Procesamiento del Lenguaje Natural, Revista nº 44, marzo de 2010, pp 11-18 recibido 08-01-10 revisado 12-02-10 aceptado 06-03-10

Un Análisis Comparativo de Estrategias para la Categorización Semántica de Textos Cortos*

A Comparative Analysis of Strategies for Semantic Short-Text Categorization

María V. Rosas, Marcelo L. Errecalde
LIDIC, UNSL, San Luis, Argentina
{mvrosas, merreca}@unsl.edu.ar

Paolo Rosso
NLE Lab. - ELiRF, UPV, España
prosso@dsic.upv.es

Resumen: La categorización de textos cortos es, hoy en día, un área importante de investigación debido a que gran parte de la información que recibimos y con la cual trabajamos habitualmente tiene esta característica (e-mails, mensajes de texto, resúmenes de noticias, entre otros). Distintos trabajos han reportado resultados interesantes en la categorización de textos incorporando información semántica a la representación de los documentos. Sin embargo, estos trabajos no se han concentrado en general en las particularidades que presentan los textos cortos. Por otra parte, los métodos de desambiguación más difundidos (basados en corpus) no siempre son adecuados en los dominios que se intentan abordar. En estos casos, la desambiguación basada en conocimiento se convierte en una alternativa interesante a considerar. En este trabajo, estudiamos la efectividad de la categorización de textos cortos, cuando se utiliza información semántica obtenida con métodos basados en conocimiento. Los resultados obtenidos con este enfoque muestran mejoras interesantes que incentivan a continuar con esta línea de investigación.
Palabras clave: categorización de textos, desambiguación, ontología, colecciones de textos cortos

Abstract: Nowadays, short-texts categorization is an important research area because most of the information we usually receive and work with have this characteristic (e-mails, text messages, news, etc.). Different studies have reported interesting results in text categorization by adding semantic information to documents' representation. However, these studies have not focused on the particularities that short texts introduce. Furthermore, the most popular disambiguation methods (corpus-based methods) not always are feasible to use in these domains. Thus, knowledge-based disambiguation methods become interesting alternatives in these cases. In this article, we study the effectiveness of short text categorization, when semantic information, obtained by knowledge-based methods, is used. The results obtained with this approach show interesting improvements that encourage to continue this line of research.
Keywords: text categorization, word sense disambiguation, ontology, short-text corpora

1. Introducción

El exceso de información disponible cada día, hace necesaria la tarea de procesar los datos de manera efectiva. En la sociedad de la información y la comunicación, el conjunto de textos en lenguaje natural con el que trabajamos tiende a aumentar pero con la particularidad de un uso acotado en el número de palabras en cada texto. En la actualidad, la comunicación escrita entre personas hace un uso constante de este tipo de textos restringidos en tamaño buscando optimizar el uso de palabras en interacciones eficientes, cortas y veloces, a través de correo electrónico, mensajes de textos, reportes internos, faxes, fragmentos de páginas web, cables de noticias, entre otros.

Debido a lo planteado previamente y al hecho de que generalmente se recibe más información de la que se desea o es posible procesar, las aplicaciones y técnicas vincu-

* El trabajo del segundo y tercer autor ha sido soportado por el proyecto TEXT-ENTERPRISE 2.0 (TIN2009-13391-C04-03).

ladas al procesamiento del lenguaje natural (PLN) juegan un papel relevante en nuestros días. Entre las distintas aplicaciones de PLN, la categorización automática de textos ha despertado un notable interés debido a la necesidad urgente de organizar, mantener y procesar toda información disponible a partir de un conocimiento más profundo del lenguaje (Sebastiani, 2005). En este sentido, existe aún un número limitado de estudios realizados sobre la categorización de textos cortos, por lo que el desarrollo de métodos efectivos para lograr mejoras en esta tarea, continúa siendo un tema abierto de investigación.

