Propuesta para un Sistema de Recuperación de Información Multilingüe Independiente del Lenguaje

Fernando Martínez Santiago, L. Alfonso Ureña López
{dofer, laurena}@ujaen.es
Grupo Sistemas Inteligentes. Departamento de Informática. Universidad de Jaén. Spain

Presentamos en este artículo los resultados obtenidos hasta la actualidad en el desarrollo de un modelo de Recuperación de Información Multilingüe (CLIR). Para ello, se estudian con cierto detalle los problemas que surgen ante la necesidad de superar la barrera lingüística, existente en sistemas de Recuperación de Información que deben tratar con colecciones multilingües. Las diversas soluciones propuestas para tales problemas tienen como factor común que son, hasta donde nos es posible, independientes de recursos lingüisticos disponibles para idiomas concretos. En esa línea, se apuesta por el Web como una fuente abundate de recursos útiles en tareas CLIR.

1 Introducción

A partir del segundo lustro de los años 90, la tarea denominada Cross Lingual Information Retrieval (de aquí en adelante, CLIR) ha ido ganando atención dentro de la comunidad RI, 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 RI 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 sistemas será frecuentemente una lista heterogénea de documentos escritos en inglés, español, francés, alemán, etc., y ordenada según la puntuación obtenida por cada documento para la consulta dada.

Si en la tarea RI se trata de seleccionar aquellos documentos relevantes para una determinada necesidad de información del usuario, en un escenario multilingüe es necesario, además, superar la barrera lingüística que surge entre el idioma de la consulta y los diversos idiomas presentes en la colección que se desea consultar (Oard, 1997). Así, cualquier intento serio de desarrollar un sistema CLIR capaz de obtener unos resultados equiparables a los obtenidos en un ambiente monolingüe, deberá preocuparse por los siguientes aspectos, a los que nos referiremos de aquí en adelante como los tres problemas CLIR (Grefenstette 1998):

i. Traducción de las consultas y/o documentos.
ii. Una vez realizada la traducción, es usual que ésta no sea única. En tal caso, ¿cuál elegir?.
iii. ¿Cómo obtener una lista única de documentos relevantes, con independencia del lenguaje utilizado en cada documento? .

Nótese que los dos primeros puntos, cómo traducir y desechar las traducciones menos precisas, son problemas típicos de los sistemas de Traducción Automática (MT, de aquí en adelante). Sin embargo, un sistema CLIR es menos exigente que un sistema MT en cuanto a la calidad de la traducción. Esto es debido a que empíricamente se ha comprobado que, mientras un sistema MT consigue sus mejores resultados cuando toma como unidad a la frase, los sistemas RI, en el estado del arte actual, parecen comportarse mejor si toman como unidad la palabra, manteniéndose poco o ninguna ligazón sintáctica entre los términos. Por lo tanto un sistema CLIR centra todos los esfuerzos en obtener un conjunto de posibles traducciones lo más preciso posible para cada una de las palabras.

Para cada uno de los tres problemas CLIR se han realizado muchas y variadas propuestas, si bien usualmente son fuertemente dependientes de los recursos lingüísticos disponibles. Por ejemplo, podemos usar un sistema MT para realizar la traducción de las consultas, pero ¿existen MT para cualquier par de idiomas?. Las soluciones aquí propuestas tienen todas un factor común: la escalabilidad a prácticamente cualquier idioma. Y la única manera de conseguir este objetivo es mediante el uso de recursos y métodos tan independientes del lenguaje como sea posible.

El resto del artículo está organizado como sigue. En el apartado siguiente se muestra un modelo formal de un sistema de recuperación de información. Posteriormente se realiza un breve estudio de cada uno de los tres problemas CLIR, así como propuestas que tan independientes de un determinado idioma como sea posible.

2 Descripción formal de un sistema RI tradicional

En este apartado se describe un formalismo que permite describir la mayoría de los sistemas de RI actuales. Un sistema de RI lo podemos representar mediante la estructura (Sheridan 1997):

$<T, \Phi, D, ff, df>$

donde:

D es la colección de documentos a indexar.

Φ es el vocabulario utilizado en los índices generados a partir de D.

