HOLOS
ISSN 1807 - 1600

REPRESENTACIÓN CONCEPTUAL PARA LA CLASIFICACIÓN MULTILINGUE DE TEXTOS

A. Borges García[*], D. Castro Castro y R. Ortega-Bueno
DATYS Tecnologías y Sistemas, Santiago de Cuba, Cuba
arisbel.borges@datys.cu[*]

Recibido 10/06/2016 - Aceptado 19/04/2018

DOI: 10.15628/holos.2018.4682

RESUMEN

Hoy en día, el porcentaje de la información disponible en inglés en *Word Wide Web* está disminuyendo, debido a que otros lenguajes como: chino, español, árabe y portugués están ganando aceptación y difusión. Este fenómeno ha provocado que el multilingüismo se convierta en uno de los principales retos para el procesamiento inteligente, gestión y recuperación de documentos. Con el fin de hacer frente a este problema de forma eficaz, los sistemas computacionales necesitan el diseño de nuevos modelos o mejorar los modelos tradicionales de representación de documentos. La disponibilidad de repositorios multilingües de conceptos y redes semánticas, ha abierto un enfoque atractivo para modelar documentos escritos en diferentes lenguas, como los vectores de conceptos en un espacio común de representación. En este trabajo se presenta una nueva representación basada en conceptos usando *Multilingual Central Repository*. Nuestra propuesta aplica una resolución de la ambigüedad semántica del sentido de la palabra de grano grueso para la selección del concepto apropiado de acuerdo con el tema y los dominios relevantes reflejados en los documentos. Evaluamos experimentalmente nuestro método en la tarea de clasificación de documentos multilingües. Los resultados obtenidos en los experimentos son alentadores y demuestran la utilidad del método propuesto.

PALABRAS-CLAVES: Representación conceptual de documentos, Clasificación multilingüe de documentos, Desambiguación conceptual por dominio.

USING CONCEPTUAL DOCUMENT REPRESENTATION FOR MULTILINGUAL TEXT CLASSIFICATION

ABSTRACT

Nowadays, the percentage of english information available in *Word Wide Web* is decreasing, because other languages such as: Chinese, Spanish, Arabic and Portuguese are gaining acceptance and dissemination. This phenomenon has caused that multilingualism become as one of the major challenges for intelligent documents processing, management, and retrieval. In order to deal with this problem efficiently, computer's system need to design new models or improve traditional models for documents representation. The availability of multilingual concepts repositories and semantic networks has opened an attractive approach to model documents written in different languages as concept vectors into a common space of representation. In this paper we present a new concept-based representation using *Multilingual Central Repository*. Our proposals apply a coarse-grained word sense disambiguation for selecting the appropriate concept according to topic and relevant domains discussed in documents. We experimentally evaluate our proposed method into a multilingual document classifications task. The results obtained in the experiments are encouraging, and demonstrate the usefulness of the proposed method.

KEYWORDS: Conceptual document representation, Multilingual document classification, Domain concept desambiguation.

HOLOS
ISSN 1807 - 1600

1 INTRODUCCIÓN

Uno de los resultados más importantes de la informatización global de la sociedad, es que internet se ha convertido en una fuente primaria de información multimodal, multilingüe, y sobre una gran variedad de formatos. En noviembre del 2015, aproximadamente 257 millones de usuarios de internet hablaban español, esto representaba un aumento de un 1 300 % en relación a los últimos 5 años; tres veces más que el inglés (De Argaez, 2015). Se puede inferir que el contenido publicado en idioma español también ha aumentado.

Además de la enorme cantidad de información y los desafíos que esto impone, la diversidad de idiomas empleados en la generación de contenido en la web actual, dificulta, en gran medida, la comprensión, unificación, filtrado, organización y síntesis de todo este contenido. Actualmente, gran número de usuarios tienen dominio sobre diferentes idiomas; sin embargo, muchos de ellos no son capaces de realizar de manera eficiente y eficaz tareas como: la desambiguación del sentido de las palabras, la construcción de resúmenes, y la categorización de textos sobre colecciones de documentos escritas en idiomas diferentes a su lengua natal.

Por tal razón, la creación de nuevos algoritmos que permitan analizar estas fuentes multilingües de manera automática y precisa, con el objetivo de transformarlas en conocimiento valioso para la toma de decisiones, resulta el punto de mira de diversos investigadores. Con el fin de resolver el problema en cuestión, resulta de gran importancia, construir un modelo matemático que permita representar y computar, las semejanzas y diferencias entre un par de documentos, escritos en diferentes idiomas.