Diferentes trabajos han analizado las ventajas de incorporar información semántica a la representación de textos obteniéndose resultados positivos en algunos de los experimentos llevados a cabo (Banerjee, Ramanathan, y Gupta, 2007). Es conveniente destacar, que en general estos estudios están enfocados en documentos donde es factible, en la mayoría de los casos, disponer de una colección de entrenamiento para la tarea de desambiguación del sentido de las palabras (*WSD* las siglas en inglés para Word Sense Disambiguation). Este enfoque basado en corpus (también conocido como supervisado) no siempre es viable de ser aplicado en dominios con la característica planteada previamente.

Una alternativa para abordar el problema anterior, es el uso de métodos de WSD basados en conocimiento que obtienen información desde recursos léxicos externos. Si bien este tipo de métodos suelen mostrar resultados de menor calidad que los obtenidos con métodos basados en corpus, constituyen en muchos casos la única alternativa realista si se desea hacer uso de información semántica en la representación de documentos (Vázquez, 2009). Teniendo en cuenta esto, se puede pensar en el enfoque basado en conocimiento como una opción apropiada para el caso que nos ocupa: la categorización de textos cortos.

El objetivo de este trabajo es determinar en que medida la información obtenida mediante técnicas de WSD basadas en conocimiento pueden beneficiar el desempeño de distintos enfoques para la categorización de textos cortos. A tal fin, elestudio experimental incluirá algunos de los algoritmos que han mostrado ser los más efectivos en la categorización de textos general, y un conjunto representativo de colecciones de documentos cortos.

El resto del trabajo está organizado de la siguiente manera: la Sección 2 presenta conceptos introductorios relacionados a nuestro trabajo; la Sección 3 detalla el diseño experimental, describiendo los conjuntos de datos utilizados y la representación de los textos con las distintas variantes en la incorporación de información semántica. Los experimentos y los resultados de los mismos se describen en la Sección 4. Finalmente, la Sección 5 presenta las conclusiones y posibles trabajos futuros.

2. Conceptos introductorios

En la mayoría de las tareas de categorización, los documentos son representados mediante el modelo de espacio vector introducido por Salton (Salton y Buckley, 1988), para la codificación de textos. En este enfoque, cada texto es representado por un vector de n-términos, donde n es el número de términos que aparecen en la colección de documentos, y cada término del vector es ponderado con un peso determinado usualmente en base a la frecuencia de ocurrencia del término en el documento y en la colección completa. En el sistema SMART (Salton, 1971), cada codificación está compuesta por tres letras: las primeras dos letras refieren, respectivamente, a TF (frecuencia de un término) e IDF (frecuencia inversa del documento) mientras que el tercer componente ($NORM$) indica si se utiliza normalización o no. Teniendo en cuenta la nomenclatura estándar SMART, se consideran cinco alternativas diferentes para la componente TF: n (natural), b (binario), l (logaritmo), m (max-norm) and a (promedio-norm); dos alternativas para el componente IDF (n y t) con n (no aplicación) y t ($tfidf$) y dos alternativas para la normalización: n (no normalización) y c (coseno). De esta forma, una codificación *ntc* representa la codificación estándar tf-idf (normalizada).

El uso de información semántica implica, en este contexto, la incorporación del *significado* de los términos a la representación. La determinación de cuál es el significado que corresponde a los distintos términos no es una tarea directa debido a los problemas de polisemia y sinonimia. Por este motivo, se requieren de métodos de WSD que, así como se explicó previamente, pueden ser clasificados en términos generales como basados en

corpus o basados en conocimiento (Agirre y Edmonds, 2006). En este trabajo, nos centraremos en métodos basados en conocimiento los cuáles requieren de algún recurso externo que, en primera instancia, puede ser cualquier base de conocimiento léxica que defina los diferentes sentidos de las palabras y relaciones entre ellas (conocida como *ontología*). La ontología más utilizada es WordNet (WN) (Miller, 1995), una combinación de diccionario y tesauro que agrupa las palabras en conjuntos de sinónimos llamados *synsets*. Cada synset representa un "concepto" léxico único, que en WN puede estar relacionado semánticamente con otros conceptos a través de relaciones de sinonimia, hiperonimia, hiponimia, etc., dando origen de esta manera a una jerarquía conceptual.

