Procesamiento del Lenguaje Natural, Revista nº **28**, mayo de 2002

SEPLN

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Artículos

Tesis

Información General

Impresos de Inscripción

Procesamiento del Lenguaje Natural, Revista nº **28**, mayo de 2002

SEPLN

ISSN: 1135-5948
Depósito Legal: B:3941-91
Distribuye: Sociedad Española para el Procesamiento del Lenguaje Natural

Procesamiento del Lenguaje Natural, núm. 28 (2002), pp. 55-62 recibido 9-4-2002; aceptado 2-5-2002

Generación de un tesauro de similitud multilingüe a partir de un corpus comparable aplicado a CLIR

Manuel García Vega
Fernando Martínez Santiago
L. Alfonso Ureña López
Mª Teresa Martín Valdivia
Departamento de Informática
Universidad de Jaén
Av. Madrid, 37. E-23071, Jaén.
{mgarcia, dofer, laurena, maite}@ujaen.es

Resumen: En este trabajo se describe un nuevo enfoque para generar de manera automática un tesauro de similitud a través de un corpus comparable con el fin de aplicarlo a tareas de recuperación de información multilingüe. Aunque la disponibilidad de recursos lingüísticos es cada vez mayor, todavía hoy en día es difícil el acceso a algunos de ellos, sobre todo en ámbitos multilingües. Incluso, la propia complejidad de la tarea CLIR requiere el uso conjunto de varios recursos para aumentar la eficacia del sistema. Los corpus comparables son uno de estos recursos multilingües especialmente interesantes por su disponibilidad y por la posibilidad de generarlos automáticamente. Sin embargo, para que sean útiles deben estar alineados al menos a nivel de documento. Para llevar a cabo esta tarea, se han utilizado técnicas de clustering. Una vez que los documentos están alineados, se genera el tesauro de similitud a partir de ellos. Los experimentos realizados muestra que los tesauros de similitud multilingües son una buena alternativa cuando otros recursos más adecuados no están disponibles.
Palabras clave: recuperación de información multilingüe, clustering, alineación de textos, recursos lingüísticos, tesauro.

Abstract: In this work, it is described a new approach to automatically generate a similarity thesaurus through a comparable corpus, with the aim of applying it to Cross Language Information Retrieval. Although the availability of linguistic resources is higher and higher, it is still difficult to have access to some of them, above all on multilingual circles. Even, the complexity itself of the task CLIR requires the global use of several resources to increase the efficiency of the system. The comparable corpus are one of this multilingual resources specially interesting due to its availability and due to its chance to be generated automatically. However, in order to make these corpora useful, they should be aligned at least at document level. In order to carry out this task, clustering techniques have been used. Once the documents are aligned, the similarity thesaurus is generated from them. The accomplished experiments show that the multilingual similarity thesaurus are a good chance when other more suitable resources are not available.
Keywords: Cross Language Information Retrieval, clustering, alignment of texts, linguistic resources, thesaurus.

1 Introducción

A finales de los años 90, la tarea denominada Cross Lingual Information Retrieval (CLIR) ha ido ganando atención dentro de la comunidad IR (Information Retrieval), hasta convertirse en nuestros días en una disciplina a la que se dedica un esfuerzo semejante al que recibe la Recuperación de Información tradicional. Un sistema CLIR básicamente es un sistema IR capacitado para operar sobre una colección de documentos multilingüe. Esto es, supuesto que un usuario consulte un sistema CLIR, éste debe recuperar todos aquellos documentos relevantes de entre los que se encuentran en la colección, con independencia del idioma utilizado tanto en la consulta como en los documentos. Así, la salida de uno de estos

ISSN 1135-5948

sistemas será frecuentemente una lista heterogénea de documentos escritos en inglés, español, francés, alemán... y ordenada según la puntuación obtenida por cada documento para la consulta dada.