La forma de representar los documentos, conocidas también como indización de documentos o indexación de documentos, es uno de los procesos más elementales y primarios indispensables en el procesamiento automático de textos. Uno de los métodos de indexado tradicional es el que obtiene un modelo basado en vector (Franco-Salvador et al, 2014), esta representación es conocida como bolsa de palabras (BOW).

En el modelo vectorial, cada documento es representado como un vector en un espacio de alta dimensionalidad y con un elevado grado de dispersión. Formalmente cada documento es definido de la siguiente forma: d = < t1:w1, t2:w2,..., tn:wn> donde ti, representa el conjunto de términos (palabras, lemas, raíces, conceptos, etc.) presentes en el documento y wi representa la importancia de ese término dentro de ese documento. Dicha importancia o peso, en muchos casos, no es más que la frecuencia de aparición de ese término dentro del texto.

Una de las ventajas del uso de esta representación es que es muy rápida y fácil de implementar. Una de las principales desventajas radica en el tamaño del vocabulario usado, como cada término es un componente del vector, se produce un incremento en la dimensión. En el entorno multilingüe el modelo basado en vector ha sido ampliamente usado (Romeo et al, 2015; Chebel et al, 2015).

HOLOS ISSN 1807 - 1600

En este trabajo se abordan los resultados que se obtienen con la introducción de un nuevo método para la representación multilingüe de documentos. Con el objetivo de realizar un análisis profundo de los mismos y poder validarlos solo se han seleccionado los idiomas inglés y español.

2 REVISIÓN BIBLIOGRÁFICA

La traducción automática se puede aplicar en el proceso de indexación (Zhou et al, 2012). Se trata de determinar inicialmente la representación del texto, en dependencia de su idioma y luego este se traduce a un idioma de interés. En este contexto existen dos esquemas, la traducción de la representación del texto a los idiomas de los documentos restantes, o la traducción de la representación de los documentos restantes al idioma del texto en cuestión. Para atenuar los inconvenientes de la escalabilidad, se ha optado por definir un idioma como forma intermedia, y todos los documentos son traducidos a una sola lengua, convirtiendo el enfoque en una tarea monolingüe.

Otra dificultad de la puesta en práctica de este enfoque es la polisemia de las palabras; a pesar de que se empleen técnicas para mitigar esta dificultad el resultado de un sistema de traducción automática de una palabra debe ser único, sin embargo un sistema de recuperación multilingüe de información puede dar más de una variante y asignarle distintos pesos (López Ostenero et al, 2004; Bikel, y Zitouni, 2012).

Recientemente se ha observado un auge en el uso de entidades nombradas (Herranz, 2013; Bermúdez, 2013), para abordar la indexación automática multilingüe. Los costes computacionales obtenidos al utilizar esos resultados son bajos, porque reducen en gran medida el espacio de representación del documento; y es esa precisamente su principal desventaja, al ignorar otros elementos léxicos muy importantes, como son los sustantivos comunes y las formas verbales.

Con el propósito de solventar las limitaciones del uso de las entidades nombradas y el de traducción de los documentos, y teniendo en cuenta la disponibilidad de redes semánticas multilingües (RSM) (Salvador, 2013), se ha estado explorando una alternativa atractiva: la utilización de los registros de índice interlingüe (ILI) para representar de manera unificada, tanto las consultas como los documentos, utilizando la comparación de conceptos, más que la comparación de palabras.

En tal sentido, las perspectivas de análisis para abordar este campo, se fundamentan generalmente en dos propuestas, a partir de enfoques orientados al empleo de los términos de las descripciones de los conceptos, realizando una expansión (Sy et al, 2012) mientras que otros abogan por la sustitución completa de los términos, lemas, o raíces, por conceptos (Cisneros et al, 2012).

Tabla 1 Subconjunto de ILI del lema player en sustantivo de la red semántica multilingüe *Multilingual Central Repository*.

ILI	Synset
10439851n	player-n#1

HOLOS ISSN 1807 - 1600

10340312n	musician-n#1
09765278n	actor-n#1

La expansión por descripciones hace más crítico el problema de la dimensionalidad. En la variante de sustitución por ILI también ocurre este fenómeno porque, como se observa en la tabla 1, un lema puede tener asociado más de un concepto, y por ende, más de un ILI. En el caso del lema player en sustantivo para la primera aserción, se refiere a "persona que participa o es experta en algún juego", en el caso del lema musician en sustantivo para la primera aserción, se describe como "alguien que toca un instrumento musical (como profesión)" mientras que el lema actor en sustantivo para la primera aserción, hace alusión a "un actor, histriónico, actriz, alguien que juega un rol como intérprete teatral", la variante de seleccionar todos los synsets (AS) se empleará en este trabajo como punto de comparación (*baseline*, en inglés).