En el presente trabajo serán evaluados tres enfoques diferentes basados en conocimiento:

1. *CIAOSENSO*: sistema basado en la idea de *densidad conceptual*, medida como la correlación entre el sentido de una palabra y su contexto. Para ello, utiliza la longitud del camino más corto que conecta dos synsets en la taxonomía de sustantivos que utiliza WordNet. El método utiliza las relaciones jerárquicas de hiperonimia e hiponimia presentes en WordNet (Rosso et al., 2003), (Buscaldi, Rosso, y Masulli, 2004).
2. *Algoritmo de Lesk*: el procedimiento determina los sentidos de las palabras que ocurren en un contexto particular basándose en una medida de solapamiento entre las definiciones de un diccionario y dicho contexto (Lesk, 1986). Una variante, denominada *Lesk Mejorado*, fue propuesta en (Banerjee y Pedersen, 2002) que considera no sólo las definiciones de las palabras a desambiguar, sino también las definiciones de aquellos términos relacionados semánticamente en la jerarquía WordNet.
3. *Método heurístico del sentido más frecuente*: sistema basado en propiedades lingüísticas aprendidas. Esta es la técnica más simple de desambiguación asignando a una palabra el sentido que ocurre más a menudo de todos los posibles sentidos de esa palabra. En este caso, los sentidos han sido obtenidos a partir de las frecuencias de ocurrencia de las palabras reportadas por WordNet.

El uso de información semántica plantea distintas alternativas respecto a cómo esta información puede ser incorporada en la representación de los documentos. En este trabajo, el enfoque tradicional basado en términos[1] será comparado con dos esquemas semánticos diferentes que referenciaremos como *"conceptos"* y *"términos+conceptos"*.

En la primera estrategia denominada "conceptos", se genera un nuevo vector reemplazando todo término de la representación original por su concepto en WN ("synset") y eliminando aquellos términos cuyo synset no existe o no pudo ser desambiguado. Cuando se habla de "términos+conceptos", al vector de términos original se le incorporan todos los conceptos de WN obtenidos en la primera estrategia.

Se debe aclarar que en el trabajo experimental, no sólo se considerarán lo conceptos directamente obtenidos del proceso de desambiguación, sino que también se hará un breve análisis del efecto de considerar aquellos conceptos disponibles siguiendo la relación de *hiperonimia* de WN. Este enfoque ya ha sido considerado en otros trabajos previos que utilizan información semántica con resultados favorables (Hotho, Staab, y Stumme, 2003), (Stein, zu Eissen, y Potthast, 2006).

3. Diseño experimental y análisis de resultados

Para los trabajos experimentales, fueron seleccionadas las siguientes colecciones de textos cortos con diferentes niveles de complejidad con respecto al tamaño de la colección, longitud de los documentos y solapamiento de vocabulario: *CICling-2002*, *EasyAbstract*, *R8+*, *R8-*, *R8*, *R8porc+* y *R8porc-*. *CICling-2002* (*CIC*) es una colección muy popular de textos cortos que ha sido reconocida como de alta complejidad debido a que sus documentos son resúmenes científicos que pertenecen a un dominio muy restringido. La colección *EasyAbstract* (*Easy*) está compuesta de documentos de corta longitud que también son resúmenes científicos, pero que tratan sobre tópicos bien diferenciados entre sí. Las colecciones previas, son colecciones de muy pocos documentos que han permitido en trabajos previos, realizar un análisis detallado

[1] Con un proceso previo de eliminación de palabras de paro (o "stopword") y lematizado de las palabras.