T es el conjunto de todos los tokens τ presentes en la colección D, usualmente las palabras o términos. Así, la función:

$\varphi: T \rightarrow \Phi,\ \tau \rightarrow \varphi(\tau)$

establece la relación entre los tokens presentes en T y su correspondiente entrada en el vocabulario de indexación Φ. Esto es, la función φ puede representar el proceso de extracción de raíces (stemming), lematización o cualquier otro que haga corresponder a cada token presente en los documentos con un elemento en el índice de documentos.

ff es la función de frecuencia de un determinado elemento índice φ_i en un documento d_j:

$ff(\varphi_i, d_j) := |\{ \tau \in T | \varphi(\tau) = \varphi_i \wedge d(\tau) = d_j \}|$

donde d es la función que hace corresponder a cada token τ con el documento que le contiene:

$d: T \rightarrow D,\ \tau \rightarrow \varphi(\tau)$

finalmente, df denota la frecuencia documental: el número de documentos que contiene un elemento índice dado:

$df(\varphi_i) := |\{ d_j \in D | \exists \tau \in T: \varphi(\tau) = \varphi_i \wedge d(\tau) = d_j \}|$

3 El primer problema CLIR: la traducción de la consulta

3.1 Enfoques tradicionales

Usualmente, un sistema CLIR debe acometer la traducción de la consulta del usuario, de los documentos, o de ambos. Así, un enfoque inmediato es el uso de una MT tal como Systran[1] o T1-Translator[2]. Los sistemas basados exclusivamente en MT (Gachot 1998) han conseguido un buen rendimiento, pero presentan ciertos inconvenientes, como el ya mencionado problema de disponibilidad para idiomas no muy extendidos. Además, aún existiendo el sistema MT, la calidad de los resultados es fuertemente dependiente del idioma origen y destino de la traducción (Savoy 2001).

Ya que la sintaxis de la traducción es algo secundario, otro enfoque válido es la traducción de la consulta, palabra a palabra, usando un diccionario electrónico (MRD, de Machine Readable Dictionary) (Ballesteros 1996, Adriani 2000, Martínez 2001c). Sin embargo, si mantenemos todas las posibles traducciones para cada término, se obtiene una traducción excesivamente ruidosa, que daña sensiblemente la precisión en la búsqueda.

El uso combinado de MT y MRD obtiene resultados muy notables: se realiza la traducción con una MT, y esta se amplía con las dos o tres traducciones más frecuentes de cada término (Savoy 2001). Sin embargo, tanto las MT como, en menor medida, los MRD, son un recurso relativamente escaso si consideramos idiomas con poca difusión en comparación con el inglés, francés, alemán o español.

Un recurso alternativo, muy apreciado en CLIR, son los corpus paralelos. En un corpus paralelo, cada documento posee una traducción al resto de los idiomas presentes en el corpus. Si conseguimos alinear el corpus a nivel de frase (dada una frase, conocemos como es traducida en el resto de los idiomas), es posible realizar traducciones término a término (Kraaij 2001), así como derivar las probabilidades de traducción (Hiemstra et al 2000). Lamentablemente, aunque existen herramientas que exploran el web en busca de estos corpus (Nie et al. 1999), no es fácil obtener corpus paralelos para según qué idiomas, y menos aún con la cobertura lo suficientemente amplia como para obtener datos fiables.

---

[1] Systran disponible en la web: babelfish.altavista.com

[2] T1-translator disponible en la web: http://t1-testdrive.sail-labs.de/index.html

3.2 Corpus comparables y Tesauros de Similitud Multilingües.

Un recurso mucho más abundante son los corpus comparables. Estos corpus, a diferencia de los paralelos, no exigen que unos documentos sean traducción de otros. Basta con que, dado un documento, existan textos en los otros idiomas que traten el mismo tema. Un buen ejemplo de estos documentos, son las noticias internacionales publicadas en los diarios. Si bien cada una de estas noticias se ha escrito independientemente de las demás, es fácil conseguir conjuntos de ellas referentes al mismo evento. El uso de los corpus comparables no es tan inmediato como el de los paralelos, pero si somos capaces de crear un corpus comparable alineado a nivel de documento, es posible derivar a partir de él los llamados Tesauros de Similitud Multilingüe (TSM).