El creciente interés en estos sistemas multilingües viene dado básicamente por dos motivos: por una parte, la popularización de Internet ha hecho de la red una enorme colección documental multilingüe, y por otra, en la sociedad de la globalización, organizaciones multinacionales generan ingentes cantidades de documentos, escritos usualmente en los idiomas que son nativos en las diversas regiones donde la organización esté presente. En ambos casos, el usuario tipo del sistema multilingüe será alguien con ciertas nociones del idioma o idiomas presentes en la colección de documentos multilingüe, pero no con la suficiente habilidad como para expresar su necesidad de información mediante una consulta precisa en cada uno de tales idiomas.

Debido a la elevada cantidad de documentos escritos en diversos idiomas, se hace necesario superar la barrera lingüística que surge cuando se intenta buscar información sobre tal colección. Para lograrlo, existen sistemas CLIR que traducen las consultas a los idiomas necesarios, o crean una colección de documentos monolingüe mediante la traducción de la colección original multilingüe, o bien realizan un enfoque mixto, traduciendo las consultas, pero manteniendo un único índice de documentos multilingüe. Si bien, la opción de traducir únicamente las consultas parece que es la predominante actualmente, dificulta la obtención de una única lista de documentos relevantes pues, en general, obtendremos tantas listas como idiomas estén presentes en la colección. La traducción documental, por su parte, presenta problemas de escalabilidad, además de resultar pesada la traducción de toda la colección, especialmente en un ambiente experimental, con frecuentes cambios y la consecuente reindexación de la colección.

Recursos tradicionales para la traducción son las Máquinas de Traducción Automáticas (MT, del inglés *Machine Translation*) tal como SYSTRAN[1] y los diccionarios electrónicos (MRD, del inglés *Machine Readable Dictionary*). Otro recurso muy apreciado para esta tarea son los llamados corpus paralelos. Estos son corpus multilingües cuyos documentos son traducción exacta los unos de los otros. Es equivalente a tener un corpus monolingüe, junto con su traducción a otros idiomas. Además, resulta muy útil alinearlos a nivel de frase, de manera que para cualquier frase es posible conocer su traducción exacta en el resto de los documentos paralelos. Podemos, de esta manera, conocer con qué frecuencia un término es traducción de otro, así como el contexto que suele acompañarle. El problema de los corpus paralelos es que se trata de un recurso difícil de encontrar, sobre todo teniendo en cuenta que para que realmente sea útil, debe tener un tamaño como mínimo de varios miles de documentos para cada idioma considerado. Un sistema exitoso que usa exclusivamente corpus paralelos para la traducción es el descrito en (Nie et al., 1999).

Mucho más fáciles de conseguir, y no carentes de interés como recurso en tareas de traducción, son los corpus comparables. La restricción ahora no es tener un corpus traducido a varios idiomas, sino significativamente más laxa: es suficiente con que los diversos corpus monolingües traten un tema común, pero no tienen por qué ser unos traducción de otros. A partir de un corpus comparable es posible generar un tesauro de similitud, el cual permite medir la similitud de dos términos en función del contexto en que aparecen (Sheridan et al., 1997).

Con independencia del enfoque que sigamos, debemos contar con recursos, técnicas y herramientas que nos apoyen en la traducción, teniendo en cuenta que los objetivos de una traducción CLIR no son exactamente los mismos que persigue la traducción automática tradicional: a un sistema CLIR le interesa especialmente que la traducción no pierda el significado de las palabras originales, aun a costa de perder la estructura sintáctica de la frase, e introducir cierta cantidad de ruido. Por esto, existe una tendencia muy fuerte a no confiar en un único recurso, sino a integrar los disponibles, tales como los sistemas MT, MRDs, bases de datos multilingües, corpus paralelos y corpus comparables.

Describimos un nuevo enfoque de generación automática de un tesauro de similitud a través de un corpus comparable para su aplicación a tareas CLIR. El paso fundamental para la correcta construcción del

[1] http://babelfish.altavista.com

tesauro consiste en que la alineación documento a documento de los corpus en ambos idiomas sea lo más perfecta posible. Para alinear los documentos, en distintos idiomas del corpus comparable, se han utilizado técnicas de clustering.