que sería dificultoso llevar a cabo si se trabaja con colecciones de gran tamaño. Desafortunadamente, si sólo estos conjuntos de datos fueran considerados no sería posible determinar si las conclusiones aplican también a colecciones de mayor tamaño. Por esta razón, otras cinco colecciones fueron consideradas en los experimentos: *R8* (Ingaramo et al., 2008), con 8 categorías obtenidas desde el conjunto de datos *Reuters-21578*, y los subconjuntos *R8+*, *R8-*, *R8porc+* y *R8porc-* diferenciándose del original por el tipo y cantidad de documentos en cada una de las 8 clases. En el caso de *R8+* los 20 documentos de mayor tamaño de cada categoría fueron seleccionados, utilizándose el mismo procedimiento para *R8-* pero teniendo en cuenta esta vez los documentos de menor tamaño. Con respecto a *R8porc+* y *R8porc-*, también se buscó en este caso generar categorías con documentos más largos en el primer caso y más cortos en el segundo. La diferencia con *R8+*, *R8-* es que ahora se tomó el 20 % de los documentos más largos de cada clase para *R8porc+* y el 20 % de los documentos más cortos de cada clase para *R8porc-*. De esta manera, estas colecciones mantuvieron el desequilibrio en la cantidad de documentos por clase que presentaba originalmente la colección *R8*. Por lo tanto, la longitud de los documentos de *R8porc+* es, en promedio, 10 veces la longitud de los documentos de *R8porc-*[2].

Como fue especificado en la Sección 1, los documentos son representados mediante el modelo de espacio vector (VSM, las siglas en inglés para Vector Space Model) introducido por Salton para la codificación de textos (Salton, 1971). El VSM utilizado para codificar cada texto fue enriquecido a partir de la incorporación de información semántica, obteniéndose los vectores de "conceptos" y "términos+conceptos". Los "conceptos" fueron obtenidos mediante los tres enfoques que ya fueron descriptos: CIAOSENSO (CIAO), Lesk Mejorado (LM) y el método heurístico del sentido más frecuente (MFS, las siglas en inglés para Most Frequent Sense).

Para realizar las comparaciones, se tomaron como base las 3 codificaciones SMART (entre las 20 posibles) que mejores resultados reportaron con las representaciones de términos originales en todos los experimentos; estas son *btc*, *ltc* y *ntc*.

Para la tarea de categorización de textos se utilizó la herramienta *Weka* (Garner, 1995) con diferentes algoritmos de aprendizaje: Naive Bayes Multinomial Updateable (*NBMU*), Naive Bayes (*NB*), Naive Bayes Multinomial (*NBM*), Complement Naive Bayes (*CNB*) y Support Vector Machine (*SVM*). Para el entrenamiento y validación de los resultados, se utilizó la validación cruzada en k pliegues (*k-fold cross validation*) con $k = 10$.

3.1. Resultados experimentales

La Figura 1 compara los mejores valores de precisión obtenidos con la representación de *términos* versus las nuevas estrategias de *conceptos* y *terminos+conceptos* para todas las colecciones. Para las tres estrategias se seleccionó la combinación "codificación-clasificador-método de WSD basado en conocimiento" que reportó, entre todas las posibles, el más alto valor de porcentaje de instancias clasificadas correctamente. Por ejemplo, para la colección CIC en la representación de conceptos el mayor valor fue determinado a partir de la codificación *ltc*, el clasificador *CNB* y el método de desambiguación *MFS*. De acuerdo a estos resultados es claro que, independientemente de la complejidad de cada colección, la incorporación de información semántica puede lograr una mejora en la precisión, con respecto a los resultados obtenidos cuando sólo los términos son considerados. En las colecciones *Easy*, *R8+*, *R8* y *R8porc+* se puede observar un comportamiento similar, esto es, la representación de "términos" es mejorada levemente por la de "conceptos" y esta última por la de "términos+conceptos" de la misma manera (o en su defecto, igualada en el caso de la colección *Easy*). Para las colecciones CIC, R8- y R8porc-, no se observa el comportamiento mencionado anteriormente. A modo de ejemplo, en el caso de la colección *CIC*, la representación de "conceptos" supera con su valor de precisión a la de "términos" pero con un porcentaje mucho mayor y teniendo la particularidad de que la de "términos+conceptos" no logra superar a la de "conceptos". Este comportamiento puede ser debido a la par-

[2]Las limitaciones de espacio nos impiden dar una descripción más detallada de estas colecciones pero es posible obtener en (Ingaramo et al., 2008) (Makagonov, Alexandrov, y Gelbukh, 2004) (Errecalde et al., 2008) (Errecalde y Ingaramo, 2008) más información acerca de sus características y enlaces para su acceso.

ticularidad de los dominios restringidos de la colección CIC (dominios que en cierta medida se solapan compartiendo algunos términos aunque este grado de ambigüedad parece resolverse en parte con los vectores basados únicamente en conceptos). No obstante esta diferencia observada entre ambos enfoque semánticos, es importante notar que el enfoque semántico de "términos+conceptos" logra superar al enfoque de sólo términos, en todas las colecciones consideradas.