Una estrategia de selección de ILI, sencilla pero con muy buenos resultados reportados en la literatura, es seleccionando el sentido más frecuente (MFS). El problema de este método consiste, en que no en todos los textos, se emplea la acepción más frecuente de la palabra.

Otros trabajos emplean una estrategia de desambiguación gruesa, empleando hiperónimos o alguna ontología de tópicos (Walker, y Amsler, 2013). Un recurso de este tipo muy empleado es *WorNet Domains* (WD) (Magnini et al, 2001). WD es un recurso léxico-semántico donde los synsets han sido anotados de forma semiautomática, con una o varias etiquetas de dominio de un conjunto de 165 etiquetas, estructuradas jerárquicamente. La anotación que provee WD brinda la posibilidad de reducir el nivel de polisemia de las palabras y agrupar aquellos sentidos que pertenecen al mismo dominio del conocimiento.

En la tabla 2 se presenta la distribución de una muestra de dominios sobre synsets de WD. Se considera la etiqueta de dominio "factotum" para aquellos conceptos que tienen un carácter general, por lo tanto pueden ser empleados en varios dominios. Tal es el caso de los conceptos asociados a día, persona, color, número, etc.

Tabla 2: Distribución de algunos dominios sobre los synsets de *WorNet Domains.*

Dominios	Cantidad de synsets
Factotum	36820
Biology	21281
Transport	2443
Comerce	637
Veterinary	92

HOLOS
ISSN 1807 - 1600

3 METODOLOGÍA

En este trabajo se propone una estrategia de selección de synsets usando *Multilingual Central Repository* MCR (Agirre et al, 2012) y el recurso *Wordnet Domains* (WD).

En la figura 1 se propone un esquema para la clasificación multilingüe. Tanto la representación de los documentos de entrenamiento y la del nuevo documento a clasificar se realiza siguiendo el mismo procedimiento.

La premisa fundamental es que se pueden seleccionar solo aquellos synsets pertenecientes a los dominios más frecuentes en el texto. Para ello, el texto es segmentado, y finalmente es etiquetado morfológicamente. Nuestra propuesta parte de la representación vectorial del texto.

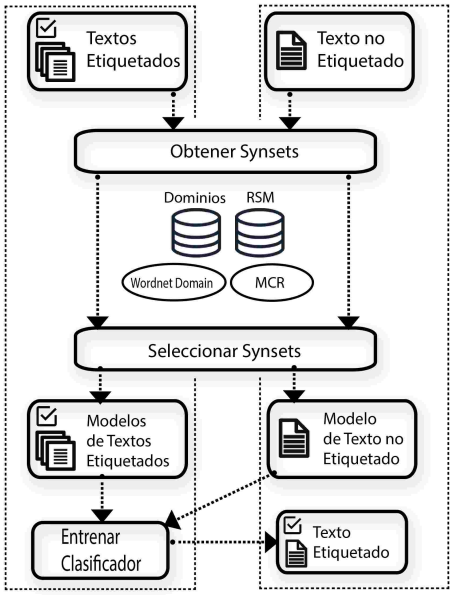

Figura 1: Esquema propuesto para la clasificación usando representación a nivel de concepto.

En el trabajo no se han tenido en cuenta las anotaciones relacionadas con el dominio "Factotum". El método se describe en el siguiente algoritmo:

Entrada:

Conjunto de lemas *L* sin synset seleccionado.

Pasos:

(1) Para cada dominio *d*, y sysnset *s* que pertenecen a WD, asignar peso cero:

$w(d) = 0, w(s)=0.$

(2) Para cada lema *l* que pertenece a *L*, obtener el conjunto S_l de todos los synsets de *l* en MCR.

(3) Para cada dominio d_s asociado a cada synset *s* que pertenece a *S* incrementar peso en 1: $w(d_s) = w(d_s) + 1.$

HOLOS ISSN 1807 - 1600

(4) Sea MFD el dominio con mayor peso

(5) Para cada lema *l* que pertenece a *L*, sea S_l' el subconjunto de S_l de todos los synset *s* asociados a MFD, sustituir lema *l* por cada synset *s*, *w(s) = w(s) * w(MFD).*

Se repite el procedimiento mientras existan lemas sin synsets seleccionados.

Con el fin de evaluar la factibilidad del método propuesto en este trabajo se ha seleccionado las máquinas de soporte vectorial SVM; este método ha demostrado, en los últimos años, una gran efectividad en la clasificación automática de textos (Del Pilar et al, 2014). Como estrategias de selección del mejor synset se han evaluado, además, el synset más frecuente (MFS), y otra variante utilizando todos los synsets (AS).
Para evaluar el desempeño del clasificador se utilizaron las medidas tradicionales de precisión (P) y cobertura, o relevancia, (R) (Van Asch, 2013; Sebastian, 2002), respecto a la clasificación realizada por un humano:

$$\mathrm{R} = \frac{VP}{VP + FN} \quad (1)$$

$$\mathrm{P} = \frac{VP}{VP + FP} \quad (2)$$

$$\mathrm{F1} = \frac{2 * P * R}{P + R} \quad (3)$$

Donde VP es la cantidad de textos clasificados correctamente (verdaderos positivos), FP es la cantidad de textos clasificados incorrectamente, y FN es el número de textos que no clasificaron (falsos negativos).

El método predictivo SVM y la medida F1 empleados se corresponden con los implementados en WEKA (Bouckaert et al, 2010). Los parámetros del clasificador son los definidos por defecto en dicha plataforma. Los resultados de las medidas calculados por WEKA tienen en cuenta el cardinal de cada clase; los valores que se muestran en este trabajo coinciden con la fórmula tradicional aplicada de manera global (F1-Macro).

4 RESULTADOS Y DISCUSIÓN

4.1 Experimento con colección de documentos de wikipedia

Se construyó una colección de pruebas donde estuviera presente el fenómeno del uso o no del sentido más frecuente. Intuitivamente, se crearon temas en cada idioma relacionados con palabras con más de un significado. También se tomó en consideración que fuera diferente el uso frecuente de esas palabras en cada idioma. Para esto, se crearon unidades textuales, a partir de los epígrafes principales de las páginas de wikipedia relacionadas con las consultas “jaguar pantera”, “jaguar auto”, “canal tv”, y “canal marítimo” para los idiomas inglés y español. Los textos seleccionados no coinciden con

HOLOS ISSN 1807 - 1600

traducciones exactas de textos del otro idioma. La dimensión de cada texto es similar a textos de género noticioso como los anotados en RCV1-RCV2 (Amini, 2009).

Los resultados de la tabla 3 indican que la dimensión de la representación usando todos los sentidos, es tres veces superior a la obtenida empleando BOW. La causa del aumento de dimensionalidad es el vocabulario polisémico que se ha empleado.

Tabla 3: Dimensión promedio de las representaciones en los idiomas ingles y español usando vector de palabras, vector de todos los sentidos, vector con los sentidos más frecuente, y vector con los sentidos de los dominios más frecuentes .

Idioma	BOW	AS	MFS	MFD
In	65	186.9	42	91
Es	66	186.2	35	93.8

La mayor reducción de dimensionalidad se arroja empleando MFS, la causa de que se reduzca la dimensión del vector es que no todas las palabras están registradas en la RSM, el caso más crítico es el español. Con MFD se reduce a la mitad la cantidad de synsets, esto indica que se han seleccionado aproximadamente dos variantes por cada lema.

Al analizar los valores de la medida F1 expuestos en la tabla 4, se puede percibir que los resultados empleando MFD, en el ambiente translingüe, superan los otros dos métodos. En el entorno monolingüe, los mismos son comparables con los obtenidos usando MFS.

Tabla 4: Valores de la medida F1 obtenidos con la colección de wikipedia.

	Monolingüe		Translingüe	
Método	In	Es	Es_In	In_Es
AS	0.47	0.65	0.58	0.50
MFS	0.55	0.68	0.53	0.61
MFD	0.55	0.63	**0.81**	**0.75**

En la tabla 5 se toman los valores de las matrices de confusión (Ej, experimento Es_In); se puede observar cómo afecta el uso del MFS. El caso más evidente ocurre en las clases "canal panamá" y "canal tv", documentos de la clase "canal panamá" clasificaron usando el sentido más frecuente en la clase "canal tv". Sin embargo, al usar nuestra propuesta, la clasificación en esa clase fue excelente. La clasificación en la clase "canal tv" usando MFD no fue buena, no obstante los problemas de la clasificación no se relacionaron con la clase "canal panamá", sino con el resto de las clases. La clasificación en la clase "jaguar auto" empleando MFD es notablemente superior a los resultados alcanzados usando AS.

Tabla 5. Matrices de confusión experimento Es-In.

HOLOS ISSN 1807 - 1600

AS				MFS				MFD				
b	*c*	*d*	*e*	*b*	*c*	*d*	*e*	*b*	*c*	*d*	*e*	<- Clasifica como
12	5	0	1	4	12	0	2	16	0	1	1	b = canal panamá
2	5	0	0	0	7	0	0	1	3	2	1	c = canal tv
1	0	6	0	1	0	6	0	1	0	6	0	d = jaguar animal
2	4	0	0	0	2	0	4	0	0	0	6	e = jaguar auto

4.2 Experimento con colección de documentos de RCV1-RCV2

RCV1-RCV2 ha sido empleada como una colección de referencia internacional. En la tabla 6 se presentan las 8 clases seleccionadas. Cada clase posee una representación de documentos en inglés y español. La cantidad de documentos en español coincide, exactamente, con el número de documentos existente en el corpus de ese idioma, y que clasifican solo en las respectivas clases. La cantidad de documentos en inglés se seleccionó de manera uniforme. Todos los parámetros, excepto la colección de pruebas, coinciden con la configuración del experimento anterior.

Tabla 6. Subconjunto de 8 clases seleccionadas de RCV1-RCV2.

Clase	ES	EN
C183	205	401
GCRIM	157	401
GDEF	83	401
GDIP	234	401
GDIS	116	401
GJOB	197	401
GSPO	84	401
GVIO	306	401

En la clasificación supervisada de documentos, al comparar métodos teniendo en cuenta la medida F1, una diferencia menor a 0.1 puntos no se considera importante (Perea Ortega et al, 2008). Teniendo en cuenta lo anterior, en el ambiente translingüe, a pesar de obtenerse resultados inferiores a los dos métodos, la diferencia no es notablemente substancial (0.04 puntos en comparación con el mejor método AS en el experimento Es_In, y 0.06 puntos en comparación con los otros dos métodos en el experimento In_Es).

Por otra parte, la diferencia entre los resultados, monolingüe y translingüe, usando MFD, es menos drástica que los otros dos métodos. Estos valores de la medida F1 están destacados en la tabla 7.

Tabla 7: Valores de la medida F1 colección de referencia internacional RCV1-RCV2.

	Monolingüe		Translingüe	
Método	In	Es	Es_In	In_Es
AS	0.85	0.66	0.65	0.64

HOLOS ISSN 1807 - 1600

MFS	0.83	0.82	0.63	0.64
MFD	0.76	0.59	**0.61**	**0.58**

En una investigación similar (Bentaallah, y Malki, 2012) se usan las 8 clases descritas en este experimento, pero realizan solo la clasificación In_Es. De igual forma, no seleccionan sentidos, es decir, se lleva a cabo un procedimiento similar al que emplear todos los sentidos (AS), con la diferencia de que se actualiza la frecuencia atendiendo a la relación, presente en *Multilingual Central Repository*, entre los sentidos. Por otro lado, el valor de la medida F1 de ese experimento no superó los 0.27 puntos, notablemente inferior a los expuestos en la tabla 7. El mismo experimento In_Es se repite traduciendo los documentos escritos en idioma inglés al español, lo que en nuestro caso sería una prueba similar a la monolingüe Es; el valor alcanzado de la medida F1 fue de 0.64 puntos. A pesar de que esa prueba supera al valor de MFD del experimento ES (0.59) de este trabajo en 0.05 puntos; es válido aclarar que se debió realizar un proceso de traducción con el consecuente consumo de tiempo, y los inconvenientes de escalabilidad del mismo explicados con anterioridad.

5 CONCLUSIONES

Con la realización de este trabajo se ha definido un método, dominio más frecuente (MFD, por sus siglas en inglés), de fácil implementación para la representación conceptual de textos escritos en español e inglés. A partir del estudio realizado durante la implementación del método, se ha podido comprobar, que la introducción del algoritmo de MFD para la selección de synsets, proporciona resultados satisfactorios durante la clasificación de textos escritos en español e inglés. La dimensionalidad se reduce prácticamente en un 50 % con relación a la estrategia AS. Los experimentos realizados con la colección de referencia internacional arrojaron, que a pesar de que los resultados fueron inferiores, la diferencia no resulta importante en relación a los otros dos métodos. Por otra parte, se pudo verificar que los resultados en ambiente monolingüe y translingüe se mantienen semejantes usando nuestra propuesta, y mejoran, incluso, en uno de los dos experimentos.

Se debe continuar experimentando el uso y la efectividad de este algoritmo con otros idiomas y con el uso de otros recursos de dominios y redes semánticas multilingües.

6 REFERENCIAS

Agirre, A. G., Laparra, E., Rigau, G., & Donostia, B. C. (2012). Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *GWC 2012 6th International Global Wordnet Conference* (p. 118).

Amini, M., Usunier, N., & Goutte, C. (2009). Learning from multiple partially observed views-an application to multilingual text categorization. In *Advances in neural information processing systems* (pp. 28-36).

Bermúdez, J. (2013). Reconocimiento conjunto de entidades nombradas y de correferencia para mejorar el acceso a la información multilingüe. *Informe de tesis doctoral. Bilbao:*

Universidad de Deusto.

Bentaallah, M. A., & Malki, M. (2012). The Use of WordNets for Multilingual Text Categorization: A Comparative Study. In *ICWIT*(pp. 121-128).

Bikel, D., & Zitouni, I. (2012). Multilingual natural language processing applications: from theory to practice. IBM Press.

Bouckaert, R. R., Frank, E., Hall, M. A., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2010). WEKA–Experiences with a Java Open-Source Project. *Journal of Machine Learning Research*, *11*(Sep), 2533-2541.

Chebel, M., Latiri, C., & Gaussier, E. (2015, September). Multilingual documents clustering based on closed concepts mining. In *International Conference on Database and Expert Systems Applications* (pp. 517-524). Springer International Publishing.

Cisneros, D. S., Bedmar, I. S., & Fernández, P. M. (2012). Prototipo buscador de información médica en corpus multilingües y extractor de información sobre fármacos. *Procesamiento del Lenguaje Natural*, *49*, 209-212.

De Argaez, E. (2015). Internet world stats. Obtenido de HTTP://www.internetworldstats.com/stats7.htm

del Pilar, S. M., Rodríguez-García, M. Á., & Valencia-García, R. (2014). Estudio de las categorías LIWC para el análisis de sentimientos en español. In *TIMM* (pp. 33-36).

Franco-Salvador, M., Rosso, P., & Navigli, R. (2014, April). A Knowledge-based Representation for Cross-Language Document Retrieval and Categorization. In *EACL* (Vol. 14, pp. 414-423).

Herranz, S. M. (2013). Estudio y nuevas estrategias en el uso de las Entidades Nombradas en el Clustering Bilingüe de noticias(Doctoral dissertation, Universidad Rey Juan Carlos).

López Ostenero, F., Gonzalo, J., & Verdejo, F. (2004). Búsqueda de información multilingüe: estado del arte. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, *8*(22).

Magnini, B., Strapparava, C., Pezzulo, G., & Gliozzo, A. (2001, July). Using domain information for word sense disambiguation. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems* (pp. 111-114). Association for Computational Linguistics.

Perea Ortega, J. M., Valdivia, M., Teresa, M., Montejo Ráez, A., & Díaz Galiano, M. C. (2008). Categorización de textos biomédicos usando UMLS. *Procesamiento del lenguaje natural. N. 40 (abril 2008); pp. 121-127*.

Romeo, S., Ienco, D., & Tagarelli, A. (2015, March). Knowledge-based representation for transductive multilingual document classification. In *European Conference on Information Retrieval*(pp. 92-103). Springer, Cham.

Salvador, F. (2013). M.: Detección de plagio translingüe utilizando una red semántica multilingüe. *Departamento de Sistemas Informáticos y Computación*.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

HOLOS
ISSN 1807 - 1600

Sy, M. F., Ranwez, S., Montmain, J., Regnault, A., Crampes, M., & Ranwez, V. (2012). User centered and ontology based information retrieval system for life sciences. *BMC bioinformatics*, *13*(Suppl 1), S4.

Van Asch, V. (2013). Macro-and micro-averaged evaluation measures. *Tech. Rep.*

Walker, D., & Amsler, R. (1986). The use of machine-readable dictionaries in sublanguage analysis. *Analyzing Language in Restricted Domains*, 69-83.

Zhou, D., Truran, M., Brailsford, T., Wade, V., & Ashman, H. (2012). Translation techniques in cross-language information retrieval. *ACM Computing Surveys (CSUR)*, *45*(1), 1.