El artículo está organizado como sigue. En el apartado siguiente nos ocuparemos de describir la tarea, dedicando dos secciones al proceso de alineación del corpus y a la generación automática del tesauro. A continuación, se presentan con detalle los experimentos y los resultados. Acabamos con las conclusiones y el trabajo futuro.

2 *Descripción de la tarea*

El concepto de tesauro de similitud fue introducido en tareas de IR para ampliar la consulta realizada por el usuario con un conjunto de términos similares a aquellos que conforman la consulta (Qiu y Frei, 1993). Un tesauro de similitud es una estructura de datos automáticamente calculada a partir de un corpus, de tal manera que, a partir de tal corpus, obtenemos para cada término, una lista de términos estadísticamente próximos o similares.

En IR tradicional, los documentos se indexan a partir de los términos que estos contienen. Así, dos documentos son más similares cuanto más se parecen sus respectivos términos. Intuitivamente, un tesauro de similitud se construye sin más que cambiar los roles de documentos y términos. Ahora son los términos los que están indexados por los documentos en los que estos aparecen, de tal manera que dos términos cualesquiera son más similares cuanto más se parecen sus índices de documentos. Esa similitud puede medirse, por ejemplo, usando la tradicional fórmula *tf·idf, y* la función coseno normalizada como medida de similitud (Salton y McGill, 1983). Finalmente, el tesauro de similitud lo obtendremos seleccionando para cada término aquellos más próximos en el sentido anteriormente descrito.

Una ventaja de los tesauros de similitud es que pueden obtenerse a partir de los motores de IR tradicionales con gran facilidad, pues las estructuras de datos utilizadas son las mismas que las de un sistema IR. Una explicación detallada de este proceso se encuentra en (Qiu y Frei, 1994)

La generación del tesauro de similitud parte de un corpus. Si el corpus que utilizamos es multilingüe, entonces lo que obtenemos es un tesauro de similitud multilingüe (Sheridan et al., 1997): para un término dado, podemos conseguir una lista de términos similares a él, en otro u otros idiomas. Para que el corpus multilingüe resulte útil, debe pertenecer a una de las tres siguientes categorías:

- Cada documento debe contener términos en varios idiomas.
- El corpus es un corpus paralelo. Un corpus paralelo es aquel cuyos documentos están traducidos a varios idiomas. Estos corpus son relativamente difíciles de conseguir, y su tamaño suele ser bastante limitado, según el conjunto de idiomas para el cual debe construirse el tesauro.
- Si un corpus paralelo no está disponible, aún podemos usar un corpus comparable. La condición para que un corpus sea comparable es menos dura que para considerarlo paralelo. Para que un corpus sea comparable, es suficiente con que esté formado por documentos expresados en más de un idioma, y además sea posible encontrar tuplas de documentos cuya temática sea la misma o muy similar, aún sin ser unos traducción de otros. Esto es, pueden haberse escrito independientemente unos de otros (Peters y Picchi, 1996).

Si bien los corpus comparables son especialmente interesantes por su disponibilidad y por la posibilidad de generarlos automáticamente (Martínez et al., 2001), requieren un paso adicional para que resulten útiles en la elaboración de un tesauro de similitud multilingüe: es necesario alinearlos a nivel de documento. Esto es, debemos conocer, para un documento dado, qué documento o documentos comparten su tema en otros idiomas.

En el modelo tradicional, una vez alineados los documentos, se obtiene un único nuevo corpus, cuyos documentos están formados por la concatenación de cada par de documentos alineados. A partir de aquí, la elaboración de tesauro de similitud se desarrolla como en el caso monolingüe.

Los tesauros de similitud multilingües se han utilizado con éxito en tareas de CLIR (Braschler y Schäuble, 2000), utilizándolos

para la pseudo-traducción de las consultas, consiguiendo resultados ligeramente inferiores a los logrados mediante un sistema de traducción automática. Ya que los tesauros de similitud multilingües pueden ser generados a partir de un corpus comparable, son una buena alternativa para aquellos casos en los que no existen o no se pueden conseguir recursos lingüísticos más escasos, como son MT, MRD...