Si bien los resultados presentados previamente, ponen de manifiesto que la incorporación de información semántica puede resultar en una mayor precisión en la categorización de textos cortos, es importante comparar ahora cuál fue el desempeño de los distintos enfoques semánticos en cada una de las distintas instancias experimentales consideradas. En las Tablas 1 y 2: se realiza esta comparación reportándose, por colección, los valores de precisión obtenidos con los diferentes métodos de WSD basados en conocimiento utilizados para "conceptos" (C) y "términos+conceptos" (T+C) escogiendo el clasificador que mostró el mejor comportamiento. Los valores en negrita indican los mejores valores obtenidos. Las colecciones *Easy* y *R8* son, mayormente, mejor categorizadas al aplicar los métodos WSD que tienen en cuenta el contexto de la palabra a desambiguar, esto es LM y CIAO; mientras que en la colección de alta complejidad *CIC* los mejores valores fueron encontrados utilizando el método MFS. Por ejemplo, si examinamos la colección *CIC*, un 75 % de precisión es obtenido con la representación de "conceptos" utilizando MFS y codificación *ntc*. Por el contrario, una tarea de categorización con precisión perfecta se obtiene aplicando, tanto para "conceptos" como para "términos+conceptos", el método de WSD *LM* o *CIAO* para la colección *Easy*. Por otra parte, si tomamos en cuenta las diferentes codificaciones utilizadas, es posible observar que en general, independientemente de la colección y método de WSD utilizados en los experimentos, en la mayoría de los casos, la *ntc* estandar es la que mejor resultados obtuvo, seguida por la codificación *ltc*.

Uno de los problemas con los enfoques semánticos como el de "términos+conceptos" es que implican un aumento considerable en la dimensionalidad de la representación de los documentos. Este problema, puede resultar muy serio cuando se deben categorizar colecciones con un tamaño grande de vocabulario, como es el caso de la colección R8, donde enfoques de este tipo pueden tener un impacto negativo en la eficiencia en tiempo y espacio de sistemas de categorización como los provistos por Weka.

Un aspecto interesante a analizar en estos casos, es en qué medida los enfoques semánticos son robustos a los métodos clásicos de reducción de dimensionalidad, para poder obtener colecciones que se puedan procesar de manera más eficiente. Para analizar este aspecto, se seleccionó la colección R8, la mayor en cuanto al tamaño del vocabulario de todos los conjuntos de datos utilizados y se aplicó uno de los filtros provistos por Weka para reducir el número de atributos en el vocabulario. En la Figura 2 se muestran los resultados obtenidos aplicando un filtro de selección de atributos con el método *Ganancia de Información*. Las representaciones fueron obtenidas mediante el método *LM* y la codificación *ntc* clasificadas con el algoritmo *CNB*. Los diferentes tamaños de vocabulario fueron determinados tomando como referencia el vocabulario de "conceptos", el cual es el menor de las tres representaciones utilizadas.

Si se consideran los distintos tamaños de vocabularios en la Figura 2, es posible observar que el enfoque de sólo conceptos supera al de "términos+conceptos" en los tamaños más pequeños, pero que a partir de un tamaño de 2500 el enfoque de "términos+conceptos" muestra los mejores valores de precisión. Estos valores no mejoran (e incluso la precisión disminuye) con vocabularios superiores a 2500, con lo que se muestra que con "términos+conceptos" se puede capturar con vocabularios relativamente pequeños, toda la información necesaria para realizar una categorización con alta precisión.

