optimiza durante el aprendizaje. Estos experimentos se realizaron con Support Vector Machines sobre la colección de textos de prueba Reuters-21578, y sus resultados fueron positivos, con especial incidencia en categorías con baja frecuencia de aparición. Es reseñable que no se realizó ningún tipo de desambiguación.
- Petridis y otros (Petridis et al., 2001) usaron synsets de WordNet como unidades de indexación con diversos algoritmos de aprendizaje sobre la colección de textos Semcor. En esta colección, todas las palabras y colocaciones han sido manualmente desambiguadas con respecto a los synsets de WordNet. Se evaluaron el enfoque de aprendizaje perezoso con el algoritmo de los K-vecinos más cercanos, el algoritmo probabilístico Bayes ingenuo, y unas Redes Neuronales sobre distintas representaciones del texto. La indexación conceptual obtuvo unos resultados considerablemente mejores que la representación con lista de palabras, siendo las Redes de Neuronas el mejor algoritmo de aprendizaje.

El trabajo de Scott sugiere que se requiere alguna clase de desambiguación. El trabajo de Fukumoto y Suzuki permite suponer que no se necesita una desambiguación total. Finalmente, el trabajo de Petridis y otros demuestra que la desambiguación perfecta es efectiva, sobre un limitado número de algoritmos de aprendizaje y una colección de textos correctamente desambiguados.

Sin embargo, hasta la fecha no se ha desarrollado ninguna serie de experimentos con el objetivo de evaluar la indexación conceptual sobre un rango representativo de algoritmos de aprendizaje y estrategias de selección de atributos. Nuestros experimentos se centran en probar el potencial de los synsets de WordNet como unidades de indexación para CT, considerando un rango representativo de estrategias de selección de atributos y algoritmos de aprendizaje. En concreto, nos concentramos en tres representaciones de los documentos: la basada en lista de palabras, la basada en conceptos, y una combinación de ambas.

4. EL DISEÑO DE LOS EXPERIMENTOS

En esta sección, describimos la configuración de nuestros experimentos, con especial atención a la colección usada como patrón de pruebas y las métricas de evaluación usadas en nuestro trabajo, y a las representaciones de texto usadas, y algoritmos de aprendizaje evaluados.

4.1 La colección de evaluación Semcor

La colección de textos Semcor es una Concordancia Semántica, un corpus etiquetado con conceptos de WordNet que se distribuye como suplemento de la propia WordNet (por ejemplo, para investigar o mostrar ejemplos del uso de conceptos). Sin embargo, Semcor ha sido adaptado y usado para evaluar tareas de RI en (Gonzalo et al., 1998), y para evaluar CT en (Petridis et al., 2001). No existe otra colección etiquetada con información conceptual en tanto detalle, y por ello, indexar con desambiguación “perfecta” es excesivamente costoso sin hacer uso de Semcor.

Semcor es un subconjunto del Brown Corpus y de la novela “The Red Badge of Courage”, con alrededor de 250.000 palabras. En Semcor, cada palabra o expresión ha sido etiquetada con su concepto adecuado en WordNet usando SGML. Es importante constatar que la información disponible en Semcor permite tanto la indexación por significados como por conceptos. Como indexación por significados, entendemos usar significados de palabras como unidades de indexación. Por ejemplo, podemos usar la tupla (coche, significado 1) o “coche\1” como unidad de indexación. La indexación por conceptos implica una normalización independiente de la palabra que permite el reconocimiento de “coche\1” y “automovil\1” como apariciones del mismo concepto, el nombre de código 02573998 en WordNet. De este modo, se aborda la sinonimia y polisemia de forma simultánea.

Semcor no precisa ser adaptado para evaluar la CT. Contiene 352 fragmentos de textos obtenidos de diversas fuentes, cubriendo 15 géneros, tales como PRENSA: REPORTAJE, RELIGIÓN, o FICCIÓN: CIENCIA. Hemos trabajado con los primeros 186 fragmentos de texto (completamente etiquetados), y los 15 géneros como categorías objetivo. Los géneros o clases objetivo no se solapan, estando cada documento en una sola

categoría. Asimismo, se trata de un problema de clasificación más cercano a la categorización basada en géneros que a la temática, aunque algunas diferencias de los géneros también se relacionan con los temas cubiertos por los textos. Finalmente, el número de documentos en cada categoría es variable, desde 2 (PRENSA: EDITORIAL) hasta 43 (EXTRANJERO).

Dado que las clases no se superponen, el problema de CT es un problema multiclase (dado un documento, seleccionar la categoría deseada entre las $m = 15$ disponibles). Sin embargo, algunos de los algoritmos de aprendizaje usados en este trabajo (por ejemplo, las Support Vector Machines) solo permiten trabajar con problemas de clasificación binarios (asignar un documento a una clase o no). El problema multiclase se puede transformar en $m = 15$ problemas binarios, en los cuáles se construye un clasificador para cada clase². Hemos preparado experimentos con problemas binarios para todos los algoritmos de aprendizaje, y con el problema multiclase sólo para aquellos que lo permiten, tal y como se describe más adelante.

4.2 Métricas de Evaluación

La efectividad de la CT se mide usualmente utilizando métricas de evolución de la RI, como la cobertura, precisión y F_1 (Sebastiani, 2002). Hemos usado F_1 , más acertada para CT que otras usadas en trabajos relacionados (como la “accuracy” o precisión en el sentido de AA en (Petridis et al., 2001)). Esta métrica es una media de la cobertura y la precisión, y puede ser promediada para las categorías por micro y macromedia. La macromedia da igual importancia a todas las categorías, mientras que la segunda da más importancia a las categorías más pobladas. En consecuencia, es importante calcular ambas medias complementarias.

Frecuentemente, las colecciones de evaluación de CT se dividen en dos partes: una para aprendizaje o entrenamiento, y otra para la prueba (el ejemplo que mejor muestra esto es la colección de entrenamiento Reuters-21568, con tres divisiones consolidadas). Cuando no hay una división de referencia, o los datos de entrenamiento son limitados, se usa un proceso de validación cruzada sobre k carpetas para estimar los valores de las métricas de evaluación. Resumiendo, los datos son divididos aleatoriamente en k grupos (preservando la distribución de documentos en clases), y se ejecutan k experimentos usando $k-1$ grupos para aprendizaje y 1 de evaluación. Los valores obtenidos en cada experimento se promedian sobre las k ejecuciones. Nuestras pruebas se han realizado con validación cruzada sobre $k = 10$ carpetas.

4.3 Enfoques para la representación de texto

Hemos evaluado tres enfoques distintos para representar el texto en CT: el modelo de lista de palabras, y el modelo de indexación conceptual, y un modelo mixto. Dado que el problema de CT sobre Semcor está parcialmente orientado al género, el modelo de lista de palabras utilizado en nuestros experimentos no hace uso de lista de parada ni de extracción de raíces. Ello se debe a que las palabras de una lista de parada (e.g. preposiciones, etc.) y las palabras en formato original (e.g. con sus correspondientes sufijos que incluyen las formas en pasado como “-ed”) pueden ser buenos indicadores de diferentes géneros de texto (Kessler et al., 1997). En el modelo de indexación conceptual hemos utilizado los conceptos o synsets de WordNet como vocabulario de representación, mientras que en el modelo mixto usamos tanto palabras como conceptos. Hemos representado los documentos de Semcor como vectores de pesos de términos, ya sean estos palabras o synsets, calculando los pesos por medio de la popular fórmula TF.IDF del Modelo de Espacio Vectorial³.

Debido al alto número de atributos o términos, que dificultan el aprendizaje en CT, se suele aplicar un proceso de selección de atributos para detectar aquellos más informativos. Un enfoque efectivo consiste en seleccionar las unidades de indexación que obtienen un valor mayor usando una métrica de calidad, como la Ganancia de Información (GI). En este trabajo, hemos evaluado los siguientes métodos de selección: sin selección (NOS), selección del 1% de atributos mejores de acuerdo a la GI (S01), selección del 10% atributos mejores según la GI (S10), y selección aquellas unidades de indexación con GI superior a 0 (S00). Estos porcentajes vienen justificados por otros trabajos (Yang y Pedersen, 1997).

² Esta clase de transformación se denomina “uno contra el resto” en la literatura (Petridis et al., 2001).

³ El proceso de indexación se ha realizado con el paquete experimental ir.jar de Mooney, disponible en <http://www.cs.utexas.edu/users/mooney/ir-course/>.

4.4 Algoritmos de Aprendizaje Automático

Es importante para nuestra hipótesis probar un rango representativo de algoritmos de aprendizaje rápidos. De aquellos probados en la literatura (Sebastiani, 2002), hemos seleccionado los siguientes⁴: el enfoque probabilístico Bayes ingenuo (Naive Bayes, NB); aprendizaje de árboles de decisión con C4.5 (C45); el método Support Vector Machines (SVM); y el algoritmo de meta-aprendizaje AdaBoost aplicado a Bayes ingenuo (ABNB). Para nuestros experimentos, hemos usado el paquete de aprendizaje WEKA⁵, con los parámetros por defecto para todos los algoritmos excepto para AdaBoost, donde usamos 10 iteraciones.

5. RESULTADOS Y ANÁLISIS

En la Tabla 1 presentamos un resumen de los resultados de nuestros experimentos. En esta tabla se presentan los resultados para los cuatro algoritmos (filas) y los tres modelos de representación (columnas) evaluados. Todos los resultados corresponden a la formulación binaria del problema de clasificación, dado que la formulación multiclase presenta resultados sensiblemente inferiores. Ello se debe a la dificultad que presenta para los algoritmos de aprendizaje separar 15 clases con un solo clasificador, resultando notablemente más fácil separar cada clase aisladamente de las demás. Asimismo, la selección de atributos en la formulación binaria es más eficaz, puesto que se ha de elegir atributos que separen adecuadamente una clase del resto, y no todas a un tiempo.

Tabla 1. Resumen de los resultados de nuestros experimentos. Se presentan los resultados para los cuatro algoritmos y los tres modelos de representación evaluados, para la formulación binaria del problema de clasificación.

| Algoritmo | Conceptos (synsets) | | Palabras | | Mixto | |
|-----------|---------------------|-------|----------|-------|-------|-------|
| | macro | micro | macro | micro | macro | micro |
| NB | 0,631 | 0,739 | 0,635 | 0,750 | 0,711 | 0,797 |
| C45 | 0,258 | 0,391 | 0,270 | 0,382 | 0,265 | 0,406 |
| SVM | 0,502 | 0,773 | 0,482 | 0,730 | 0,477 | 0,739 |
| ABNB | 0,638 | 0,759 | 0,638 | 0,760 | 0,706 | 0,800 |

En cada columna de la Tabla 1 se presentan los valores de F_1 obtenidos por macromedia (macro) y por micromedia (micro). Se muestran los mejores valores para cada algoritmo, según el método de selección más efectivo para cada representación. Para el algoritmo Bayes ingenuo, el método de selección más efectivo es S00 en todos los casos. Para el algoritmo C4.5, el método de selección más efectivo es S00 para conceptos, y S01 para las demás representaciones. En el caso del algoritmo Support Vector Machines, el método más efectivo de selección es S01 en todos los casos. Finalmente, en el caso de AdaBoost con Bayes ingenuo, el método de selección más efectivo es S00 para las representaciones basadas en conceptos y en palabras, y S10 para el modelo mixto.

Es importante señalar que la indexación mixta se revela más efectiva que los otros modelos de representación en todos los casos, excepto para las SVM. Dado que la indexación basada en palabras supera a la indexación basada en conceptos en dichos casos, se puede afirmar que los conceptos en exclusiva nos son la representación más efectiva, pero contribuyen decisivamente a mejorar la efectividad de la representación.

Los mejores resultados se obtienen con la representación mixta y los algoritmos Bayes ingenuo (NB) y AdaBoost sobre Bayes ingenuo (ABNB). Se aprecia que el algoritmo de meta-aprendizaje AdaBoost logra mejorar el valor de F_1 por micromedia, a costa de empeorar el valor por macromedia. Ello implica que AdaBoost aumenta la efectividad de Bayes ingenuo sobre las categorías más frecuentes, y la disminuye sobre las menos frecuentes.

⁴ Por brevedad, hemos omitido las referencias. Pueden encontrarse algunas en (Sebastiani, 2002).

⁵ Disponible en <http://www.cs.waikato.ac.nz/ml/weka/>.

6. CONCLUSIONES

A partir de los resultados presentados en las secciones anteriores, se pueden extraer dos conclusiones principales:

- En primer lugar, ninguno de los métodos de indexación (por palabra o por conceptos) es aisladamente superior al otro. Intuitivamente, y de acuerdo con nuestra discusión previa, el segundo ha de ofrecer mejores resultados, al abordar los problemas de polisemia y sinonimia del lenguaje natural. Sin embargo, los resultados experimentales no confirman esta hipótesis de manera firme.
- En segundo lugar, la combinación de representaciones se revela superior a la utilización de palabras o conceptos por separado. En este caso, los resultados experimentales confirman la intuición de que aportar más información al sistema de categorización (la extraída de la BDL WordNet) permite que éste aumente la efectividad.

Conviene señalar que los resultados son muy prometedores pero no concluyentes, debido a las limitaciones de la colección Semcor (incluyendo su tamaño y la orientación a género tanto como a tema). En próximos experimentos pretendemos concentrarnos en la colección Reuters-21578, cuyas dimensiones y orientación permiten obtener conclusiones más generales y precisas. Por otra parte, esta colección no está etiquetada con respecto a los significados de WordNet, lo que implica la utilización de algún algoritmo de desambiguación que limitará necesariamente las conclusiones sobre ella.

Finalmente, nos gustaría recalcar que en estos experimentos hemos usado exclusivamente los conceptos que aparecen explícitos en los documentos de la colección Semcor. Los experimentos reflejados en (Fukumoto y Suzuki, 2001) evidencian que un uso limitado de las relaciones semánticas disponibles en WordNet puede producir mejoras en la efectividad. En el futuro abordaremos este enfoque.

REFERENCIAS

- Caropreso, M.F., S. Matwin, y F. Sebastiani, 2001. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. *Text Databases and Document Management: Theory and Practice*. Idea Group Publishing, pp. 78–102.
- Cavnar, W.B. y J.M. Trenkle, 1994. N-gram based text categorization. *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161–175, Las Vegas, US.
- Fukumoto, F. y Y. Suzuki, 2001. Learning lexical representation for text categorization. *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*.
- Gómez, J.M., 2002. Evaluating Cost-Sensitive Unsolicited Bulk Email Categorization. *ACM Symposium on Applied Computing*, Madrid.
- Gómez, J.M., Puertas E., Carrero, F. y Buenaga, M. de., 2003. Categorización de texto sensible al coste para el filtrado de contenidos inapropiados en Internet. *Procesamiento del Lenguaje Natural*, No. 31, pp. 13-20.
- Gonzalo, J., F. Verdejo, I. Chugur, y J. Cigarrán, 1998. Indexing with WordNet synsets can improve text retrieval. *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*.
- Junker, M. y A. Abecker, 1997. Exploiting thesaurus knowledge in rule induction for text classification. *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing*, pp. 202–207.
- Kessler, B., G. Nunberg, y H. Schütze, 1997. Automatic detection of text genre. *Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics*, pp. 32–38, Madrid, ES.
- Kilgarriff, Adam y Joseph Rosenzweig, 2000. Framework and results for english SENSEVAL. *Computers and the Humanities*, Vol. 34, Nos. 1–3, pp. 15–48.
- Lewis, David D, 1992. Representation and learning in information retrieval. Ph.D. tesis, Department of Computer Science, University of Massachusetts, Amherst, US.
- Liu, J. y T.S. Chua, 2001. Building semantic perceptron net for topic spotting. *Proceedings of 37th Meeting of Association of Computational Linguistics*.
- Mihalcea, Rada y Dan I. Moldovan, 2000. Semantic indexing using WordNet senses. *Proceedings of ACL Workshop on IR and NLP*.
- Miller, George A., 1995. WordNet: A lexical database for English. *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41.

- Mladenic, D., 1998. Turning Yahoo into an Automatic Web-Page Classifier. *Proceedings of the 13th European Conference on Artificial Intelligence*, pp. 473-474.
- Petridis, V., V.G. Kaburlasos, P. Fragkou, y A. Kehagias, 2001. Text classification using the σ -FLNMAP neural network. *Proceedings of the 2001 International Joint Conference on Neural Networks*.
- Riloff, E., 1996. Using learned extraction patterns for text classification. *Connectionist, statistical, and symbolic approaches to learning for natural language processing*, pp. 275-289. Springer Verlag.
- Salton, G., 1989. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison Wesley.
- Scott, S., 1998. Feature engineering for a symbolic approach to text classification. Master's thesis, Computer Science Dept., University of Ottawa, Ottawa, CA.
- Sebastiani, F., 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, Vol. 34, No. 1, pp. 1-47.
- Stokoe, Christopher, Michael P. Oakes, y John Tait, 2003. Word sense disambiguation in information retrieval revisited. *Proceedings of the 26th ACM International Conference on Research and Development in Information Retrieval*.
- Voorhees, Ellen M., 1998. Using WordNet for text retrieval. *WordNet: An Electronic Lexical Database*, MIT Press.
- Yang, Y. y J.O. Pedersen, 1997. A comparative study on feature selection in text categorization. *Proceedings of the 14th International Conference on Machine Learning*.