Un tesauro de similitud (Qiu, 1995) es una estructura de datos en la cual obtenemos para un término dado, términos similares o semánticamente próximos a él. Informalmente, la medida de similitud entre dos términos la obtenemos a partir de cómo dos términos son indexados por los documentos. Esto es, tradicionalmente en RI medimos las similitud entre documentos a partir de cómo son indexados por sus términos. En un tesauro de similitud se intercambian los papeles de términos y documentos. Un poco más formalmente, usando la notación introducida en el apartado anterior, un sistema de R.I. lo notamos:

<T, Φ, D, ff, df>

entonces un tesauro de similitud lo podríamos representar como una estructura dual a la anterior:

<T, Φ', D', ff', df'>

donde $\Phi' := D$, son los elementos índice (ahora los documentos), y $D' := \Phi$, son los elementos a recuperar (los términos), y las funciones ff' y df' las definimos como:

$ff'(d_j, \varphi_i) := |\{\tau \in T | d(\tau)=d_j \wedge \varphi(\tau)=\varphi_i\}| = ff(\varphi_i, d_j)$

$df'(\varphi_i) := |\{\varphi_i \in \Phi | \exists \tau \in T : d(\tau)=d_j \wedge \varphi(\tau)=\varphi_i\}| \approx$ longitud de d_j

Si ahora aplicamos sobre este espacio dual algún método concreto de R.I, lo que obtenemos es un tesauro de similitud

Si el corpus de que partimos es multilingüe, entonces lo que obtenemos es un TSM (Sheridan, 1997): dado un término en un idioma, podemos obtener términos similares en otros idiomas. Esto permite realizar seudo-traducciones de las consultas en nuestro sistema CLIR. Este enfoque ya ha sido utilizado con éxito en tareas CLEF partiendo de un corpus comparable lo suficientemente amplio (Braschler et al, 2001).

3.3 Una propuesta para la creación de TSM a partir de la Web

La Web es una inacabable fuente potencial para la creación de corpus. Pero dada la poca estructuración de la Red, ¿cómo extraer la información relevante de la Web?, ¿cómo conocer qué documentos tratan el mismo tema?.

La extracción de información estructurada de la Red (Grefenstette 1998, Pierre 2001) es una tarea difícil de automatizar, pues necesita la compresión de la máquina del documento HTML, el cual está diseñado para ser leído por las personas, no por las máquinas (Bernes-Lee, 1998). Sin embargo, si es posible establecer un puente entre el estilo del documento HTML y la compresión de la máquina, siempre que tal estilo "delate" de alguna manera la información allí almacenada. Esto se puede expresar a través de reglas sencillas como "El elemento <TITLE> es el título del documento", o algunas más complejas como "el texto con estilo de letra arial 12pt, que aparece después del elemento <META content="text"> , se corresponde con el autor

Regla aplicable	***Documento HTML***	***Documento normalizado***
<translate> <tag_in> texto </tag_in> <tag_out> CUERPO <tag_out> </translate>	 <texto>Este texto <B>se mantiene</B> </texto>pero éste otro <B>se desecha</B> <texto>y éste de aquí se añade</texto>	<CUERPO> Este texto se mantiene y éste de aquí se añade </CUERPO>
<translate> <tag_in closed="false" attr_name="NAME" attr_value="TITULO" attr_get="CONTENT"> META </tag_in> <tag_out> TITULO </tag_out> </translate>	<META name= "TITULO" content="Finaliza la cumbre europea"> <META name="FECHA" content=" 12/11/2000">	<TITULO> Finaliza la cumbre europea </TITULO>

Tabla 1 - ejemplo de reglas WebReader.

del artículo". Evidentemente, esta descripción del texto HTML depende de cada sitio, y debe ser realizada por un experto humano. En esta línea, hemos propuesto una herramienta, que llamamos WebReader (Martínez 2001a) , que genera un corpus con un buen nivel de estructuración, y poco ruido, a partir de un documento de especificación que describe los sitios web a partir de los cuales extraer el corpus (Tabla 1).