El último aspecto considerado en nuestro estudio fue el impacto de agregar los hiperónimos de los conceptos obtenidos, un enfoque propuesto en trabajos previos que utilizan información semántica (Hotho, Staab, y Stumme, 2003), (Stein, zu Eissen, y Potthast, 2006).

Dado que un análisis detallado de la hiperonimia escapa a los alcances de este trabajo, nuestro análisis se restringió a aquella colección que mayor dificultad había presentado a los distintos métodos (CIC) y se tomó como nivel de hiperonimia el corres-

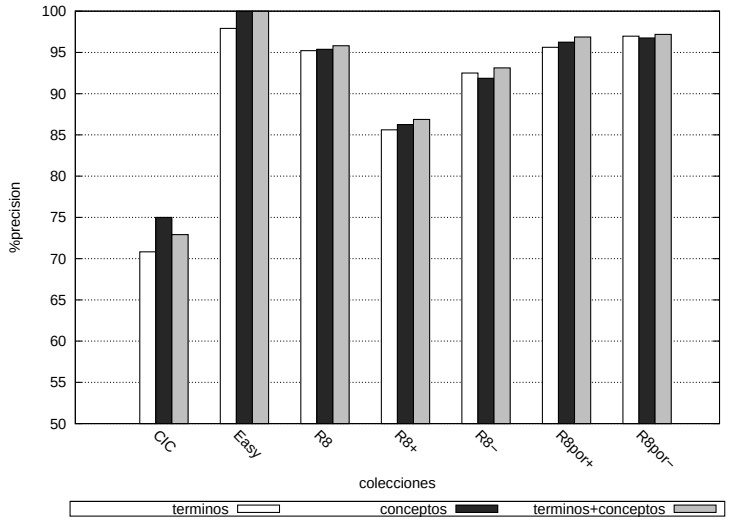

Figura 1:
Sin Información Semántica vs. *Información Semántica* para todas las colecciones

	CICling-2002			EasyAbstracts			R8		
C	LM	$CIAO$	MFS	LM	$CIAO$	MFS	LM	$CIAO$	MFS
btc	62.50	62.50	58.33	87.50	93.75	87.50	94.56	94.43	94.06
ltc	60.41	64.58	66.66	91.65	95.83	91.66	**95.38**	95.12	94.86
ntc	62.50	66.66	**75.00**	97.91	**100**	95.83	**95.38**	95.21	**95.38**
T+C	LM	$CIAO$	MFS	LM	$CIAO$	MFS	LM	$CIAO$	MFS
btc	60.41	58.33	60.41	93.75	91.66	89.58	95.38	95.34	95.77
ltc	58.33	66.66	70.83	93.75	93.75	93.75	95.55	95.64	95.25
ntc	66.66	70.83	**72.91**	**100**	**100**	97.91	**95.81**	95.60	95.68

Tabla 1: Mejores valores de precisión diferenciando por colección cada sistema de WSD

	R8+			R8-			R8porc+			R8porc-		
C	LM	$CIAO$	MFS	LM	$CIAO$	MFS	LM	$CIAO$	MFS	LM	$CIAO$	MFS
btc	85.00	84.37	83.12	90.00	90.00	90.00	95.00	94.37	92.5	95.68	96.46	96.32
ltc	85.62	**86.25**	85.00	90.62	90.00	**92.5**	95.62	**96.25**	95.00	95.89	96.32	96.11
ntc	**86.25**	83.75	81.25	91.87	90.62	91.25	**96.25**	93.75	93.12	96.11	96.11	**96.54**
T+C	LM	$CIAO$	MFS	LM	$CIAO$	MFS	LM	$CIAO$	MFS	LM	$CIAO$	MFS
btc	86.25	85.62	84.37	**92.50**	90.62	90.62	96.25	95.62	94.37	97.84	97.19	**97.89**
ltc	**86.87**	85.00	85.62	91.87	91.25	90.62	**96.87**	95.00	95.62	97.62	97.19	97.19
ntc	81.87	83.12	82.50	**92.50**	92.12	**92.50**	91.87	93.12	91.87	97.40	97.19	96.32