El trabajo realizado muestra la consecución de un tesauro de similitud multilingüe a partir de un corpus comparable inglés-español, no alineado. En concreto se trata de todas las noticias publicadas durante el año 1994 por los Los Angeles Times por un lado, y la agencia EFE por otro. Ambas colecciones están disponibles para los participantes en el foro CLEF (del inglés, *Cross Language Evaluation Forum*) (Peters, 2000). Partiendo de tal colección de datos, presentamos un nuevo método para la necesaria alineación del corpus y la posterior generación del tesauro de similitud.

2.1 Alineación de corpus

Para generar el tesauro de similitud multilingüe, necesitamos una correspondencia biunívoca entre los documentos de los dos corpus. Esta relación permitirá comparar sus palabras con sólo ordenar adecuadamente los pesos de las palabras, como veremos en el apartado siguiente.

Hemos usado algoritmos de clustering para realizar esta alineación. La idea básica es agrupar los dos corpus en uno sólo y calcular los diferentes clusters, agrupando los documentos del mismo idioma, para cada cluster encontrado. De esta forma, se obtienen dos documentos por cluster, uno por cada idioma y relacionados biunívocamente, es decir, necesitamos convertir el corpus comparable en un corpus comparable alineado por documento.

El problema fundamental se encuentra en que los corpus están escritos en diferentes idiomas y el clustering no es directamente aplicable puesto que ambas lenguas son prácticamente disjuntas. Para resolverlo, se han aplicado dos criterios (Sheridan et al., 1997):

- Filtrado de todos los documentos, reduciéndolos exclusivamente a los nombres propios que contengan, términos que, obviamente, son compartidos por los dos idiomas en un alto grado.
- Ya que la colección a alinear son noticias emitidas por un periódico y una agencia, una restricción útil es considerar candidatos a los documentos con noticias publicadas exactamente el mismo día.

Dado que pretendemos realizar todo el trabajo de manera automática, pero sin el uso recursos externos, hemos tomado como nombres propios, todas aquellas palabras que empiezan por mayúscula. Esto no supone una tara para la resolución del problema, ya que, como los idiomas español e inglés no tienen prácticamente palabras en común, los términos que empiecen por mayúscula que no sean nombres propios (palabras después de un punto, primera palabra de un párrafo, palabras después de puntos suspensivos, etc.) no es probable que aparezcan en los documentos del otro idioma, por lo que no se tendrán en cuenta, si bien disminuirán homogéneamente los pesos de todos los demás términos.

Así pues, después de pasar este filtro, se puede aplicar cualquier técnica de clustering para encontrar las parejas de documentos bilingües. Para elegir el algoritmo adecuado, debemos tener en cuenta que lo importante de la tarea no es encontrar todos los clusters sino que los encontrados sean correctos, esto es, cuando dos documentos se agrupen, los documentos originales deben tratar del mismo asunto.

Para ello, necesitamos un algoritmo de agrupamiento que nos permita determinar cuánto de parecidos son los documentos para ajustar su valor hasta alcanzar resultados aceptables en el proceso de agrupamiento. Además necesitamos que el tiempo requerido para el cálculo sea suficientemente bajo como para poder usar grandes colecciones de textos y encontrar los cluster en un tiempo razonable.

El algoritmo SLINK (Romesburg 1984) se ajusta perfectamente a la tarea especificada. Se calcula la similitud entre todos los documentos del corpus bilingüe, obteniendo una matriz simétrica cuadrada de similitudes. Se trata de agrupar aquellos documentos de mayor similitud, recalculando las similitudes entre el nuevo cluster con el resto de documentos como el máximo de las similitudes entre los documentos agrupados con cada uno de ellos. La condición de parada del algoritmo será obtener un solo cluster, caso improbable, o

alcanzar una similitud por debajo de un umbral establecido.

Este umbral debe ser lo suficientemente alto como para garantizar la corrección de los clusters encontrados. En nuestro caso, con los vectores de los documentos normalizados, se ha escogido un umbral de 0,5 de similitud.

De todos los clusters encontrados, sólo nos interesan aquellos formados por documentos de los dos idiomas, por lo que aún es necesario procesarlos para formar un nuevo corpus bilingüe alineado documento a documento. Para cada cluster con documentos en ambos idiomas, generamos dos ficheros nuevos, uno con la unión de los documentos en español y el otro con los de inglés. Estos dos ficheros generados, forman una pareja de documentos alineados del nuevo corpus comparable.

En el proceso realizado, hemos obtenido un total de 1.061 documentos para cada idioma. Este número puede incrementarse o disminuirse cambiando la similitud umbral, permitiendo uniones entre documentos con mayor o menor parecido. Aunque el uso de umbrales menores pudiera parecer contraproducente, debemos resaltar que aún nos queda un segundo grado de libertad en el proceso de generación de los tesauros de similitudes, por lo que podríamos disminuir este umbral, a costa de poner más restricciones en la elección de las palabras del tesauro. En nuestros experimentos, el número de documentos generados se ha mostrado satisfactorio, por lo que un umbral de similitud de 0,5 es adecuado.

2.2 Generación del tesauro

La generación del tesauro de similitud toma como partida el corpus comparable bilingüe generado en el apartado anterior. Para nuestros experimentos necesitamos calcular, para una palabra dada en español, su correspondiente tesauro expresado con términos en inglés. En el modelo introducido por Sheridan, la creación del tesauro de similitud multilingüe a partir de un corpus comparable alineado a nivel de documento requiere aunar cada par de documentos alineados en uno solo, obteniendo así un nuevo corpus único. En este trabajo se ha optado por otro método: se mantienen los corpus alineados por separado, se calcula el peso de cada término en cada documento siguiendo el esquema *tf·idf*, y, finalmente, se buscan aquellos términos cuyos vectores de pesos son más similares en el otro idioma. A diferencia del modelo original, la similitud entre los términos no la medimos en función de cómo éstos son indexados por los documentos, sino cómo los documentos son indexados por los términos: dos términos son similares si tienden a indexar los documentos de forma parecida (Rijsbergen 1979).

La similitud entre palabras puede calcularse de una manera adecuada a partir de una interpretación inversa de los pesos de las palabras. En definitiva se trata de decidir si dos palabras son similares en función de los documentos en los que aparecen. Cuando dos palabras figuran en los mismos documentos, deducimos que son usadas en el mismo contexto, por lo que es posible considerarlas en el mismo tesauro de similitud.

Podemos aplicar esta misma idea en nuestros experimentos, ya que, como los documentos están alineados uno a uno, podemos concluir que si una palabra en español aparece en una serie de documentos emparejados con un número igual de documentos en los que aparece otra palabra en inglés, ambos términos están en los mismos contextos y podrían formar parte del mismo tesauro de similitud.

Así pues, debemos construir los ficheros inversos correspondientes al corpus comparable obtenido, uno para los documentos en español y otro para los de inglés. Usaremos el Modelo de Espacio Vectorial (MEV) para indexar las palabras, donde cada término de un idioma dado tendrá un vector asociado y normalizado con los tradicionales pesos *tf·idf* de la palabra con respecto a cada uno de los documentos de la colección del mismo idioma.

A la colección de partida, se le hace un filtrado estándar de *stopping* y *stemming* reduciendo la dimensión del espacio vectorial y mejorando la solución. Después de preprocesar la colección, se calculan los pesos de todas las palabras de los dos idiomas generando dos ficheros, uno para cada lengua, que son el punto de partida del cálculo de los tesauros de similitud.

Hemos optado por calcular tesauros de 100 pares término–similitud, para poder parametrizar en la evaluación los valores mínimos de similitud entre palabras. En principio, podemos optar por elegir las k mejores palabras o bien, aquellas que superen cierto umbral de similitud. La segunda solución es más versátil, ya que permite, en

conjunción con el umbral propuesto en el cálculo de los clusters, flexibilizar el cálculo sin descartar ninguna palabra por un error en el proceso de agrupamiento. De hecho, el algoritmo usado podría haber sido cualquier otro, ya que lo que estamos calculando son los mejores clusters y no todos. El algoritmo SLINK destaca por su sencillez y eficaz implementación y el umbral elegido asegura que los clusters obtenidos son correctos.

3 *Descripción del entorno experimental*

Para probar el algoritmo resultante se ha usado el corpus correspondiente al anuario de 1994 de Los Angeles Times (LAT), formado por más de 100.000 documentos, y un conjunto de 40 consultas en español e inglés junto con los correspondientes juicios de relevancia, pertenecientes a las jornadas CLEF del año 2000. Para simular un ambiente multilingüe, el conjunto de consultas en español debe ser traducido al inglés, lanzando las consultas así traducidas contra la colección LAT. Para medir la bondad de nuestro tesauro de similitud, hemos realizado tres experimentos:

- Realizar la traducción utilizando el sistema de traducción automática SYSTRAN.
- Traducir palabra a palabra utilizando EuroWordNet (Vossen, 1999).
- "Traducir" cada palabra usando el tesauro de similitud.

Es importante remarcar que el objetivo de nuestros experimentos no es obtener mejores resultados que los obtenidos a través de una MT o un MRD, sino conseguir un método alternativo a estos, con un rendimiento similar, y mucho más fácil de conseguir.

Debido al pequeño tamaño del tesauro de similitud multilingüe que hemos generado, ciertos términos en español no han podido ser traducidos al inglés. Con la finalidad de medir la repercusión de esta incidencia, hemos usado dos conjuntos de consultas: uno formado por las 40 consultas originales, conjunto de consultas I, y el otro en el que sólo aparecen aquellas consultas que no contienen términos sin traducir, conjunto de consultas II (tabla 1). En total son 13 palabras no vacías en 13 consultas distintas las que quedan sin traducción.

En la tabla 2 se muestran los resultados obtenidos, demostrando que los tesauros de similitud multilingües son una buena alternativa cuando otros recursos más adecuados no están disponibles, siempre que tengan una cobertura lo suficientemente amplia. Si bien, la mejor precisión es la alcanzada con SYSTRAN, éste es un recurso escaso, y cuyo rendimiento es fuertemente dependiente del par de idiomas involucrados en la traducción. Por su parte, el rendimiento obtenido con EuroWordNet es de un orden similar al alcanzado con el tesauro de similitud multilingüe.

	Consultas	**Términos**
Conjunto de Consultas I	40	82
Conjunto de Consultas II	27	69

Tabla 1: Conjuntos de consultas utilizados

	I	**II**
SYSTRAN	0,25	0,26
EuroWordNet	0,17	0,19
Tesauro de Similitud Multilingüe	0,10	0,15

Tabla 2: Precisión media obtenida con el conjunto de consultas I y II

En la tabla 1 se observa que son 13 términos sobre un total de 82 los que quedan sin traducir, lo que supone que la cobertura del tesauro roza el 90%. A pesar de que es una cobertura bastante alta, el hecho de que ciertos términos queden sin traducción explica que la diferencia de precisión entre las colecciones de consultas I y II sea del 5% para el caso del tesauro de similitud multilingüe, mientras que utilizando otros recursos (MT, MRD), la diferencia entre ambas colecciones es poco relevante: 1% y 2% para MT y MRD respectivamente. También resulta interesante observar cómo la precisión obtenida por el tesauro de similitud es de tan sólo un 4% por debajo de la que alcanzamos con EuroWordNet, siempre que las consultas no

contengan términos sin traducir. Este dato nos anima a confiar en que un corpus comparable lo suficientemente amplio permitiría incluso superar el rendimiento obtenido mediante un MRD.

Los datos expuestos en la tabla 2 para el caso de tesauro de similitud multilingüe son los obtenidos sustituyendo cada término en español por aquellos cuyo coeficiente de similitud es superior o igual a 0,5. En la tabla 3 se observa cómo varía la precisión media alcanzada sobre la colección de consultas II para varios valores de corte.

Similitud	Precisión
0,1	0,091
0,2	0,120
0,3	0,138
0,4	0,142
0,5	0,153
0,6	0,142
0,7	0,072
0,8	0,062
0,9	0,047

Tabla 3: Coeficiente de Similitud-Precisión

4. *Conclusiones y futuros trabajos*

En este artículo hemos mostrado un nuevo método para la alineación de corpus comparables a nivel de documento, fundamentado en el algoritmo SLINK y una interpretación alternativa de los tesauros de similitud, basada en cómo los documentos son indexados por los términos, y no al revés.

Uno de los aspectos más interesantes de los tesauros de similitud es la posibilidad de generarlos a partir de un corpus comparable. Aunque en este trabajo se ha utilizado un corpus comparable disponible (LAT y EFE), nuestro siguiente objetivo será crear una colección de entrenamiento de mayor cobertura, a partir de la bastísima fuente documental disponible en Internet, e intentar escalar el sistema a las lenguas de la Comunidad Europea, creando así un sistema CLIR que no se apoye en recursos escasos tales como MT, MRD, corpus paralelos...

Otro aspecto a explorar es la conveniencia de la unidad de indexación utilizada. Tradicionalmente, un tesauro de similitud mide la relación entre términos en función de los documentos que lo contienen. Pero quizás ésta sea una unidad excesivamente grande. El uso de párrafos o incluso frases, esperamos que aporte una mayor precisión. Sin embargo, si partimos de un corpus comparable, es más complejo alinear el corpus a nivel de párrafo o de frase que a nivel de documento, Se requieren, pues, técnicas adicionales para acometer tal tarea.

Por último, la posibilidad de aplicar otros algoritmos de clustering para encontrar el mejor alineamiento entre documentos también será fruto de posteriores investigaciones.

Bibliografía

Braschler, M., Schäuble, P. Experiments with the Eurospider Retrieval System for CLEF 2000. En Proceedings of CLEF 2000. Lecture Notes in Computer Science, Springer Verlag, 2001.

Braschler, M., Ripplinger, B. Schäuble, P. Experiments with the Eurospider Retrieval System for CLEF 2001. Proceedings of CLEF 2001 Cross-Language Text Retrieval System Evaluation Campaign, 2001.

Martínez, F., Ureña, L.A., García, M. WWW como fuente de recursos lingüísticos para su uso en PLN. En Procesamiento del Lenguaje Natural, número 27, páginas 141-147, 2001.

Nie, J.-Y., Simard M., Isabelle P. y Durand R. Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts in the Web. Proceedings of SIGIR '99, Berkeley, CA, 1999.

Peters, C. (Ed.). Cross-Language Information Retrieval and Evaluation Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, 2000.

Qiu, Y., Frei, H. Concept Based Query Expansion. Proceedings of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA, pages 160 - 169, 1993.

Qiu, Y., Frei, H. Improving the Retrieval Effectiveness by a Similarity Thesaurus.

Technical Report 225, ETH Zurich, Department of Computer Science,1994.

Romesburg, H.C. Cluster Analisys for Researchers; Lifetime Learning Publications, California, 1984.

Van Rijsbergen, C.J. Information Retrieval, Second ed., London: Butterworths, 1979.

Salton, G. y McGill. Introduction to Modern Information Retrieval. McGraw-Hill, New York, 1983.

Sheridan, P., Braschler, M., Schäuble, P. Cross-language information retrieval in a multilingual legal domain. Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries, pages 253 - 268, 1997.

Vossen, P. EuroWordNet: A Multilingual Database for Information Retrieval. Third Delos Wokshop Cross-Language Information Retrieval, pp. 85-94. European Research Consortium For Informatics and Mathematics, 1997.