Método de traducción	AvgP
SysTran	0.26
EuroWordNet	0.19
Tesauro de Similitud Multilingüe	0.15

Tabla 2 – Precisiones obtenidas

WebReader ha sido utilizada con éxito para generación de un corpus comparable inglés/español formado por unos 150.000 documentos provenientes de las ediciones en línea de diversos diarios norteamericanos, ingleses y españoles. Tal corpus ha sido alineado a nivel de documento, para así realizar de un TSM, que ha sido aplicado en tareas CLIR para la seudo-traducción de consultas. El rendimiento obtenido es similar al alcanzado mediante un MRD. En la Tabla 2 se muestra la precisión obtenida sobre una colección de más de 100.000 documentos en inglés (colección Los Angeles Times 1994), a partir de un juego de 40 consultas en español.

3.4 La importancia de las multipalabras en la traducción

Un problema presente en la traducción de consultas basada en MRD o TSM es que es una traducción palabra a palabra, por lo que es frecuente que las llamadas multi-palabras sean mal traducidas. Una multipalabra es una sucesión de palabras cuyo sentido no es igual a la suma de los sentidos de las palabras que la conforman. Tal es el caso de "Casa Blanca", "Unión Europea" o "estado de sitio". La mala traducción de una multipalabra reduce la

AvgP. sin detección de multipalabras	AvgP. Con detección de multipalabras
0.375	0.390

Tabla 3 – Detección de multipalabras con una Red Neuronal

Consula	AvgP. Original	AvgP. Multi-palabras	Multi-palabras detectadas
#7	0.396	0.445	"world soccer"
#9	0. 102	0.202	"war ii" "ii war" "war rwanda" "world war"
#3	0.391	0.322	"decisions made", "hard soft"
#32	0.412	0.251	"women priest", "change direction"

Tabla 4 –Precisiones obtenidas

precisión de un sistema CLIR en torno al 40% (Hull y Grefestette 1996). Existen diccionarios que son muy ricos en estas multipalabras, pero en pro de la independencia de recursos lingüístos, estamos desarrollando un método basado exclusivamente en los corpus generados con WebReader. Tal método consta de dos fases: una primera fase de detección de multipalabras en el idioma en el que se realiza la consulta, y una segunda fase de traducción de tales multipalabras.

La detección de multipalabras dista de ser un problema trivial, y generalmente requiere la integración de varias fuentes de información (estadística, sintáctica y semántica) (Maynard y Ananiadou 2000). Nosotros proponemos el uso de estructuras de representación de información tales como las redes neuronales y redes bayesianas para la integración de recursos de origen principalmente estadístico (coocurrencia de términos, similitud de términos...) (Martínez et al. 2002a), pues tal información es fácilmente extraíble de un corpus tal como el descrito en el anterior punto. Los resultados que hemos obtenido muestran que si bien la detección de multipalabras pueden duplicar la precisión de una consulta dada, también puede reducir a la mitad tal precisión, si se marca como multipalabra una expresión que realmente no lo es (Tabla 3 y 4). Es por esto que nuestros esfuerzos se encaminan a la mejora de la precisión en la detección de las multipalabras, aun en detrimento de la cobertura.

Una vez marcadas las multipalabras, el siguiente paso es su traducción. Tradicionalmente, la traducción de una multipalabras está basada en la traducción por separado de cada palabra (con un MRD o un

TSM) de que consta la multipalabra, para luego, entre todas las posibles combinaciones de traducciones, escoger aquella que es más probable según algún criterio estadístico (coocurrencia de los términos traducidos, similitud de pesos de los términos, etc) (Adriani 2000, Ballesteros y Croft 1997). Realmente, este problema de la traducción de las multipalabras es muy similar al de su detección, pero sobre un espacio mucho más reducido. Si en el caso de la detección, debemos encontrar multipalabras sobre un corpus, ahora se trata de encontrar cual es la mejor candidata a multipalabra de entre todas las posibles traducciones de una expresión dada. Es por lo tanto de nuevo aplicable el mismo esquema basado en la integración de recursos mediante una red neuronal o una red bayesiana.

4 El segundo problema CLIR: filtrado de traducciones

Una vez realizada la traducción, ¿qué traducciones mantener y cuales no?. Una posible solución es calcular las probabilidades de traducción de un término por otro, de tal manera que podemos eliminar aquellas traducciones menos probables(Hiemstra, 2000). En esta línea, nosotros hemos propuesto un método (Martínez et al., 2001b, d) para filtrar traducciones del español al inglés combinando EuroWordNet (Vossen 1998) y SemCor. A continuación se expone brevemente cual es el método seguido, y sus limitaciones.

EuroWordNet permite conocer la traducción de un término a otros idiomas, para cada sentido WordNet del término original. Por otra parte SemCor es un subconjunto del Brown Corpus, donde cada término está manualmente etiquetado con su sentido WodNet. Con esta información, la forma de proceder es la siguiente:

Traducir el término en español E por I_1, $I'_1,...I^n_1$, I_2, $I''_2..I^m_2$, usando EuroWordNet, de tal manera que conocemos que cualquier I^j_1 es traducción de E con el sentido 1, I^k_2 es traducción de E con el sentido 2, y así siguiendo.

Para cada término en inglés I^j_k, calcular la probabilidad de que actúe con el sentido k, pues esa será la probabilidad de que sea traducción de E

Retener aquellas traducciones de E más probables.

Traducción	AvgP
EuroWordNet	0,1701
EuroWordNet+filtrado	0,1941

Tabla 4 - Uso de Prob. de Traducción

La idea subyacente es que si un término es traducción de otro, lo es porque comparten un determinado significado en alguna de sus acepciones. Cuanto más común sea esa acepción en I, más probable es que sea una traducción correcta del E. Este enfoque presenta un limitación evidente: sólo estamos evaluando la probabilidad de que I_k actúe con cierto sentido k, pero desconocemos si entre todas las acepciones de E, la que presenta en un determinado contexto es justamente k, y no otra. Esto es, el siguiente paso sería desambiguar el término original(Ureña 2001), obteniendo así que la traducción correcta de E son aquellos términos I_j, $I'_j,...I^n_j$, pues es j el sentido con el actúa E. Luego, usando el algoritmo antes propuesto, del conjunto I_j, $I'_j,...I^n_j$ mantendremos sólo aquellos en los cuales el sentido j es más usual en ellos, pues sólo en ese caso son traducción de E.

Todo este proceso, en el estado en que se encuentra actualmente, es fuertemente dependiente de EuroWordNet y SemCor. Sin embargo, el método subyacente no depende de un recurso concreto. Nuestros pasos actualmente se encaminan a la creación de un desambiguador independiente del idioma, y a la obtención de las probabilidades de traducción a través de un TSM.

5 El tercer problema CLIR: Cálculo del RSV en dos pasos

Un enfoque usual en CLIR es traducir la consulta a cada idioma presente en el corpus, para a continuación ejecutar diversas ejecuciones monolingües, una por idioma. Finalmente, es necesario obtener un único ranking de documentos, mezcla de los obtenidos por separado. Pero, ¿cómo realizar tal mezcla?. Este es un problema que dista de resultar trivial, pues la puntuación alcanzada por cada documento (RSV, del inglés *Retrieval Status Value*) son calculados no sólo en función de la idoneidad del documento y el modelo RI seguido, sino que también es determinante el resto del corpus monolingüe al cual pertenece tal documento. Existen diversos enfoques de "normalización" de los RSV (Powell et al. 2000), pero aún así se genera

una pérdida grande de precisión en el proceso (según la colección, entre el 20 y el 40%) (Savoy 2001, Voorhees 1995), y siendo quizás por esto que los sistemas CLIR basados en traducción de documentos, suelen conseguir resultados sensiblemente mejores que aquellos que tan sólo traducen la consulta.

El planteamiento propuesto calcula el RSV en dos fases, preselección y reordenamiento, y está orientado a sistemas basado en traducción de consultas, con independencia de la técnica usada en la traducción.

I. La fase de preselección de documentos se corresponde con la traducción y lanzamiento de la consulta sobre cada colección monolingüe, D_i, como es usual en los sistemas CLIR basados en la traducción de consultas. Esta fase produce dos resultados:
   a. resultado de unir todos los documentos recuperados para cada idioma, obtenemos una única colección multilingüe de documentos preseleccionados (colección D').
   b. Resultado del proceso de traducción, obtenemos para cada término de la consulta original, su traducción al resto de los idiomas. Al conjunto de términos que son unos traducciones de los otros, lo llamamos concepto. Un concepto es independiente del idioma. Así, obtenemos un vocabulario Φ, formado por todos los conceptos presentes en la consulta.
II. La segunda fase consiste en reindexar la colección D', considerando el vocabulario Φ. Creamos un índice de conceptos, no de términos, ya que todos los términos pertenecientes a un mismo concepto se tratan como ocurrencias del mismo concepto. Así, si por ejemplo en la consulta aparece el término "casa", es traducido por "house", "casa" ocurre un total de 100 veces en los documentos recuperados, y "house" 150, entonces, la frecuencia del término sería 250. A efectos prácticos, en esta segunda fase cada ocurrencia de "casa", se trata exactamente igual que cada ocurrencia de "house", sobre la colección D'.

Por último, lanzamos la consulta sobre el índice creado en II, consulta que estará formada por conceptos, no por términos, con lo que es independiente del lenguaje.

Un poco más formalmente, el método podría describirse como sigue: para cada colección monolingüe partimos de ya conocida estructura:

$$<T_i, \Phi_i, D_i, ff, df>, 1<=i<=N$$

donde N es el número de idiomas presentes en la colección multilingüe a indexar.

Sea el conjunto $Q = \{Q_i, 1<=i<=N\}$, una consulta junto con sus traducciones, de tal forma que Q_i es la consulta expresada en el mismo idioma que la colección D_i.

Tras haber lanzado cada traducción Q_i, contra su correspondiente estructura $<T_i, \Phi_i, D_i, ff, df>$, es posible obtener un nueva y única estructura:

$$<T', \Phi', D', ff, df>$$

donde:

- D' es el conjunto multilingüe de documentos recuperados como consecuencia de lanzar la consulta Q.
- Φ' es un nuevo vocabulario de indexación, calculado a partir de los conceptos que aparecen en Q. Cada elemento índice $\varphi_j \in \Phi$' es el conjunto formado de la siguiente manera:

  $$\varphi_j := \{\varphi(\tau_{ij}), 1<=i<=N\}, 1<=j<=M$$

  Puesto que cada consulta Q_i es traducción de las demás, es posible alinear las consultas a nivel de término. Sea

  $$\tau_j := \{\tau_{ij} \in Q_i, 1<=i<=N\}, \ 1<=j=M, \ M=|Q|$$

  dónde τ_{ij} es el término j-ésimo de la consulta Q, traducido al idioma i. Así, τ_j representa el concepto j de la consulta Q, con independencia del lenguaje, y φ_j el elemento índice que se deriva de él.
- T' es el conjunto de conceptos τ_j , y representa el vocabulario de D'.
- ff y df se interpretan como es usual.

En cierta forma, este método toma algo de los sistemas CLIR que traducen el corpus, pero en vez de traducir el corpus completo, tan sólo traduce las palabras que aparecen en la consulta, y sobre el juego de documentos recuperado, no sobre el corpus completo. Estas dos simplificaciones permiten plantear el sistema en tiempo de consulta: el reindexado necesario en la segunda fase es factible en términos computacionales por el pequeño tamaño de la colección D', y el, en general, escaso vocabulario de indexación Φ'

(aproximadamente los términos de la consulta Q por el número de idiomas presentes en D').

Algunas consideraciones sobre el modelo propuesto son las siguientes:

- Es fácilmente escalable a varios idiomas.
- El sistema requiere alinear la consulta original y sus traducciones a nivel de término. Este proceso, dependiendo del enfoque seguido para la traducción, puede resultar más o menos trivial.
- En el modelo expuesto, un término junto con su traducción se tratan exactamente de la misma manera. Así, la frecuencia del concepto j-ésimo de la consulta será

$$ff(\varphi_j,d_k) := \Sigma_i ff(\varphi_{ij},d_k),\ \forall\varphi_{ij}\in\varphi_j,\ 1<=i<=N$$

donde

$$ff(\varphi_{ij},d_k) := |\{\ \tau_{ij}\in T|\varphi(\tau_{ij})=\varphi_{ij}\wedge d(\tau_{ij})=d_j\}|$$

Esto no siempre es lo más adecuado, puesto que es usual no pesar de la misma manera el término original y el/los traducido/s. Por ejemplo, puede ocurrir que para un idioma i dado, mantengamos más de una traducción para un determinado concepto de la consulta original. En consecuencia, la frecuencia de ese concepto se verá incrementada artificialmente en aquellos documentos expresados en el idioma i. En estos casos puede resultar interesante dividir la frecuencia de cada término en un determinado idioma por el número de traducciones mantenidas para ese concepto en ese idioma. Esto lo podemos modelizar como sigue:

$$ff(\varphi_j,d_k) := \Sigma_i ff'(\varphi_{ij},d_k),\ \forall\varphi_{ij}\in\varphi_j,\ 1<=i<=N$$

$$ff'(\varphi_{ij},dk) := ff(\varphi_{ij},dk)\ *\ w(\tau_{ij})$$

$ff(\varphi_{ij},d_k)$ representa, como es usual, la frecuencia de la traducción i-ésima del concepto j. Entonces $w(\tau_{ij})$ representa nuestra confianza en que τ_{ij} sea efectivamente traducción del concepto τ_j. Así, para los términos de la consulta original, sin traducir, valdrá 1.

En cierta forma, este método toma algo de los sistemas CLIR que traducen el corpus, pero en vez de traducir el corpus completo, tan sólo traduce las palabras que aparecen en la consulta, y sobre el juego de documentos recuperado, no sobre el corpus completo. Estas dos simplificaciones permiten plantear el sistema en tiempo de consulta: el reindexado necesario en la segunda fase es factible en términos computacionales por el pequeño tamaño de la colección D', y el, en general, escaso vocabulario de indexación Φ' (aproximadamente los términos de la consulta Q por el número de idiomas presentes en D').

El modelo aquí propuesto aún se encuentra en etapa experimental, por lo que aun no tenemos resultados que midan la bondad del enfoque.

6 Conclusiones y trabajo futuro

Se ha esbozado aquí un modelo CLIR independiente del idioma en un alto grado, gracias al uso de recursos lingüísticos abundantes, tal como son los corpus comparables. Hemos repasado los tres problemas CLIR tradicionales, mostrando resultados ya obtenidos en ambientes bilingües para los dos primeros problemas CLIR. Por último, se ha presentado con detalle un nuevo enfoque para el tercer problema CLIR.

Como trabajo futuro, tres son las puntos que nos hemos marcado: medir la bondad del cálculo del RSV en dos pasos, conseguir mejores traducciones para más idiomas, a partir de corpus comparables de mayor cobertura extraídos del Web, e integrar las soluciones propuestas en un único sistema CLIR.

7 Bibliografía

Adriani M. Dictionary-based CLIR for the CLEF Multilingual Track, 2000. In Working Notes of the Workshop in Cross-Language Evaluation Forum (CLEF), Lisbon, September.

Ballestreros L, Croft W.B., 1997. Resolving Ambiguity for Cross-language Retrieval. In Proceedings of the 20 th International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, (pp 84-91).

Ballesteros L., Croft W.B., 1996: Dictionary-based methods for cross-lingual information retrieval. In Proceedings of the 7th International DEXA Conference on Database and expert Systems Applications, pp 791—801.

Berners-Lee. T. Berners-Lee. Semantic Web Road. Map. www.w3.org/DesignIssues/Semantics.html, 1998.

Braschler M., B. P. Ripplinger, Schäuble. Experiments with the Eurospider Retrieval System for CLEF 2001. Carol Peters, editor, Proceedings of the CLEF 2001 Cross-

Language Text Retrieval System Evaluation Campaign. @Springer-Verlag, 2001.

Gachot, D. A., Lange, A., Yang, J. The SYSTRAN NLP Browser: An Application of Machine Translation Technology in Cross-Language Information Retrieval. In G. Grefenstette, editor, Cross-Language Information Retrieval. pp.105-118, 1998.

Grefenstette G. The problem of Cross-Language Information Retrieval. In Cross-Language Informatioin Retrieval, chapter 1, (pp 3-4). Kluwer Academic Publishers, 1998.

Grefenstette G. "The WWW as a Resource for Example-Based MT Tasks". conference. In ASLIB'99 Translating and the Computer 21, London, UK, Nov 10-11, 1999.

Hiemstra D., Kraaij W., Pohlmann R. and Westerveld T., "Twenty-One at CLEF2000: Translation resources, Merging Strategies and Relevance feedback, In: C. Peters, editor, Cross-language Information Retrieval and Evaluation, Proceedings of CLEF 2000 workshop, pp. 102-116, LNCS 2069, 2001, © Springer-Verlag.

Hull D. A. and Grefenstette G. Experiments in multilingual information retrieval. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval , 1996.

Kraaij, W: TNO at CLEF-2001: Comparing translation resources , Proceedings of CLEF 2001 workshop , Darmstadt 2001 .

Martínez F., Ureña A. y García M. WWW como Fuente de Recursos Lingüísticos para su Uso en PLN. In Proc. SEPLN, Sep. 2001. Pp 141—151, . 2001a.

Martínez Santiago F., Ureña López A., Díaz Galiano M, García Vega, M., Martín Valdivia M. SINAI at CLEF: Evaluating Translation Probabilities with SemCor. Carol Peters, editor. Proceedings of the CLEF 2001 Workshop. @Springer-Verlag, 2001b.

Martínez-Santiago F, Ureña López A. LLajú: Un sistema de Recuperación de Información Multilingüe basado en EuroWordNet. Procesamiento del Lenguaje Natural, Revista nro. 27, septiembre 2001c.

Martínez Santiago F., Ureña López A., Díaz Galiano M, García Vega, M., Martín Valdivia M. Uso de SemCor como recurso lingüístico en tareas CLIR. Actas de la IX Conferencia de la AEPIA (CAEPIA), 2001d.

Martínez Santiago F., Ureña López A., Díaz Galiano M, Rivas, V., Martín Valdivia M. Using Neural Networks for Multiword Recognition in RI. ISKO'2002, pendiente de publicar, 2002a.

Maynard D. and Ananiadou S. TRUCKS: a model for automatic term recognition, Journal of Natural Language Processing, December 2000.

Nie J., Simard M., Isabelle P., and Durand R. Cross-language information retrieval based on parallel texts and automatic mining parallel texts from the Web. In ACM SIGIR'99, (pp.74-81) , 1999.

Oard D. Cross-Language Text Retrieval Research in the USA. Presented at 3rd ERCIM DELOS Workshop, Zurich, Switzerland, 1997.

Pierre J. On the Automated Classification of Web Sites. Linköping Electronic Articles in Computer and Information Science. Vol. 6(2001): nr 0.

Powell L., French J. C., Callan J., Connell M. and Viles C. L., Measuring the Impact of Database Selection on Distributed Searching, Proc. 23rd ACM SIGIR Conference on Information Retrieval (SIGIR 2000), July 2000, pp. 232-239.

Qiu Y. Automatic Query Expansion Based on A Similarity Thesaurus. PhD Thesis, Swiss Federal Institute of Technology (ETH), 1995

Savoy J., 2001. Report on CLEF-2001 Experiments.Experiments with the Eurospider Retrieval System for CLEF 2001. Carol Peters, editor, Proceedings of the CLEF 2001 Cross-Language Text Retrieval System Evaluation Campaign. @Springer-Verlag, 2001

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

Voorhees E. M., Gupta N.K. & Jhonson-Laird B. The collection fusion problem. In Proceedings of TREC'3, (pp. 95-104). Gaithersburg: NIST Publication #500-225, 1995.

Ureña A, Buenaga M y Gómez J.M., "Integrating Linguistic Resources in TC through WSD". En Computer and the Humanities, vol 35, 2, pp. 214-213, 2001.

Vossen P. EuroWordNet: A multilingual database with lexical semantic networks. Dordrecht: Kluwer. , 1998.