Tabla 2: Mejores valores de precisión diferenciando por colección cada sistema de WSD

	Sin hiperónimos					Con hiperónimos				
	$NBMU$	CNB	NBM	NB	SVM	$NBMU$	CNB	NBM	NB	SVM
MVP	35.41	66.66	41.66	**50**	**60.41**	**54.16**	**72.00**	**60.41**	**50**	45.83

Tabla 3: Resultados *sin hiperónimos* vs *con hiperónimos* para la colección CICling-2002 con CIAO como WSD

pondiente a los mejores resultados reportados en (Hotho, Staab, y Stumme, 2003) (nivel de hiperonimia 5 en la jerarquía de WN). Los conceptos en este caso se obtuvieron mediante el método CIAO. En la Tabla 3 se resumen los mejores valores obtenidos para *CIC* con los distintos algoritmos en el caso de no usar hiperónimos (izquierda) y con el uso de hiperónimos (derecha). Observando los mejores valores de precisión (MVP) obtenidos en cada caso, podemos apreciar que en algoritmos como $NBMU$, CNB y NBM el uso de hiperónimos muestra algunas mejoras sobre los resultados sin hiperóni-

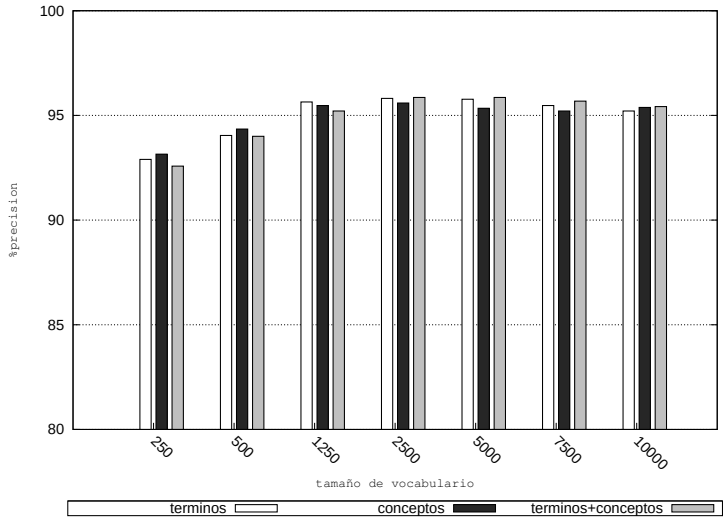

Figura 2:
Reducción del tamaño de vocabulario para la colección R8

mos, en NB los resultados son similares y SVM muestra un claro deterioro en la precisión al introducir los hiperónimos. Por lo tanto, podemos concluir que la efectividad del uso de este tipo de información en la representación de los documentos, depende significativamente del método de categorización utilizado y se requiere de un estudio más detallado para determinar la conveniencia o no de incorporar este tipo de información en la categorización de textos cortos.

4. Conclusiones y trabajos futuros

El objetivo principal de este trabajo fue determinar si la incorporación de información semántica en la representación de documentos, mediante métodos de WSD basados en conocimiento, ayuda a mejorar la tarea de categorización de colecciones de textos cortos. Dos estrategias fueron consideradas: "conceptos" y "términos+conceptos". Se evaluaron diferentes codificaciones y diferentes clasificadores, como también si la reducción de vocabulario afecta el comportamiento de la tarea de categorización. Se concluye que el uso de información semántica en la representación de los documentos, a través de métodos basados en conocimiento, puede ser beneficioso para la tarea de categorización de documentos cortos, en especial el enfoque que denominamos "términos+conceptos". Este enfoque demostró además que si bien, en general involucra un aumento en la dimensionalidad de la representación, es muy robusto a la aplicación de métodos de reducción de vocabulario y puede obtener una muy alta precisión con altos niveles de reducción de vocabulario.

Con respecto al desempeño de los distintos métodos de WSD en estos casos, teniendo en cuenta que no hay un solo método de WSD basado en conocimiento que presente los mejores resultados para todas las colecciones, se puede afirmar que un óptimo WSD general no puede ser determinado en base a los experimentos realizados. Por tal motivo, el mejor valor de precisión depende de la codificación y clasificador utilizado.

Por último, las ventajas del uso de hiperónimos en la representación de documentos cortos no han quedado demostradas y es un punto que requiere investigación futura. En este sentido, una extensión interesante es la planteada en (Hotho, Staab, y Stumme, 2003) para la tarea de categorización no supervisada (o "clustering") que consiste en agregar a la representación de "conceptos" aquellos términos que no pudieron ser desambiguados.

Otras extensiones posibles a nuestro trabajo, es la consideración de otras colecciones de textos cortos en los experimentos, la utilización de otras técnicas de reducción de vocabulario e incorporar nuevos métodos de WSD basados en conocimiento como por ejemplo otras variantes del algoritmo de Lesk (Simplificado y aplicando "Simulated aneal-ing") y otros métodos heurísticos.

Bibliografía

Agirre, Eneko y Philip Edmonds, editores. 2006. *Word Sense Disambiguation: Algorithms and Applications*, volumen 33 de *Text, Speech and Language Technology*. Springer.

Banerjee, Satanjeev y Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. En *CICLing*, páginas 136–145.

Banerjee, Somnath, Krishnan Ramanathan, y Ajay Gupta. 2007. Clustering short texts using wikipedia. En *SIGIR*, páginas 787–788.

Buscaldi, Davide, Paolo Rosso, y Francesco Masulli. 2004. The upv-unige-ciaosenso wsd system. En *SENSEVAL-3: 3rd International Workshop on the Evaluation of Systems, Association for Computational Linguistics for the Semantic Analysis of Text*, páginas 77–82, Barcelona, Spain.

Errecalde, Marcelo, Leticia Cagnina, Diego Ingaramo, y Paolo Rosso. 2008. A discrete particle swarm optimizer for clustering short-text corpora. En *BIOMA08*, páginas 93–103.

Errecalde, Marcelo y Diego Ingaramo. 2008. Short-text corpora for clustering evaluation. Informe técnico, LIDIC.

Garner, Stephen. 1995. Weka: The waikato environment for knowledge analysis. En *In Proc. of the New Zealand Computer Science Research Students Conference*, páginas 57–64.

Hotho, Andreas, Steffen Staab, y Gerd Stumme. 2003. Ontologies improve text document clustering. En *ICDM*, páginas 541–544.

Ingaramo, Diego, David Pinto, Paolo Rosso, y Marcelo Errecalde. 2008. Evaluation of internal validity measures in short-text corpora. En *Proc. of the CICLing 2008 Conf.*, volumen 4919 de *LNCS*, páginas 555–567. Springer-Verlag.

Lesk, Michael. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. En *Proceedings of the 5th International Conference on Systems Documentation*.

Makagonov, Pavel, Mikhail Alexandrov, y Alexander Gelbukh. 2004. Clustering abstracts instead of full texts. En *Proc. of TSD-2004*, volumen 3206 de *LNAI*, páginas 129–135.

Miller, George. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Rosso, Paolo, Francesco Masulli, Davide Buscaldi, Ferran Pla, y Antonio Molina. 2003. Automatic noun sense disambiguation. En *CICLing*, páginas 273–276.

Salton, Gerard. 1971. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc.

Salton, Gerard y Chris Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523.

Sebastiani, Fabrizio. 2005. Text categorization. En Laura Rivero Jorge Horacio Doorn, y Viviana Ferraggine, editores, *Encyclopedia of Database Technologies and Applications*. Idea Group, páginas 683–687.

Stein, Benno, Sven Meyer zu Eissen, y Martin Potthast. 2006. Syntax versus semantics: Analysis of enriched vector space models. En Benno Stein y Odej Kao, editores, *3rd International Workshop on Text-Based Information Retrieval (TIR 06)*, páginas 47–52. University of Trento, Italy, August.

Vázquez, Sonia. 2009. *Resolución de la ambigüedad semántica mediante métodos basados en conocimiento y su aportación a tareas de PLN*. Ph.D. tesis, Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante.