ARTÍCULO

Búsqueda de información multilingüe: estado del arte

Fernando López-Ostenero, Julio Gonzalo, Felisa Verdejo
Departamento LSI, ETSI.Informática, UNED
C/Juan del Rosal, 16
28040 MADRID
{flópez,julio,felisa} @lsi.uned.es

Presentamos un estado del arte en el problema de la búsqueda de información multilingüe, con especial atención a los distintos recursos lingüísticos utilizados y a los aspectos interactivos de la búsqueda de documentos en idiomas desconocidos por el usuario.

Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial. Vol. VIII, No.22 (2004),pp. 11-35
ISSN: 1137-3601. © AEPIA (http://www.aepia.org/revista).

Búsqueda de información multilingüe: estado del arte

Fernando López-Ostenero, Julio Gonzalo, Felisa Verdejo
Departamento LSI, ETSI.Informática, UNED
C/Juan del Rosal, 16
28040 MADRID
{flopez,julio,felisa}@lsi.uned.es

Resumen

Presentamos un estado del arte en el problema de la búsqueda de información multilingüe, con especial atención a los distintos recursos lingüísticos utilizados y a los aspectos interactivos de la búsqueda de documentos en idiomas desconocidos por el usuario.

Abstract

This paper summarizes the state of the art in Multilingual Information Retrieval, paying special attention to the linguistic resources used and to the interactive aspects of searching documents in unknown languages.

1 Introducción

Tradicionalmente, la Recuperación de Información se ha entendido como el proceso, totalmente automático, en el que, dada una consulta (expresando las necesidades de información del usuario) y una colección de documentos, se devuelve una lista ordenada de documentos supuestamente relevantes para la consulta. Un motor de búsqueda ideal recuperaría todos los documentos relevantes (lo que implica una *cobertura* completa) y sólo aquellos documentos que son relevantes (*precisión* perfecta). Este modelo tradicional lleva consigo muchas restricciones implícitas; entre ellas, la suposición de que la consulta y el documento están escritos en el mismo idioma. La mayoría de los motores de búsqueda en Internet, de hecho, tienen la limitación de encontrar documentos sólo en el idioma en que se escribe la consulta. Algunos incorporan sistemas de traducción automática para traducir los documentos encontrados, que sólo resultan útiles cuando éstos ya han sido localizados, pero no facilitan un medio efectivo para salvar la barrera del idioma en el proceso de búsqueda.

El término *Acceso Multilingüe a la Información* hace referencia a un concepto más amplio, aunque más adaptado a la realidad de Internet, que el concepto clásico de recuperación de información: ayudar al usuario a buscar información (no ya documentos) procedente de fuentes heterogéneas (textuales o de contenido multimedia) por encima de las barreras idiomáticas. Diversas líneas de investigación abordan los distintos aspectos que se engloban en este concepto incluso dentro del mismo marco del Procesamiento del Lenguaje Natural: *Recuperación Multilingüe de Información*, *Recuperación de Información Multimedia* (ya sea sobre video, audio o imágenes digitales), *Recuperación Interactiva*

Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial, No. 22 (2003), pp. 11-35.
ISSN: 1137-3601. © AEPIA (http://www.aepia.org/revista).

de Información, *Sistemas de Pregunta y Respuesta...* etc.

En este artículo, nos centramos el estudio de la recuperación de información translingüe, que trata el problema de encontrar documentos que están escritos en idiomas distintos al de la consulta. En general, sólo se conoce bien el comportamiento del inglés, y para conseguir una recuperación translingüe eficiente es necesario disponer de buenos sistemas de búsqueda monolingüe en idiomas de otras características, por tanto en primer lugar es necesario estudiar las características propias de cada idioma a la hora de efectuar la recuperación monolingüe de documentos. En segundo lugar, hablaremos de búsqueda bilingüe cuando la consulta éste en un idioma origen y los documentos en un único idioma destino. Finalmente, hablaremos de búsqueda multilingüe cuando la consulta éste en un idioma origen y los documentos distribuidos en varias colecciones de idiomas diferentes. En este caso, el problema consiste en devolver un único ranking de documentos relevantes escritos en todos los idiomas considerados.

Por último, hablaremos de recuperación translingüe interactiva cuando estudiemos qué tipo de asistencia puede proporcionar un sistema de recuperación translingüe para que un usuario formule sus consultas, identifique información relevante y sea capaz de refinar sus necesidades de búsqueda sobre información escrita en idiomas que desconoce.

Ya en 1969 Salton planteó por primera vez el problema de encontrar documentos escritos en un idioma diferente al de la consulta y propuso una aproximación consistente en la utilización de un tesauro bilingüe (creado manualmente) entre alemán e inglés (Salton, 1970). Los resultados obtenidos fueron prácticamente iguales a los realizados con una búsqueda monolingüe, debido a que el tesauro utilizado había sido construido manualmente (de tal forma que no existía ambigüedad en los términos de indexación) y la correspondencia entre los términos de indexación entre ambos idiomas era perfecta. De esta forma el problema de la ambigüedad de las palabras

Pero no fue hasta 1996 cuando, con la creación de las primeras campañas de evaluación comparada sistemática de este tipo de sistemas, se inicia como un área de investigación propia.En ese año se organizó un workshop específicamente dedicado a la recuperación translingüe de información en el SIGIR [1]. A partir de este evento se organizan con carácter regular las siguientes actividades internacionales:

- Desde 1997 se creó un "track" especial en el marco del TREC [2] para la evaluación de este tipo de sistemas.

  Inicialmente la evaluación se limitó a un sistema bilingüe (involucrando dos idiomas de entre inglés, francés o italiano) para, posteriormente ser extendida a una evaluación en un entorno totalmente multilingüe. El resultado de los tracks de recuperación de información translingüe del TREC es la primera gran colección para la evaluación de sistemas de recuperación translingüe de información.

- En 1998 se crea el workshop NTCIR [3], donde se evalúan, entre otras cosas, sistemas translingües entre el inglés y el chino, japonés o coreano, adoptando muchas de las ideas en las que el TREC fue pionero.

- En el año 2000 el track de recuperación translingüe se separó del TREC creándose el CLEF [4] (Peters, 2001) donde se realiza el estudio de sistemas multilingües de recuperación de información que utilicen idiomas europeos, mientras que en el TREC se mantuvo un track de recuperación de información translingüe específicamente dedicado a idiomas asiáticos.

A lo largo de todas estas evaluaciones comparadas se han desarrollado y contrastado con éxito una serie de técnicas y recursos que hacen de la recuperación translingüe de información un tema de investigación relativamente maduro.

En este artículo analizamos las diversas técnicas que han venido utilizándose para salvar la ba-

[1] Special Interest Group on Information Retrieval (grupo de interés especial en la recuperación de información de la ACM) http://www.acm.org/sigir/

[2] Text REtrieval Conference (Conferencia sobre recuperación de textos) http://trec.nist.gov/

[3] NII-NACSIS Text Collection for IR systems (colección textual para sistemas de recuperación de información) http://research.nii.ac.jp/~ntcadm/index-en.html

[4] Cross-Language Evaluation Forum (Foro para la Evaluación de la Recuperación de Información Translingüe) http://www.clef-campaign.org

rrera idiomática en una búsqueda translingüe de información.

Comenzaremos viendo diversas técnicas que son utilizadas para mejorar la recuperación de información monolingüe (sección 2) en idiomas que no presentan las características del inglés.

En la sección 3 veremos los enfoques que se han utilizado para traducir las consultas introducidas por el usuario a los diferentes idiomas en los que están escritos los documentos. Estos enfoques dependen, sobre todo, de los recursos que se utilicen (aisladamente o en combinación): diccionarios bilingües, corpora, programas de traducción automática, tesauros... A continuación, en la sección 4, veremos los principales enfoques alternativos a la traducción de la consulta: traducción de los documentos, traducción bidireccional e indexación conceptual.

Finalmente, en la sección 5 revisaremos las investigaciones sobre los aspectos interactivos de las búsquedas translingües.

2 Aspectos monolingües

A lo largo de la investigación en recuperación de información se han aplicado con éxito diversos modelos (como el *modelo de espacio vectorial*, la *Realimentación mediante Pseudo-Relevancia* o la *Indexación mediante semántica latente*) a búsquedas realizadas sobre consultas y documentos escritos casi siempre en inglés.

Al enfrentarnos a idiomas que presentan características distintas al inglés (idiomas más flexivos, idiomas aglutinativos o incluso idiomas que no marcan una separación explícita entre las palabras) es necesario mejorar la búsqueda monolingüe sobre esos idiomas para poder realizar una búsqueda translingüe efectiva.

Veamos diferentes técnicas que son utilizadas en el momento de la indexación de los documentos para mejorar las búsquedas:

2.1 Stemming

Una de las técnicas que ha demostrado ser de gran ayuda en la recuperación de información monolingüe es el **stemming**. Consiste en la obtención de la raíz de las palabras, de forma que el proceso de indexación se lleve a cabo sobre ellas en lugar de sobre las palabras originales. Asumiendo que dos palabras que tengan la misma raíz representan el mismo concepto, esta técnica permite a un sistema de recuperación de información relacionar términos presentes en la consulta y en los documentos que pueden aparecer bajo diferentes variantes morfológicas. Además, reduce apreciablemente el espacio de indexación.

Existen diversos stemmers para inglés basados en la eliminación de sufijos derivacionales (Lovins, 1968; Dawson, 1974; Porter, 1980). También existen stemmers para otros idiomas como francés (Savoy, 1999), castellano (Figuerola et al., 2002), árabe (Abu-Salem et al., 1999), holandés (Kraaij and Pohlmann, 1994), griego (Kalamboukis, 1995) e incluso latín (Schinke et al., 1996). En general, estos algoritmos no llevan a cabo ningún análisis morfológico sofisticado, sino que se basan en un conjunto sencillo de reglas que truncan las palabras hasta obtener una raíz común.

Una alternativa es el aprendizaje de las reglas de truncamiento a partir de grandes corpora. Un ejemplo en este sentido es (Bacchin et al., 2002) donde se evalúa SPLIT: un algoritmo de stemming independiente del idioma basado en métodos estadísticos. Analizando un conjunto de palabras, que forman parte del idioma, SPLIT detecta los sufijos y prefijos que las forman y selecciona como raíz de cada palabra el prefijo más probable. Para realizar la evaluación de este algoritmo, se aplicó a un conjunto de documentos en italiano y se comparó la precisión de los resultados de la búsqueda utilizando SPLIT como stemmer y otro stemmer específicamente diseñado para este idioma disponible en la página web de Snowball (Porter, 2001). Los resultados mostraron que la calidad de SPLIT era comparable a la del stemmer de italiano.

2.2 Segmentación de compuestos

En los idiomas idiomas aglutinativos, como alemán y holandés, se unen palabras para formar otras más largas. Por ejemplo la palabra holandesa “wereldbevolkingsconferentie”

está compuesta por "wereld" (mundo), "bevolking" (población) y "conferentie" (conferencia), y se traduce como "Conferencia sobre la población mundial".

Diversos estudios muestran que la descomposición de estas palabras en lemas individuales produce una significativa mejora en las búsquedas en este tipo de idiomas al considerar cada elemento de la palabra compuesta como un término (Kraaij and Pohlmann, 1998; Monz and de Rijke, 2002).

Una alternativa a la descomposición empleando métodos lingüísticos (que exigen disponer de herramientas adecuadas en precisión, cobertura y eficiencia) es el uso de métodos estadísticos. En (McNamee and Mayfield, 2001) se presenta una aproximación a la recuperación multilingüe de información utilizando recursos independientes del idioma. Los documentos de cada uno de los idiomas son indexados utilizando 6-gramas [5]. Se realizan dos búsquedas monolingües, una empleando los 6-gramas y otra con palabras (sin ningún tipo de procesamiento adicional), cuyos resultados se combinan para ofrecer una única lista de documentos. Los resultados obtenidos fueron los mejores sobre idiomas aglutinativos en el `CLEF'2000`, quedando incluso por delante de otros sistemas que utilizaban algoritmos específicos para descomponer las palabras.

Esta estrategia también ha sido probada con otros idiomas como el árabe (Mayfield et al., 2001), llegando a alcanzar una eficiencia superior al 90% de la búsqueda monolingüe equivalente utilizando, en este caso, 4-gramas.

2.3 Segmentación de palabras

En los idiomas asiáticos, como japonés, coreano y chino, los límites de las palabras no se marcan de manera explícita en el texto escrito. Por ello es necesario identificar las palabras individuales para mejorar el proceso de búsqueda.

A la hora de indexar los textos escritos en estos idiomas, existen dos aproximaciones principales:

- Indexación basada en texto segmentado: que incluye la indexación de palabras y/o de sintagmas.
- Indexación de caracteres: basada en n-gramas. Fundamentalmente se utilizan bigramas, ya que en japonés, chino y coreano, la longitud media de las palabras es de, aproximadamente, dos caracteres al ser, fundamentalmente, idiomas silábicos.

Algunos estudios han mostrado que las búsquedas textuales en chino y coreano basadas en la indexación mediante bigramas obtienen resultados comparables (y, en ocasiones, incluso mejores) a las basadas en indexación mediante palabras (Lee and Ahn, 1996; Kwok, 1997; Chen et al., 1999).

En (Ozawa et al., 1999) se argumenta que los bigramas son insuficientes cuando se indexan documentos conteniendo lenguaje técnico, donde la longitud de las palabras es superior a la media. Se comprueba que un método adaptativo de segmentación que produce n-gramas de varias longitudes, supone una mejora substancial con respecto a la utilización de bigramas.

A pesar de los resultados anteriores no parece existir un claro consenso acerca de cual de las dos aproximaciones (n-gramas o palabras) es mejor para la indexación de textos en este tipo de idiomas. En muchas ocasiones la combinación de ambas presenta una clara mejora sobre ambas (Fukushima and Akamine, 1999).

3 Enfoques basados en la traducción de la consulta

A la hora de realizar una búsqueda translingüe de información, nos enfrentamos a la siguiente situación: la consulta y los documentos no están escritos en el mismo idioma. Es, por tanto, necesario efectuar alguna forma de traducción para poder realizar una búsqueda en la que tanto consulta como documentos se encuentren en el mismo idioma.

La traducción de la consulta es la opción más frecuente. Por ejemplo los 9 participantes que realizaron experimentos en recuperación translingüe en el TREC-10 emplearon esta

[5] Los n-gramas son conjuntos de n caracteres que aparecen juntos en el texto.

técnica (Gey and Oard, 2001). Esto es debido, principalmente, a que la consulta es sensiblemente más pequeña que los documentos y, por ello, el coste computacional de su traducción es mucho menor (Hull and Grefenstette, 1996).

En (Grefenstette, 1998) se identifican los tres problemas principales a los que se enfrenta un sistema de búsqueda translingüe de información al traducir la consulta:

1. Saber cómo un término escrito en un idioma puede ser expresado en otro idioma.
2. Decidir cuáles de las posibles traducciones de cada término son las adecuadas en ese contexto.
3. Saber cómo pesar la importancia de las diferentes traducciones que se consideran adecuadas.

Los dos primeros retos son compartidos por los sistemas de traducción automática. Sin embargo, un sistema de traducción automática debe dar una única traducción para cada término, mientras que un sistema de recuperación translingüe de información puede dar varias y asignarles distintos pesos.

En esta sección veremos diferentes recursos y cómo se han utilizado a la hora de traducir las consultas. Estos recursos no son utilizados por separado, cada uno puede aportar información complementaria al problema de la traducción.

En el apartado 3.1 comenzaremos viendo los problemas que plantea la utilización de versiones electrónicas de diccionarios bilingües, así como una perspectiva histórica sobre su uso. En el 3.2 veremos cómo se ha utilizado la información proveniente de corpora (ya sea paralelo o comparable) para realizar el proceso de traducción. En el apartado 3.3 discutiremos sobre el uso de programas de traducción automática y en el 3.4 contaremos el uso de tesauros multilingües. Finalmente en el apartado 3.5 se abordará el problema de la fusión que aparece en un entorno multilingüe.

3.1 Diccionarios

La utilización de versiones electrónicas de diccionarios bilingües como recurso de traducción palabra por palabra, ha sido ampliamente estudiada en la literatura. Sin embargo su uso directo no resuelve por completo el problema de encontrar las traducciones de los términos, debido a las siguientes razones:

- La cobertura del diccionario puede no ser completa, por lo que algunos términos no son traducidos. Esto sucede frecuentemente con los términos técnicos que no son de uso común. La **terminología** específica de un determinado dominio del conocimiento no suele estar contemplado en los diccionarios de uso común.
- No contemplan todas las posibles **variantes morfológicas** de una palabra. Por ejemplo un diccionario puede contener el término "*asintótico*" pero quizá no contenga "*asintóticamente*". Este problema puede ser mitigado empleando la técnica de stemming comentada en la sección anterior.
- En ocasiones es necesario traducir los nombres propios de personas (el nombre "*Yeltsin*" se escribe "*Eltsine*" en francés) o localizaciones ("*Letonia*" se escribe "*Latvia*" en inglés) y estas traducciones pueden no estar contempladas en el diccionario. Este problema está relacionado con el "**reconocimiento de entidades**".
- Para cada contexto, sólo algunas traducciones son apropiadas. Por ejemplo la palabra inglesa "*spring*" tiene diversas traducciones en castellano con significados muy distintos entre sí: "*muelle*", "*primavera*", "*manantial*"... La **polisemia** de las palabras dificulta la traducción y no se cuenta con métodos automáticos que puedan resolverla satisfactoriamente.
- La traducción errónea de los términos es particularmente perjudicial en los conceptos representados por **expresiones multipalabra**. Por ejemplo la palabra castellana "*banco*" se traduce frecuentemente por "*bank*" en inglés. Sin embargo la expresión "*banco de peces*" ha de traducirse por "*school of fish*".

Por todas estas razones la utilización de un diccionario como único recurso de traducción reduce drásticamente la efectividad de las

búsquedas translingües. Diversos trabajos como (Hull and Grefenstette, 1996; Ballesteros and Croft, 1996) comprueban que substituyendo cada término por todas las traducciones ofrecidas por el diccionario se reduce la efectividad entre un 40 y un 60% respecto de la misma búsqueda realizada en un contexto monolingüe.

Con respecto a la polisemia (Davis, 1997) propone utilizar la categoría gramatical de las palabras de la consulta para elegir entre las posibles traducciones de los términos: por ejemplo la palabra inglesa "*object*" puede actuar como nombre y ser traducida al castellano como "*objeto*", "*objetivo*" o "*complemento*", mientras que si actúa como verbo puede traducirse por "*objetar*" u "*oponerse*". Utilizando un diccionario bilingüe con información sobre la categoría gramatical para traducir las consultas, Davis comprobó que esta estrategia incrementaba en un 37% la precisión con respecto a la estrategia de substituir cada término por todas las traducciones ofrecidas por el diccionario.

En (Ballesteros and Croft, 1997) se intenta mejorar la efectividad de las traducciones utilizando diccionarios de traducción de expresiones multipalabra [6]. Cuando estas traducciones de sintagmas eran correctas, las búsquedas eran un 150% más eficientes que aquellas realizadas utilizando consultas traducidas únicamente palabra por palabra. Por desgracia este tipo de diccionarios no es frecuente, y sólo un pequeño porcentaje de consultas contenía términos de este tipo.

(Pirkola, 1998) estudia los efectos de diferentes factores:

- **Formulación de la consulta**: comparó consultas escritas en lenguaje natural con consultas formadas únicamente por las palabras y sintagmas más relevantes de la consulta. La precisión de las búsquedas fue mayor con las consultas expresadas en lenguaje natural.
- **Proceso de traducción**: utilizó dos diccionarios bilingües para realizar la traducción: uno de propósito general y otro con información específica sobre el dominio de la medicina y la salud. Probó varias formas de combinar estos diccionarios, comprobando que la que mejores resultados daba era la de utilizar las suma de todas las traducciones proporcionadas por ambos diccionarios (eliminando traducciones duplicadas).
- **Estructura de la consulta tras la traducción**: comparó la utilización de consultas sin ningún tipo de estructura (una simple lista de todas las traducciones) con el uso de consultas estructuradas mediante los operadores proporcionados por el motor de búsqueda Inquery (Callan et al., 1992). Las traducciones provenientes de un mismo término se agruparon mediante un operador de sinonimia y los términos multipalabra se identificaron con un operador de proximidad. La estructuración de la consulta resultó ser el factor que incrementó en mayor medida la precisión de las búsquedas, superando en algunos casos el 50% de incremento.

(Sperer and Oard, 2000) se plantean la utilización de un diccionario bilingüe estructurado en el que las traducciones de cada término se encuentran agrupadas en conjuntos con un significado claramente similar. No existen muchos diccionarios bilingües que presenten esta estructura, por lo que los autores desarrollan, además, un método que permite dotar de esta estructura a cualquier diccionario bilingüe empleando criterios lingüísticos (similitud entre las palabras según WordNet (Miller, 1990)), morfológicos (agrupar las palabras que comparten la misma raíz) y ortográficos (agrupar las palabras que se diferencien en un único carácter).

Compararon la estructuración de la consulta propuesta por (Pirkola, 1998) con otras alternativas, empleando para ello diferentes operadores del lenguaje de consulta de Inquery y los conjuntos de traducciones agrupadas. Los resultados mostraron que la traducción de las consultas con la estructuración propuesta por Pirkola obtenían una mayor precisión que la traducción utilizando los diccionarios estructurados.

En (Gollins and Sanderson, 2001) se propone utilizar dos idiomas pivote para realizar la traducción cuando no se dispone de un diccionario directo. En primer lugar traducen consultas del alemán al español y al holandés utilizando EuroWordnet (Vossen, 1998). Posteriormente estas consultas se vuelven a traducir al inglés

[6] Formadas por secuencias de nombres y parejas nombre-adjetivo.

y se combinan las traducciones para producir lo que los autores denominan una triangulación léxica. Los resultados demuestran que utilizar un idioma pivote para traducir entre dos idiomas provoca una mayor pérdida de eficiencia que la utilización de un diccionario directo, al igual que los resultados obtenidos en (Ballesteros, 2000).

En (Boughanem et al., 2002) se realiza una selección de las traducciones empleando las traducciones inversas: sólo aquellas traducciones que pueden volver a traducirse al término de partida son seleccionadas. Los resultados muestran que esta simple estrategia puede ser más efectiva que otras más complejas como la desambiguación de traducciones empleando corpora paralelo.

3.2 Utilización de corpora

3.2.1 Corpora paralelo

El concepto de *Corpora Paralelo* hace referencia a varias colecciones de documentos escritas en diferentes idiomas en las que se puede relacionar cada documento de una colección con un documento de otra colección que son traducción el uno del otro. En ocasiones la información de traducción es más fina, refiriéndose no a documentos completos, sino a partes de documentos.

EBMT

Un ejemplo de utilización de corpora paralelo alineado a nivel de sintagmas es la llamada *traducción mediante ejemplo* (Example-Based Machine Translation). Se parte de un corpus que contiene información acerca de la traducción de los sintagmas y frases contenidos en él, como base para traducir cualquier otro texto (Brown, 1996; Brown, 1997; Carl and Hansen, 1999; Collins, 1999; Nirenburg et al., 1994)

La traducción mediante ejemplo ha demostrado ser una técnica muy eficiente como método para traducir las consultas (Yang et al., 1998), alcanzándose en la búsqueda translingüe una eficiencia similar a la de la búsqueda monolingüe equivalente. La principal desventaja que presenta esta técnica es la necesidad de disponer de corpora paralelo alineado a nivel de sintagmas, lo que la restringe a dominios específicos en los que existan este tipo de recursos.

PRF

La técnica denominada *Pseudo-Relevance Feedback* [7] o PRF (Buckley et al., 1995) es utilizada en la recuperación monolingüe de información para expandir la consulta con términos potencialmente útiles. Consiste en asumir que los documentos que ocupan los primeros puestos del ranking devuelto por el sistema son relevantes para la consulta (sin ninguna intervención por parte del usuario). No siempre es efectiva, ya que entre estos documentos pueden aparecer algunos que no sean relevantes. Diversos estudios (Hersh et al., 1994; Srinivasan, 1996) han encontrado evidencias a favor y en contra de su uso.

Si se dispone de corpora paralelo es posible extender la técnica de PRF a un entorno multilingüe, sin más que utilizar la información acerca de la correspondencia de los documentos (o partes de documentos) que son el uno traducción del otro. Podemos encontrar varios estudios acerca de la eficiencia de la PRF en un entorno multilingüe (Carbonell et al., 1997; Braschler et al., 2000a) con resultados que muestran una mínima pérdida con respecto a la búsqueda monolingüe utilizando también esta misma técnica.

GVSM

Uno de los modelos utilizados en la recuperación de información es el modelo de espacio vectorial (VSM) (Salton and Buckley, 1983). En él las consultas y los documentos son representados mediante vectores y la similitud entre ellos se mide utilizando los ángulos que forman éstos. Una extensión a este modelo, el modelo de espacio vectorial generalizado (GVSM) (Wong et al., 1985), puede ser adaptada a la recuperación translingüe de información (Carbonell et al., 1997) utilizando corpora paralelo alineado a nivel de documentos. Los resultados muestran que la pérdida de eficiencia de la búsqueda translingüe es de, aproximadamente, un 9%.

LSI

Una extensión del GVSM es la llamada *Latent Semantic Indexing* [8] (LSI) (Deerwester et al.,

[7] Realimentación mediante pseudo-relevancia.
[8] Indexación mediante semántica latente.

1990). Mientras que en el VSM la base ortogonal del espacio vectorial está formada por palabras y en el GVSM por documentos, en LSI se utiliza la combinación lineal de las dimensiones originales que posea un mayor significado. Al igual que el GVSM, es posible extender LSI a un entorno multilingüe (Dumais et al., 1996), sin más que utilizar corpora paralelo para calcular la nueva base del espacio vectorial.

En (Carbonell et al., 1997) se realiza por primera vez una comparación entre GVSM y LSI, tanto en sus versiones monolingües como translingües. Los resultados muestran que GVSM se comporta mejor que LSI en ambos escenarios, pero presenta una pérdida de eficiencia superior a LSI al pasar de monolingüe a translingüe (un 9% para GVSM frente a un 1% para LSI).

Todas estas técnicas requieren la utilización de corpora paralelo (a diferente nivel de alineación) para poder trabajar en un entorno multilingüe. En muchas ocasiones no es posible disponer de corpora de estas características, por lo que su utilidad queda limitada por este hecho.

3.2.2 Construcción automática de corpora paralelo

Una posible solución a la falta de corpora paralelo es utilizar los motores de búsqueda en la red para encontrar páginas web que tengan versiones en dos idiomas diferentes. De esta manera se podría construir corpora paralelo entre diversos idiomas.

En (Resnik, 1998) se implementó un prototipo llamado STRAND con el que se realizó una búsqueda de documentos escritos en inglés y castellano. Tras eliminar los errores y las páginas duplicadas se encontraron 90 parejas de webs candidatas a ser traducción la una de la otra. Tras una evaluación manual se vió que únicamente 24 podían realmente considerarse como traducciones correctas.

De las 90 parejas candidatas, STRAND marcó 17 como traducciones correctas. Comparando estos resultados con la evaluación manual se comprobó que 15 de las 17 traducciones seleccionadas por STRAND eran realmente traducciones correctas. Estos datos suponen una precisión del $88'2\%$ y una cobertura del $62'5\%$.

En (Chen and Nie, 2000) se implementa otro sistema llamado PTMiner con el que se construye un corpora paralelo para inglés y francés (con, aproximadamente, un 95% de precisión en las alineaciones) y otro para inglés y chino (que alcanzó una precisión del 80%).

Los experimentos utilizando el corpora paralelo así construido volvieron a mostrar que la traducción de las consultas utilizando la información de corpora paralelo permite lograr una precisión mucho mayor que la obtenida con las consultas traducidas mediante el uso de un diccionario bilingüe.

En un estudio posterior (Nie et al., 2001) se empleó el mismo sistema PTMiner para obtener corpora paralelo en los pares de idiomas inglés-italiano e inglés-alemán. Junto con los datos previamente obtenidos para inglés y francés se realizaron unos experimentos en el marco del CLEF para comprobar la utilidad del corpora obtenido. Los resultados muestran que la utilización de corpora paralelo extraído de la web es un recurso que resulta muy útil al ser utilizado en la traducción de las consultas para la recuperación translingüe de información.

Los corpora obtenidos por PTMiner se pusieron a disposición de los participantes en el `CLEF`, y fueron usados con éxito por varios grupos como información complementaria en sus sistemas.

3.2.3 Corpus monolingüe

Una alternativa al uso de corpora paralelo consiste en utilizar la propia colección de documentos como corpus de referencia. Esto no resuelve el problema de encontrar las traducciones de los diferentes términos (el corpus está escrito en un único idioma) pero puede ser utilizado como un apoyo para seleccionar y pesar las diferentes traducciones ofrecidas por un diccionario bilingüe como en (Chen and Gey, 2001).

(Ballesteros and Croft, 1998) proponen la utilización de estadísticas de coocurrencia de términos sobre un corpus en el idioma de los documentos como método para desambiguar las posibles traducciones de sintagmas.

La hipótesis de los autores plantea que las traducciones correctas de los términos de la consulta coocurrirán frecuentemente en un corpus

del idioma de los documentos, mientras que las incorrectas no lo harán.

Para comprobarlo compararon dos técnicas para realizar la desambiguación de sintagmas [9]:

- Mediante corpora paralelo: se utiliza corpora paralelo (alineado a nivel de documentos) para realizar la desambiguación. Con la consulta original se recuperan 30 documentos en el corpus del idioma de la consulta. De los documentos equivalentes a los recuperados (en el corpus del idioma de los documentos) se extraen 5000 términos y se ordenan utilizando una medida basada en la frecuencia de aparición en los documentos de los que provienen. Después, las traducciones de cada término de la consulta son ordenadas según su aparición en esta lista de 5000 términos.

- Mediante coocurrencia estadística: para cada una de las posibles combinaciones de traducciones de parejas de términos se obtiene una puntuación basada en la frecuencia de aparición de las traducciones en el corpus del idioma de los documentos. La combinación que obtiene la mayor puntuación se elige como mejor traducción.

Para comparar ambas técnicas se realizaron búsquedas translingües partiendo de consultas en castellano y recuperando documentos en inglés. Los resultados demostraron que la desambiguación mediante coocurrencia estadística alcanzaba una efectividad similar (e incluso superior) a la desambiguación mediante corpora paralelo. En conjunto se logró un 90% de la efectividad de la búsqueda monolingüe.

Así pues la desambiguación mediante coocurrencia estadística demuestra ser una buena alternativa a la utilización de corpora paralelo para realizar la desambiguación de las traducciones de sintagmas, especialmente cuando la obtención de corpora paralelo para el par de idiomas considerados resulte difícil.

[9] formados por secuencias de nombres y parejas nombre-adjetivo.

3.2.4 Corpora comparable

Otra alternativa al uso de corpora paralelo es la utilización de *Corpora Comparable*. La obtención de corpora comparable es más sencilla, ya que sólo se requiere que los corpus en distintos idiomas tengan una temática similar, pero no que haya documentos equivalentes entre idiomas.

Uno de los primeros trabajos en aprovechar corpora comparable es (Peters and Picchi, 1997), donde se describe su utilización en un sistema multilingüe de recuperación de información para expandir las consultas, no sólo con las traducciones de cada término, sino también con un vocabulario que define un contexto probable en ambos idiomas. Así cuando el diccionario no ofrecía ninguna traducción para un término, la búsqueda translingüe es posible al haberse enriquecido la consulta con el contexto de dicho término aprendido del corpus.

Existen diversos trabajos que tratan sobre la alineación de corpora comparable. Las diferentes aproximaciones utilizan análisis lingüístico sofisticado (Braschler and Schäuble, 1998), métodos estadísticos que consideran la frecuencia de las palabras en ambos corpus (Chen, 1993; Kay and Röscheisen, 1993) y la longitud del texto analizado (Gale and Church, 1991) o, incluso, la alineación de cognados [10] (Simard et al., 1992). Estas alineaciones pueden ser utilizadas para generar recursos de traducción que pueden ser aprovechados en la recuperación translingüe de información (Fung, 1995).

Los llamados "*tesauros de similitud*" (Qiu and Frei, 1993) son otra forma de aprovechar corpora comparable para recuperación de información multilingüe. Mientras que los tesauros son construidos de forma manual por especialistas en el tema que cubren (ver sección 3.4), los tesauros de similitud extraen relaciones de proximidad temática de forma automática a partir del vocabulario presente en la colección a indexar. La utilización de estos tesauros para realizar expansiones de la consulta puede suponer una mejora sustancial en la eficiencia de las búsquedas monolingües (Qiu, 1995) y multilingües (Braschler and Schäuble, 2001). Esta

[10] Dos palabras en distinto idioma se denominan cognados si provienen de la misma palabra o estructura. Normalmente los cognados tienen estructuras fonológicas similares.

técnica, por tanto, combina el análisis del corpus que forman los documentos con la idea de tesauro.

Con la llegada de los foros de evaluación de sistemas de recuperación translingüe de información se crean los primeros corpora comparable específicamente diseñados para la evaluación de este tipo de sistemas (incluyendo información acerca de la relevancia de los documentos). El corpora comparable creado en el TREC contiene, mayoritariamente, noticias de periódicos en alemán, francés, inglés, e italiano (Braschler et al., 1999). El sucesor de este corpora ha sido el creado en el CLEF, cuya tercera edición contiene más de un millón de documentos en 8 idiomas: alemán, castellano, finlandés, francés, holandés, inglés, italiano y sueco (Peters, 2002a).

3.3 Programas de traducción automática

Otro recurso ampliamente utilizado para la traducción son los programas comerciales de traducción automática, siempre que exista uno disponible para el par de idiomas considerados. En la octava edición del TREC, al menos la mitad de los grupos participantes emplearon el sistema de traducción automática Systran de alguna forma en sus experimentos (Braschler et al., 2000b). Sin embargo otros métodos basados en la combinación de corpus y diccionarios obtuvieron mejores resultados.

Los experimentos acerca de la efectividad de estos programas a la hora de traducir la consulta no aportan datos concluyentes. (Oard, 1998) sugiere que la efectividad puede depender de la longitud de las consultas: para consultas cortas (entre 1 y 3 términos) no parece haber diferencia entre esta aproximación y la utilización de diccionarios para la traducción. Para consultas largas (formadas por varias frases) sí se aprecia diferencia. (Nie, 1999) comprueba que con consultas basadas en frases, la traducción mediante Systran da mejores resultados en las búsquedas que otros métodos de traducción basados en diccionarios o corpus.

Esto es debido a que los sistemas de traducción automática hacen uso de la estructura sintáctica del texto. Si las consultas están formadas por frases, los sistemas de traducción consiguen una traducción mejor que si la consulta está formada por términos independientes sin estructura.

Aparte de este problema, el uso de sistemas de traducción automática depende de la existencia de un traductor entre los idiomas considerados. La creación de estos traductores es costosa, y por eso sólo existen para los pares de idiomas más demandados por el mercado.

(Jones and Lam-Adesina, 2002) utilizaron un sistema comercial para la traducción de consultas en francés, alemán, italiano, castellano, chino y japonés al inglés. Vieron que las diferencias entre la búsqueda monolingüe y las translingües dependían bastante del idioma de partida oscilando entre un $2'3\%$ de pérdida en el caso del francés y un $29'5\%$ para el chino.

(Kraaij, 2002) realizó una comparación sistemática de tres tipos de recursos para la traducción de las consultas en una búsqueda translingüe: diccionarios, corpora paralelo (obtenido de la web utilizando el sistema PTMiner) y traducción automática (utilizando Babelfish [11]). Los resultados mostraron que los tres métodos alcanzaron, al menos, el 90% de la eficiencia de una búsqueda monolingüe. Además encontraron, al igual que en (Jones and Lam-Adesina, 2002), que la diferencia de eficiencia dependía bastante del par de idiomas considerados.

3.4 Tesauros

Un tesauro [12] está formado por la colección de términos o palabras clave que se utilizan para realizar la indexación de los documentos (ya sea ésta manual o automática), así como las relaciones semánticas que los unen.

La utilización de tesauros en el campo de la recuperación de información se centra en el enriquecimiento de la consulta con términos relacionados que aparecen realmente en los documentos, aunque hay otros muchos aspectos en los que pueden ser utilizados (Soergel, 1997):

- Proporcionan un vocabulario controlado

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

[12] Del sustantivo latino Tesaurus-Tesauri: tesoro, depósito de riqueza. Se toma la acepción del primer diccionario analógico inglés: "Thesaurus of English Words and Phrases".

para expresar las consultas, por lo que se elimina el problema del desconocimiento por parte del usuario de los términos que aparecen realmente en los documentos.

- Permiten dar una mejor estructuración a los resultados. Por ejemplo la construcción de un resumen temático estructurado del documento, describiendo los temas principales del mismo así como los diferentes subtemas tratados, empleando para ello conjuntos de términos semánticamente relacionados (Loukachevitch and Dobrov, 2000).

- Su estructuración jerárquica hacen posible su utilización en un entorno de búsqueda interactivo. Los usuarios pueden identificar los diferentes conceptos navegando por la jerarquía y, de esta forma, precisar su búsqueda.

- Un tesauro multilingüe sobre un dominio determinado permite la traducción de términos específicos de ese dominio que quizá no puedan encontrarse en un diccionario bilingüe. Un ejemplo de tesauro multilingüe sobre el dominio médico es el metatesauro de UMLS [13] (National Library of Medicine, 1997).

Los tesauros construidos para la indexación manual de los documentos describen un idioma artificial (basado en uno real) sobre un dominio específico, incluyendo información adicional con anotaciones para los indexadores sobre los términos que lo componen. Estos tesauros no resultan apropiados para ser utilizados en un entorno automático de indexación (Salton, 1989), al carecer de la información necesaria que aporta el sentido común de las personas que realizan la indexación manual.

Los tesauros multilingües fueron el primer tipo de recursos específicamente diseñados para la recuperación de información translingüe. Un ejemplo lo podemos encontrar en el tesauro EuroVoc de la Comunidad Europea, abarca 9 idiomas y se utiliza en la actualidad para la recuperación de documentos europeos (EUROVOC, 1995). (Loukachevitch and Dobrov, 2002) apuntan los requisitos que han de tenerse en cuenta a la hora de desarrollar estos tesauros para el procesado automático de documentos textuales:

- Es necesario describir de forma precisa las diferentes variantes de un mismo concepto en diferentes idiomas. Algunos conceptos se describen con una palabra en un idioma, mientras que en otros son necesarias varias (por ejemplo la palabra rusa "*dissident*" es equivalente a "*political dissident*" en inglés).

- Es preciso, además, describir de forma manual extensos conjuntos de sinónimos para cada concepto analizado en cada uno de los idiomas considerados.

- Se requiere detallar la mayor cantidad posible de términos multipalabra que definan un concepto determinado. De esta forma se podrán utilizar como base para realizar una desambiguación léxica.

En (Sheridan et al., 1997) se construye un tesauro de similitud multilingüe sobre dos colecciones en el dominio de la ley federal suiza, que contienen documentos escritos en francés, alemán e italiano. Los resultados muestran que las búsquedas translingües realizadas, empleando este tesauro de similitud para traducir la consulta, presentan una mínima pérdida de precisión frente a las equivalentes búsquedas monolingües. Basándose en este tesauro se creó EUROSPIDER [14], un motor de búsqueda multilingüe sobre el dominio de la ley federal suiza, que es utilizado en la actualidad por los profesionales del sistema legal de ese país.

La utilización de tesauros en la recuperación de información translingüe queda supeditada a disponer de un tesauro multilingüe que cubra el dominio de las colecciones documentales que van a ser utilizadas. En el caso de los tesauros de similitud es necesario disponer de corpora paralelo (o comparable) para poder construir uno multilingüe. Por estas razones, los tesauros no son moneda común en la recuperación translingüe de información. Sin embargo, en dominios en los que el uso de tesauros está generalizado, como en medicina, la situación es bien distinta; por ejemplo, en el marco del proyecto *Muchmore*, financiado por la Comisión Europea y la National Science Founda-

[13] Unified Medical Language System: sistema unificado de terminología médica.

[14] http://www.eurospider.com

tion americana, se ha estudiado con detalle la combinación de tesauros específicos (UMLS) y recursos genéricos (EuroWordNet) en recuperación de información multilingüe mediante indexación conceptual. En (Volk et al., 2002), por ejemplo, se demuestra cuantitativamente la utilidad del tesauro UMLS en una tarea de CLIR en dominios médicos, especialmente en combinación con otras fuentes de información semántica.

3.5 El problema de la fusión

Trabajando en un entorno multilingüe la traducción de las consultas no se realiza a un único idioma, sino que deben ser traducidas a todos los idiomas en los cuales estén escritos los documentos, para así poder realizar búsquedas monolingües en cada uno de esos idiomas.

Esto representa un problema a la hora de mostrar al usuario los resultados de las búsquedas, ya que no se tiene una única lista de documentos ordenados por relevancia, sino que se dispone de varias de ellas. El problema de mezclar estas listas en una única se conoce con el nombre de *fusión de listas de documentos* y aún no ha sido resuelto por completo.

Un método bastante simple de fusión consiste en asumir que la relevancia es comparable entre las diferentes colecciones de documentos, por lo que se mezclan las diferentes listas de documentos utilizando su relevancia para ordenarlos (Kwok et al., 1995; Moffat and Zobel, 1995). Este método se conoce con el nombre de *raw scoring*, sin embargo las diferencias entre las colecciones o, incluso los pesos de las diferentes consultas invalidan la asunción de que la relevancia sea comparable entre las distintas colecciones (Voorhees et al., 1995).

Una primera aproximación para que esta medida sea comparable entre las colecciones es realizar una normalización de la relevancia dividiendo por la relevancia máxima obtenida en cada búsqueda. Una variante a este método consiste en restar la relevancia mínima obtenida en cada lista y dividir por la diferencia entre la relevancia máxima y la mínima (Powell et al., 2000). Sin embargo esto soluciona el problema sólo parcialmente, ya que la normalización se realiza de forma independiente en cada una de las listas de documentos provenientes de las distintas colecciones.

Otra forma de realizar la fusión de las listas de documentos es utilizar un algoritmo del tipo *round-robin* y tomar el primer elemento de cada una de las N listas de documentos, ordenarlos según su relevancia, y esos serían los N primeros documentos de la lista fusionada. A continuación se repetiría el proceso con los segundos elementos para obtener los N siguientes documentos y así hasta terminar. Esta solución, sin embargo, adolece del mismo problema: para calcular la posición de un determinado documento sólo se tiene en cuenta la colección a la que pertenece.

En (Martínez-Santiago et al., 2002) se propone una estrategia que tiene en cuenta el peso relativo de cada término de la consulta para realizar una reindexación de los documentos formando una nueva colección multilingüe, sobre la que se realiza una nueva búsqueda empleando los términos originales de la consulta junto con sus traducciones.

Se realizó un experimento de recuperación multilingüe de información involucrando 5 idiomas (alemán, castellano, francés, inglés e italiano), donde se comparó esta estrategia de fusión con otras ya estudiadas como el uso de un algoritmo round-robin o una fusión basada en la normalización de la relevancia. Los resultados mostraron que la estrategia de reindexación obtiene mejores resultados que las otras dos estrategias comparadas.

Otra aproximación consiste en mezclar desde un principio todos los documentos en una única colección. Esto puede hacerse de dos formas distintas:

- sin tener en cuenta su carácter multilingüe (Gey et al., 1999; McNamee and Mayfield, 2002). De esta forma la consulta es traducida a todos los idiomas necesarios y, en lugar de realizar varias búsquedas monolingües, se mezclan todas estas traducciones realizándose una única búsqueda y, por tanto, obteniéndose una única lista. Esta estrategia tampoco parece resolver el problema, ya que los resultados obtenidos son peores que combinando las listas obtenidas por varias búsquedas.

- añadiendo a cada término de indexación una marca identificativa del idioma al que pertenece (Nie, 2002). Por ejemplo, la palabra "*chair*" significa "*carne*" en francés y "*silla*" en inglés. Si estas palabras se marcan como "*chair_f*" y "*chair_e*" en el momento de la indexación, se puede identificar de qué idioma proviene cada una de ellas y no se recuperarían documentos erróneos. Sin embargo los experimentos realizados por (Nie and Jin, 2002) no obtuvieron buenos resultados debido a problemas con los pesos asignados a cada término ya que los métodos de traducción no funcionaron por igual.

4 Otros enfoques

4.1 Traducción de documentos

La traducción de los documentos al idioma en el cual va a realizarse la consulta presenta una serie de ventajas desde el punto de vista teórico (Dumais et al., 1996; Oard, 1998):

- Las traducciones serán más precisas al contar con una mayor información acerca del contexto en el que se utilizan las palabras.
- La degradación de la información que se produce debido a los errores en la traducción afectará en menor medida al proceso de búsqueda si la traducción se realiza sobre los documentos.
- Cuando se está en un entorno multilingüe desaparece el problema que supone realizar la fusión de distintas listas relevantes de documentos.

Sin embargo, en la práctica, el tamaño de la colección de documentos normalmente va a requerir un elevado coste computacional y, posiblemente, una gran cantidad adicional de espacio de almacenamiento, por lo cual esta opción resulta menos práctica que la traducción de las consultas.

En (Oard, 1998) se comprueba de manera práctica que un sistema comercial de traducción automática puede emplear aproximadamente unos diez meses en proporcionar la traducción de $250,000$ documentos. Esto es, a todas luces, inviable para un sistema real de recuperación de información.

Una alternativa es producir una traducción menos precisa que, aunque no sirva para ser leída, si sea suficiente para aplicar sobre ella técnicas de recuperación de información. En este sentido en (Oard et al., 2001) se realiza un experimento (en el marco del CLEF) en el cual los documentos son "traducidos" término a término con la siguiente estrategia llamada *traducción compensada*: si tiene más de una traducción se traduce por sus dos traducciones más frecuentes según el diccionario utilizado, si tiene una única traducción ésta se copia dos veces y si no tiene traducciones (por ejemplo los nombres propios) se copia dos veces el término original en el documento traducido.

El resultado es un documento que contiene dos términos por cada uno de los términos del documento original, de esta forma no se varía la importancia que tiene cada término original. Tras traducir todos los documentos con esta estrategia, las diferentes colecciones documentales traducidas fueron indexadas por separado utilizando Inquery.

Se echan de menos, sin embargo, enfoques intermedios entre la traducción palabra por palabra (demasiado imprecisa) y la traducción automática (demasiado costosa). En general, la posibilidad de traducir los documentos ha recibido mucha menos atención de la que merece. Al fin y al cabo, una vez que el sistema de recuperación encuentra documentos en el idioma destino, es necesario informar al usuario sobre su contenido, y para ello se necesita algún tipo de traducción automática, eficiente y precisa.

4.2 Traducción bidireccional

La traducción de los documentos al idioma de la consulta y la traducción de la consulta al idioma (o idiomas) de los documentos, representan dos enfoques opuestos de combinar las técnicas de recuperación de información con las de traducción automática.

Según (McCarley, 1999) estos dos enfoques no tienen por qué ser mutuamente exclusivos. Pa-

ra comprobarlo realizaron dos experimentos de recuperación translingüe entre francés e inglés (uno en cada sentido).

Se compararon los resultados obtenidos con la traducción de las consultas, la traducción de los documentos y un sistema híbrido que combinó los resultados producidos por ambas aproximaciones de la siguiente forma: la relevancia de un documento es la media de la relevancia obtenida con la traducción de la consulta y la relevancia obtenida con la traducción de los documentos (previa normalización de ambas).

Se observó que las búsquedas que involucraban una traducción en el sentido francés → inglés obtuvieron mejores resultados con independencia de las unidades de traducción (documentos o consultas). Así pues aunque la traducción de los documentos presente ventajas teóricas, éstas van a depender de la calidad de la traducción entre el par de idiomas considerados. Los resultados del sistema híbrido fueron superiores a los de las dos aproximaciones individuales, no influyendo el sentido en el que se realizan las traducciones.

4.3 Indexación conceptual

Otra posibilidad consiste en realizar la traducción tanto de las consultas como de los documentos a un vocabulario de indexación conceptual independiente del idioma. Los documentos se traducen a una representación independiente del idioma en la cual son indexados. Posteriormente se realiza la traducción las consultas a la misma representación y se lleva a cabo la búsqueda.

En (Gilarranz et al., 1997) se propone utilizar los synsets [15] de EuroWordnet como unidades de indexación en un entorno multilingüe. WordNet (Miller, 1990) es una base de datos léxica en inglés que contiene información semántica sobre relaciones entre las diferentes palabras que la componen. Partiendo de WordNet se desarrolló EuroWordnet (Vossen, 1998) que contiene información sobre las relaciones semánticas entre palabras de diversos idiomas europeos, así como relaciones multilingües entre los conceptos de los diferentes idiomas. La indexación en términos del Índice Interlingua de EuroWordnet (ILI) presenta las siguientes ventajas técnicas:

- Al tener un único espacio de índices para todos los idiomas, se evita el problema de la fusión de resultados de búsqueda para cada idioma destino.
- Es más escalable que los enfoques de traducción cuando crece el número de idiomas.
- Utiliza la desambiguación semántica automática para resolver de forma directa algunos de los problemas tradicionales de las palabras como índices de búsqueda, como la identificación de términos sinónimos, la diferenciación de los distintos sentidos de una palabra, etc.
- Permite sar las relaciones conceptuales de EuroWordNet para expandir las consultas (conceptos más genéricos, más específicos, partes de, etc.).

En (Diekema et al., 1999) se realizó un experimento de recuperación translingüe entre inglés (utilizando los synsets de WordNet) y francés (previa construcción de una base de datos léxica equivalente a WordNet, pero en francés). Los datos parecen indicar que las consultas realizadas con indexación conceptual alcanzan, en muchas ocasiones, la misma precisión que las búsquedas monolingües. Sin embargo debido a errores en la implementación del sistema los resultados globales parecen indicar justo lo contrario.

En (Ruiz et al., 2000) se continúa el experimento anterior, corrigiéndose los errores detectados. Las búsquedas translingües (entre los mismos idiomas francés e inglés) logran una precisión del 75%, en ambas direcciones, de la precisión alcanzada por las búsquedas monolingües correspondientes.

El Proyecto ITEM [16] exploró la viabilidad de integrar diferentes técnicas de procesamiento del lenguaje natural en un motor de búsqueda

[15] Un *synset* representa un concepto y, para cada idioma, contiene una serie de palabras que hacen referencia a dicho concepto.

[16] Proyecto ITEM: recuperación de Información Textual en un Entorno Multilingüe (CICyT TIC96-1243-C03-01)

de información en un entorno multilingüe. Empleando el ILI de EuroWordnet para indexar tanto los documentos de los diferentes idiomas como las consultas se puede realizar una búsqueda a nivel conceptual y de forma independiente del idioma.

La experiencia con este motor de búsqueda indicó que la indexación conceptual tiene ciertas ventajas respecto a las aproximaciones basadas en traducción, además de las ya citadas: por ejemplo, la expansión automática de las consultas empleando las relaciones de EuroWordnet permite traducir conceptos sin una representación directa en el otro idioma. Por ejemplo "*grand jury*" en inglés no tiene un concepto equivalente en castellano. Sin embargo un hiperónimo suyo es "*jury*" que tiene una traducción directa como "*jurado*". Así se puede paliar la pérdida de información que implica el proceso de traducción.

Sin embargo, se comprobaron también los problemas del enfoque: por un lado, las técnicas de desambiguación automática no han alcanzado todavía un grado suficiente de madurez para un enfoque tan ambicioso (Senseval-2, 2001). Por otro lado, es difícil encontrar el nivel de representación conceptual adecuado para la tarea. Las expresiones multipalabra, que pueden representar conceptos complejos, no son adecuadas para la recuperación monolingüe, en la que es mejor indexar todos los componentes (ver sección 2.2). Sin embargo, este tipo de expresiones son mucho mejores que las palabras individuales a la hora de traducir. De esta forma, si una expresión multipalabra está en EuroWordNet, puede ser traducida adecuadamente a otros idiomas, pero sus componentes no serán indexados. Y si no está en EuroWordnet, los conceptos que la forman no serán representativos en los idiomas destino. Este problema deriva del hecho de considerar la indexación y la traducción como una única tarea.

5 Interactividad en la Recuperación de Información Multilingüe

Un sistema automático de búsqueda translingüe es sólo un componente de un proceso completo de búsqueda y utilización de la información. Desde la perspectiva de los usuarios no sirve de nada que un sistema translingüe de recuperación de información recupere con mayor o menor precisión documentos, por ejemplo, en chino, si el usuario no es capaz de reconocer aquellos que le interesan (Oard, 2001), ni refinar su búsqueda global a partir de resultados que no comprende.

En general, los estudios realizados sobre sistemas translingües o multilingües de recuperación de información no han considerado la interactividad con el usuario como pieza fundamental de diseño, por este motivo la investigación en este campo está aún en sus comienzos.

En esta sección vamos a ver los diferentes trabajos que se han realizado en este campo. Las primeras investigaciones (apartados 5.1 al 5.5) emplearon metodologías de trabajo diferentes entre sí, hasta la llegada del `iCLEF` (apartado 5.6), donde se ha proporcionado una infraestructura y una metodología específicas para la realización de experimentos interactivos de recuperación translingüe de información.

5.1 Los trabajos iniciales de Oard y Resnik

Oard y Resnik realizaron una de las primeras investigaciones sobre la usabilidad de un sistema translingüe de búsqueda de información de cara al usuario (Resnik, 1997; Oard and Resnik, 1999): realizar una traducción palabra por palabra al inglés de una serie de documentos en japonés y comprobar si una serie de usuarios era capaz de clasificar los documentos traducidos agrupándolos por temas similares.

Se comprobó que los usuarios eran capaces de realizar esta tarea de una manera mucho más efectiva que un clasificador automático, pero con menor precisión que otros usuarios que examinaron unas traducciones perfectas en inglés de los documentos originales.

Posteriormente en (Taylor and White, 1998) se propone utilizar un sistema de traducción automática para realizar la misma tarea, aunque no se realiza ninguna evaluación al respecto.

Oard y Resnik proponen separar los procesos

de búsqueda y utilización de la información en un entorno multilingüe. El sistema debe proporcionar inicialmente al usuario la capacidad de expresar su necesidad de información en su propio idioma y, con su ayuda, trasladar ésta al idioma en el cual se encuentran los documentos:

- utilizando un diccionario que contenga la definición de los términos en el propio idioma del usuario.
- traducir cada término de la consulta y mostrar al usuario las traducciones inversas de cada posible traducción, para proporcionarles información acerca de los contextos en los que dicha traducción puede ser utilizada en su propio idioma.

Una vez que el usuario ha comunicado al sistema su necesidad de información y ésta se encuentra expresada en el idioma de los documentos, el sistema puede realizar una búsqueda automática. La información contenida en los documentos así recuperados habrá de ser mostrada al usuario en su propio idioma. La visualización de esta información debe cumplir, según los autores, dos tareas fundamentales:

- Facilitar los juicios de relevancia al usuario, de manera que no le resulte excesivamente complicado decidir si un determinado documento le puede, o no, ser útil.
- Proporcionar nuevo vocabulario en el idioma de los documentos para que el usuario pueda refinar su búsqueda.

Con una metodología de evaluación consistente en medir la precisión y la cobertura sobre la selección manual realizada por los usuarios se propone medir la eficacia de estos sistemas en ayudar a sus usuarios a realizar búsquedas translingües de información.

5.2 MULINEX

El proyecto MULINEX (Erbach et al., 1997) se presenta como el primer sistema interactivo completo de búsqueda translingüe. En él se desarrolló una interfaz de recuperación translingüe de información inicialmente en Alemán, Francés e Inglés, pero fácilmente aplicable a otros idiomas.

Este sistema tenía una interfaz de traducción de la consulta y ofrecía resúmenes de los documentos que podían ser traducidos al idioma elegido.

Una vez completada la interfaz, se llevaron a cabo diversos estudios (Capstick et al., 1998b; Capstick et al., 1998a) para determinar la manera más eficiente de presentar los resultados. El sistema de búsqueda translingüe demostró funcionar perfectamente. Sin embargo, debido a que los usuarios que intervinieron en estas evaluaciones generalmente tenían conocimiento acerca de los diferentes idiomas sobre los que se realizaban las búsquedas, las características de traducción del sistema apenas fueron utilizadas.

5.3 Los trabajos de Ogden y Davis

Al igual que en los trabajos de Oard y Resnik estos autores proponen la separación entre los procesos de búsqueda y la posterior utilización de la información, concentrándose en dos tareas fundamentales de cara al usuario: ayudarle a expresar su consulta en otros idiomas y presentarle documentos en el suyo propio.

En (Ogden et al., 2000) se describe un experimento realizado dentro del marco del TREC 8 en el que con la ayuda de un interfaz un usuario monolingüe de habla inglesa es capaz de construir consultas en Italiano, Francés y Alemán (idiomas que le resultaban totalmente desconocidos).

A la vista de los resultados se comprobó que las consultas producidas por el usuario monolingüe alcanzaban hasta un 85% de la precisión obtenida con las consultas traducidas manualmente (salvo en Alemán, donde se llegaba a un 70%).

Este experimento demostró que un usuario monolingüe es capaz de expresar su consulta en otro idioma con una precisión bastante razonable asistido por una interfaz de ayuda a la traducción.

En (Ogden et al., 1999a) se realiza un experimento en el que un usuario monolingüe traduce sus consulta del inglés al alemán utilizando

diccionarios on-line. Los documentos recuperados le son mostrados utilizando una interfaz de thumbnails (Ogden et al., 1999c). Por último los diez primeros documentos son traducidos al inglés utilizando BabelFish [17] y sobre esas traducciones el usuario seleccionó aquellos que le parecieron relevantes. La precisión obtenida en esta selección manual fue del 86%.

Basándose en todos estos resultados, los autores desarrollaron un prototipo denominado Keizai (Ogden et al., 1999b) con el que un usuario de habla inglesa puede recuperar textos en Japonés y Coreano.

5.4 Selección Documental interactiva entre Inglés y Japonés

En (Suzuki et al., 2001) se presenta un estudio sobre la selección interactiva de documentos en un sistema translingüe de recuperación de información entre los idiomas Inglés y Japonés.

Los autores realizaron dos experimentos interactivos: en el primero comprobaron la habilidad de los usuarios para juzgar la relevancia sobre una traducción palabra por palabra, comprobando que eran capaces de emitir juicios de relevancia sobre estas traducciones con bastante precisión, en consonancia con los resultados obtenidos por Oard y Resnik. En el segundo estudiaron un método de mostrar los documentos recuperados consistente en resumir traducciones automáticas de los documentos (a un tamaño del 30% del original). En esta ocasión la precisión alcanzada por los usuarios fue menor que en el primer experimento.

Lamentablemente las personas que participaron en ambos experimentos no fueron las mismas, por lo que no se puede concluir con certeza que los resúmenes de las traducciones sean menos efectivos que las traducciones palabra por palabra.

5.5 WebSite Term Browser

En (Peñas, 2002) se presenta un sistema de navegación con sintagmas multilingüe. Se parte de una serie de colecciones documentales que son previamente procesadas con técnicas superficiales de tratamiento del lenguaje natural, extrayendo de ellas sintagmas nominales, con los que son indexados los documentos.

Cuando el usuario introduce la consulta, el sistema busca aquellos sintagmas que estén más relacionados con los términos de la consulta, en cualquiera de los idiomas indexados (castellano, catalán, francés, inglés e italiano). Para ello se expande cada término mediante EuroWordNet y diccionarios bilingües; la restricción de coocurrencia dentro de cada sintagma filtra el posible ruido producido por la expansión. Después, el sistema busca los documentos que contengan los sintagmas encontrados, mostrándolos por separado en cada idioma. A diferencia de las experiencias anteriores, en este caso se presupone que el usuario tiene cierto conocimiento pasivo de los idiomas de búsqueda, y puede reconocer expresiones de búsqueda útiles en otros idiomas y usar los documentos encontrados.

La evaluación del sistema se realizó diseñando una interfaz de búsqueda que presentaba, a la vez, los sintagmas sugeridos por WTB y los resultados devueltos por Google para la consulta del usuario. Se indexó el dominio de una universidad, y se permitió la utilización libre del sistema para profesores, alumnos y cualquier otro usuario de la web. El análisis de más de mil sesiones de búsqueda reales reveló que los sintagmas sugeridos por WTB se utilizaban tan a menudo como los resultados devueltos por Google, lo que indica la utilidad de la información de sintagmas como complemento a las tradicionales listas de documentos. Por desgracia, la utilidad de los aspectos estrictamente translingües del enfoque hubo de medirse por otras vias indirectas (como la comparación entre el proceso de extracción de términos con tesauros preexistentes en dominios educativos), porque los usuarios que accedieron al sistema se limitaron casi exclusivamente a preguntar y leer información en castellano.

5.6 iCLEF: un foro de evaluación de sistemas interactivos translingües

El CLEF (Cross-Language Evaluation Forum) se ha ocupado de desarrollar y mantener una infraestructura para la evaluación de sistemas de

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

recuperación de información sobre idiomas europeos (Peters, 2002b). Para ello se han creado una serie de datos reutilizables con el fin de medir las características de estos sistemas de recuperación de información.

Los datos proporcionados por el CLEF consisten en:

- Colecciones documentales en varios idiomas europeos. Todas ellas formadas por noticias publicadas en diversos medios de comunicación en el año 1994 o 1995.
- Una serie de consultas sobre diversos temas que están tratados en las colecciones documentales, expresadas en una amplia variedad de idiomas que no sólo cubren aquellos en los que se encuentran escritos los documentos.
- Juicios de relevancia nativos para todas las consultas en cada uno de los idiomas contemplados.

Aprovechando todos estos datos se organiza un "track" especialmente destinado a la evaluación de diferentes aspectos interactivos en la recuperación de información translingüe.

En dicho track cada sistema presentado se compara con un sistema de contraste, mezclando las consultas, sistemas y usuarios con un diseño de cuadrados latinos para minimizar los efectos que una determinada combinación consulta-sistema pudiera tener sobre los resultados.

5.6.1 iCLEF'2001: Selección documental

El propósito de los experimentos realizados en el marco de la primera edición del iCLEF (Oard and Gonzalo, 2002) fue investigar en nuevas formas de presentación de documentos escritos en un idioma que el usuario desconoce o no es capaz de leer con fluidez.

En esta primera edición del iCLEF se presentaron tres prototipos diferentes de presentación de documentos, los cuales fueron comparados con una traducción automática proporcionada por Systran Professional 3.0.

En (Wang and Oard, 2002) podemos ver la descripción del sistema presentado por la Universidad de Maryland. En este prototipo se comparaban las traducciones ofrecidas por el Systran con un sistema de traducción palabra por palabra donde se cada palabra original en francés era substituida por aquella de sus traducciones que fuera más frecuente en el *Brown Corpus*.

En (Bathie and Sanderson, 2002) se describe el experimento llevado a cabo por la Universidad de Sheffield. En dicho experimento se llevó a cabo una comparación entre la selección documental monolingüe y translingüe. Los evaluadores seleccionaron documentos en francés (previamente traducidos por el Systran) en uno de los sistemas, y en el otro seleccionaron documentos en inglés sin mediar traducción alguna.

En (López-Ostenero et al., 2002b) se describe el experimento llevado a cabo por la UNED. En dicho experimento se aprovechan los sintagmas nominales extraidos con el WebSite Term Browser para realizar un resumen translingüe de los documentos basándose en dichos sintagmas.

A la vista de los resultados obtenidos por estos tres experimentos se pueden sacar las siguientes conclusiones:

- Todos los experimentos obtienen una medida de precisión superior a la obtenida siguiendo la estrategia de marcar todos los documentos como relevantes. Esto demuestra que los sistemas empleados ayudan al usuario a juzgar los documentos.
- La selección documental monolingüe obtiene mejores resultados de precisión y cobertura que la selección documental utilizando las traducciones del Systran.
- Para los documentos en francés se ve que las traducciones proporcionadas por el Systran ofrecen más ayuda al usuario que las realizadas palabra por palabra.
- El sistema de traducción mediante sintagmas es un punto intermedio de complejidad entre la traducción palabra por palabra y la traducción ofrecida por Systran. Sin embargo obtiene resultados cualitativamente mejores según las medidas oficiales del iCLEF que estas dos aproximaciones.

5.6.2 iCLEF'2002: Búsquedas interactivas

El objetivo del iCLEF'2002 (Gonzalo and Oard, 2002) fue proporcionar un marco de referencia común para realizar experimentos comparando dos sistemas de recuperación de información translingüe que permitan a un usuario que desconoce el idioma de los documentos realizar una expansión interactiva de la consulta, una selección interactiva de documentos (al igual que el año anterior), o ambas opciones a la vez.

Cinco fueron, en esta ocasión, los sistemas que se presentaron, aunque sólo tres de ellos permitían realizar una búsqueda translingüe completa:

En (Petrelli et al., 2002) se presenta el experimento llevado a cabo para el iCLEF'2002 por la Universidad de Sheffield. El prototipo participante se enmarca dentro del proyecto Clarity, que pretende el desarrollo de un sistema de recuperación translingüe de información entre idiomas para los que se disponen de escasos recursos de traducción. En este caso los documentos estaban escritos en finlandés, mientras que las consultas se realizaban en inglés.

El propósito fundamental de este experimento fue el comprobar si la traducción de la consulta por parte del usuario debía ser considerada una tarea del proceso de búsqueda o si, por el contrario, resulta más beneficioso ocultar el proceso de traducción al usuario.

Debido a diversos errores, los resultados cuantitativos no permiten establecer conclusiones, aunque sí es destacable el hecho de que los usuarios del sistema manifestaron no encontrarse cómodos seleccionando traducciones para términos individuales.

En (He et al., 2002) tenemos la descripción del experimento interactivo presentado por la Universidad de Maryland en el que se buscaron documentos en alemán utilizando consultas en inglés. Los dos sistemas comparados fueron:

1. Traducción automática de la consulta: utilizando la estructura propuesta por (Pirkola, 1998) para la traducción de las consultas (ver sección 3.1) a partir de todas las traducciones posibles para cada término.
2. Traducción de la consulta con la asistencia del usuario: la tarea de seleccionar las traducciones adecuadas de cada palabra recae en los usuarios.

Para que los usuarios tuvieran información con la que seleccionar las diferentes traducciones se les proporcionaron diversas pistas, como las traducciones inversas de cada término o frases (extraídas de corpora paralelo) que mostraban el contexto en el que cada término original podía recibir una traducción determinada.

Los resultados muestran que la traducción asistida por el usuario obtiene una mayor eficiencia que la traducción automática estructurada. Una conclusión adicional a la que llegaron los autores es que, en general, las pistas ofrecidas para seleccionar de entre las posibles traducciones resultan ser de ayuda para los usuarios, aunque el grado de utilidad varía según la consulta, la colección y los recursos de traducción disponibles.

De igual modo sugieren que utilizar los mismos recursos de traducción para la traducción de las consultas y de los documentos, ofrecerá una mayor ayuda al usuario de cara a la reformulación de la consulta.

En (López-Ostenero et al., 2002a) se amplía el sistema presentado en la anterior edición dotándole de la posibilidad de especificar una consulta seleccionando sintagmas nominales que estuvieran relacionados con la misma. La traducción mediante sintagmas de los documentos permitía a los usuarios realizar, de manera sencilla, una expansión de la consulta sin más que añadir a la misma aquellos sintagmas de los documentos que el usuario considerase relevantes.

Este sistema se comparó con un interfaz de ayuda a la traducción de términos individuales contenidos en la consulta, mostrandole al usuario las traducciones inversas de cada posible término de traducción.

El uso de sintagmas nominales para hacer la traducción obtuvo mejores resultados no sólo en la búsqueda, sino también de cara a los usuarios. Los usuarios interactúan con sintagmas en su propio idioma, dejando el proceso de traducción al sistema. Al igual que en (Petrelli et al., 2002) los usuarios se manifestaron en

contra de tener que seleccionar interactivamente los términos de traducción.

La discriminación de las traducciones utilizando un criterio de coocurrencia estadística demuestra ser más efectiva que la selección interactiva de traducciones cuando el usuario no entiende el idioma en el que se encuentran escritos los documentos.

Los resultados apoyan, además, la hipótesis planteada en (He et al., 2002), ya que el uso del mismo recurso de traducción (los sintagmas nominales) para la consulta y para los documentos aportó una mayor ayuda al usuario en el proceso de refinamiento de la consulta.

6 Recapitulación

La necesidad de realizar búsquedas translingües es un hecho, y la demanda de este tipo de búsquedas aumentará en los próximos años con el crecimiento de la Web.

Los experimentos realizados han demostrado que la recuperación translingüe es perfectamente realizable y con un nivel de eficiencia cercano a una búsqueda monolingüe. La tarea de obtener una lista de documentos en un mismo idioma ordenada según la relevancia que tengan para una consulta escrita en un idioma diferente, ya ha sido básicamente resuelta (Oard, 2002), aunque la eficiencia de los sistemas depende de la pareja de idiomas que se considere.

Sin embargo aún quedan diversos problemas a los que se debe dirigir la investigación. Algunos de ellos se mencionaron en el Workshop "CLIR: a research roadmap" en el ámbito del SIGIR'2002:

- **Dominio**: la mayoría de las técnicas empleadas han sido probadas sólo sobre noticias de periódicos (en las colecciones TREC, CLEF y NTCIR) y no se sabe si serán efectivas fuera de él.

- **Eficiencia**: el coste computacional que supone una traducción adecuada de las consultas puede resultar excesivo para un entorno real de búsqueda, aparte que la calidad de las traducciones aún no es óptima.

- **Unificación**: actualmente los sistemas de recuperación translingüe de información presentan dos claras separaciones:

  - **Traducción y búsqueda**: los procesos de traducción y búsqueda se realizan, normalmente, por separado. De esta forma la incertidumbre de las traducciones no influye en el proceso de búsqueda.
  - **Diferentes idiomas**: cuando se realiza una búsqueda multilingüe, el problema de fusionar los resultados de cada una de las búsquedas monolingües en una única lista ordenada aún no ha sido resuelto.

  En (Nie, 2002) se propone la creación de un único modelo de forma que integre estas diferencias y se pueda abordar la recuperación multilingüe de información de una manera similar a la recuperación monolingüe.

- **Interacción**: los usuarios reales de los sistemas de búsqueda están interesados en la información contenida en los documentos, no en la lista ordenada que proporcionan los sistemas.

Por último, la capacidad de encontrar documentos en varios idiomas a partir de una única consulta debe todavía combinarse con otras aplicaciones y técnicas, como la Extracción de Información o la Búsqueda de Respuestas, en el camino a un *Acceso a la información multilingüe* sin limitaciones. Aunque se haya avanzado mucho en algunos aspectos puntuales del problema en los últimos años, lo cierto es que aún queda un largo camino hasta que el manejo combinado de fuentes en distintos idiomas sea algo natural y rutinario para el usuario tipo de un sistema de información.

Agradecimientos

Este trabajo ha sido financiado parcialmente por la Comisión Interministerial de Ciencia y Tecnología, proyecto Hermes (TIC2000-0335-C03-01).

Referencias

Abu-Salem, H., Al-Omari, M., and Evens, M. (1999). Stemming methodologies over individual queries words for an Arabian information retrieval system. *JASIS*, 50:524–529.

Bacchin, M., Ferro, N., and Melucci, M. (2002). University of Padua at CLEF 2002: Experiments to evaluate a statistical stemming algorithm. In *Proceedings of CLEF 2002*.

Ballesteros, L. (2000). Cross Language Retrieval via transitive translation. In Croft, W. B., editor, *Advances in Information Retrieval: Recent Research from the CIIR*, pages 203–234. Kluwer Academic Publishers.

Ballesteros, L. and Croft, W. B. (1996). Dictionary Methods for Cross-Lingual Information Retrieval. In *Database and Expert Systems Applications*, pages 791–801.

Ballesteros, L. and Croft, W. B. (1997). Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval. In *Research and Development in Information Retrieval*, pages 84–91.

Ballesteros, L. and Croft, W. B. (1998). Resolving Ambiguity for Cross-Language Information Retrieval. In *Proceedings of the SIGIR'98*, pages 64–71.

Bathie, Z. and Sanderson, M. (2002). iCLEF at Sheffield. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *LNCS*, pages 336–354. Springer.

Boughanem, M., Chrisment, C., and Nassr, N. (2002). Investigation on Disambiguation in CLIR Aligned Corpus and Bi-directional Translation-Based Strategies. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *LNCS*, pages 158–168. Springer.

Braschler, M., Kan, M., Schuble, P., and Klavans, J. (2000a). The Eurospider Retrieval System and the TREC-8 Cross-Language Track. In *Proceedings of TREC8*, pages 367–376. NIST, Gaithesburg, MD.

Braschler, M., Krause, J., Peters, C., and Schäuble, P. (1999). Cross-Language Information Retrieval (CLIR) Track Overview. In *Proceedings of TREC7*, pages 25–32. NIST, Gaithesburg, MD.

Braschler, M., Peters, C., and Schäuble, P. (2000b). Cross-Language Information Retrieval (CLIR) Track Overview. In *Proceedings of TREC8*, pages 25–34. NIST, Gaithesburg, MD.

Braschler, M. and Schäuble, P. (1998). Multilingual Information Retrieval Based on Document Alignment Techniques. In Nikolau, C. and Stephanidis, C., editors, *Research and Advanced Technology for Digital Libraries, Second European Conference ECDL'98*, pages 183–197.

Braschler, M. and Schäuble, P. (2001). Experiments with the eurospider retrieval system for clef 2000. In *Proc. CLEF 2000*. Springer-Verlag.

Brown, R. D. (1996). The Pangloss-Lite Machine Translation System. In *Expanding MT Horizons: Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas*, pages 268–272.

Brown, R. D. (1997). Automated Dictionary Extraction for Knowledge-Free Example-Based Translation. In *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation*.

Buckley, C., Salton, G., Allan, J., and Singhal, A. (1995). Automatic Query Expansion Using SMART: TREC 3. In *Proceedings of TREC3*, pages 69–80. NIST, Gaithesburg, MD.

Callan, J., Croft, W., and Harding, S. (1992). The Inquery Retrieval System. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 78–83. Springer-Verlag.

Capstick, J., Diagne, A. K., Erbach, G., and Uszkoreit, H. (1998a). MULINEX: Multilingual Web Search and Navigation. In *Industrial Applications of Natural Language Procesing*.

Capstick, J., Erbach, G., and Uszkoreit, H. (1998b). Design and Evaluation of a Psychological Experiment on the Effectiveness of Document Summarisation for the Retrieval of Multilingual WWW Documents. In *Working Notes of the AAAI Spring symposium Intelligent Text Summarisation*.

Carbonell, J. G., Yang, Y., Frederking, R. E., Brown, R. D., Geng, Y., and Lee, D. (1997). Translingual Information Retrieval: A Comparative Evaluation. In *IJCAI (1)*, pages 708–715.

Carl, M. and Hansen, S. (1999). Linking Translation Memories with Example-Based Machine Translation.

Chen, A., Gey, F., Kishida, K., Jiang, H., and Liang, Q. (1999). Comparing multiple methods for Japanese and Japanese-English text retrieval. In *Proceedings of the First NTCIR Workshop*, pages 49–58.

Chen, A. and Gey, F. C. (2001). Translation Term Weighting and Combining Translation

Resources in Cross-Language Retrieval. In *Proceedings of TREC10*. NIST, Gaithesburg, MD.

Chen, J. and Nie, J.-Y. (2000). Parallel Web Text Mining for Cross-Language IR. In *Proceedings of RIAO 2000 conference*.

Chen, S. F. (1993). Aligning Sentences in Bilingual Corpora using Lexical Information. In *Meeting of the Association for Computational Linguistics*, pages 9–16.

Collins, B. (1999). *Example-Based Machine Translation: An Adaptation-Guided Retrieval Approach*. PhD thesis, Trinity College, Dublin.

Davis, M. (1997). New Experiments in Cross-Language Text Retrieval at NMSU's Computing Research Lab. In *Proceedings of TREC5*, pages 447–454. NIST, Gaithesburg, MD.

Dawson, J. (1974). Suffix removal and word conflation. *ALLC bulletin*, 2:33–46.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Diekema, A., Oroumchian, F., Sheridan, P., and Liddy, E. D. (1999). TREC-7 Evaluation of Conceptual INterlingua DOcument Retrieval (CINDOR) in English and French. In *Proceedings of TREC7*, pages 169–180. NIST, Gaithesburg, MD.

Dumais, S., Landauer, T., and M.L.Littman (1996). Automatic Cross-Linguistic information retrieval using latent semantic indexing. In *SIGIR'96 Workshop on Cross-Linguistic Information Retrieval*.

Erbach, G., Neumann, G., and Uszkoreit, H. (1997). MULINEX: Multilingual Intexing, Navigation and Editing Extensions for the World-Wide Web. In Hull, D. and Oard, D., editors, *AAAI Symposium on Cross-Language Text and Speech Retrieval*.

EUROVOC (1995). Thesaurus EUROVOC: Vol 1-3 / European Communities. Luxembourg: Office for Official Publications of the European Communities.

Figuerola, C. G., Gómez, R., Rodríguez, A. F. Z., and Berrocal, J. L. A. (2002). Spanish Monolingual Track: The Impact of Stemming on Retrieval. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *LNCS*, pages 253–261. Springer.

Fukushima, T. and Akamine, S. (1999). A character-based indexing and word-based ranking method for Japanese text retrieval. In *Proceedings of the First NTCIR Workshop*, pages 179–182.

Fung, P. (1995). Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus. In *Proceedings of the 3rd Workshop on Very Large Corpora*.

Gale, W. A. and Church, K. W. (1991). A Program for Aligning Sentences in Bilingual Corpora. In *Meeting of the Association for Computational Linguistics*, pages 177–184.

Gey, F., Jiang, H., Chen, A., and Larson, R. (1999). Manual Queries and Machine Translation in Cross-Language Retrieval and Interactive Retrieval with Cheshire II at TREC-7. In *Proceedings of TREC7*, pages 527–540. NIST, Gaithesburg, MD.

Gey, F. C. and Oard, D. W. (2001). The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic using Englis, French or Arabic Queries. In *Proceedings of TREC10*. NIST, Gaithesburg, MD.

Gilarranz, J., Gonzalo, J., and Verdejo, F. (1997). Language-independent text retrieval with the EuroWordNet Multilingual Semantic Database. In *Proceedings of the Workshop on Multilinguality in Software Industry: the AI contribution, in IJCAI'97 (International Joint Conference on Artificial Intellligence)*.

Gollins, T. and Sanderson, M. (2001). Sheffield University CLEF 2000 Submission - Bilingual Track: German to English. In Peters, C., editor, *Cross-Language Information Retrieval and Evaluation, Workshop, CLEF 2000*, volume 2069 of *LNCS*, pages 245–252. Springer.

Gonzalo, J. and Oard, D. W. (2002). The CLEF 2002 Interactive Track. In *Proceedings of CLEF 2002*.

Grefenstette, G. (1998). *The problem of Cross-Language Information Retrieval*, chapter in Cross-Language Information Retrieval. Kluwer Academic Publishers.

He, D., Wang, J., Oard, D., and Nossal, M. (2002). Comparing User-assisted and Automatic Query Translation. In *Proceedings of CLEF 2002*.

Hersh, W., Buckley, C., Leone, T., and Hickman, D. (1994). Oshumed: an interactive retrieval evaluation and new large text collection for research. In *Proceedings of SIGIR'94*, pages 192–201.

Hull, D. A. and Grefenstette, G. (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, pages 49–57.

Jones, G. F. and Lam-Adesina, A. M. (2002). Exeter at CLEF 2001: Experiments with Machine Translations for Bilingual Retrieval. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *LNCS*, pages 59–77. Springer.

Kalamboukis, T. (1995). Suffix stripping with modern Greek. *Program*, 29:313–321.

Kay, M. and Röscheisen, M. (1993). Text-translation alignment. *Computational Linguistics*, 19(1):121–142.

Kraaij, W. (2002). TNO at CLEF-2001: Comparing Translation Resources. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *LNCS*, pages 78–93. Springer.

Kraaij, W. and Pohlmann, R. (1994). Porter's stemming algorithm for Dutch. In Noordman, L. and de Vroomen, W., editors, *Informatiewetenschap, Tilburg, STINFON*.

Kraaij, W. and Pohlmann, R. (1998). Comparing the effecto of syntactic vs. statistical phrase index strategies for Dutch. In *Proceedings ECDL'98*, pages 605–617.

Kwok, K. (1997). Comparing representations in Chinese information retrieval. In *Proceedings of SIGIR'97*, pages 34–41.

Kwok, K. L., Grunfeld, L., and Lewis, D. D. (1995). TREC-3 Ad-Hoc, Routing Retrieval and Thresholding Experiments using PIRCS. In *Proceedings of TREC3*, pages 247–256. NIST, Gaithesburg, MD.

Lee, J. H. and Ahn, J. S. (1996). Using n-grams for corean text retrieval. In *Proceedings of SIGIR'96*, pages 216–224.

Loukachevitch, N. V. and Dobrov, B. V. (2000). Thesaurus-Based Structural Thematic Summary in Multilingual Information Systems. *Machine Translation Review*, 11:10–20.

Loukachevitch, N. V. and Dobrov, B. V. (2002). Cross-Language Information Retrieval Based on Multilingual Thesauri Specially Created for Automatic Text Processing. In *Proceedings of Workshop on Cross-Language Information Retrieval: A Research RoadMap. SIGIR 2002*.

Lovins, J. (1968). Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31.

López-Ostenero, F., Gonzalo, J., Peñas, A., and Verdejo, F. (2002a). Interactive Cross-Language Searching: phrases are better than terms for query formulation and refinement. In *Proceedings of CLEF 2002*.

López-Ostenero, F., Gonzalo, J., Peñas, A., and Verdejo, F. (2002b). Noun phrase translations for Cross-Language Document Selection. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *LNCS*, pages 320–331. Springer.

Martínez-Santiago, F., Martín, M., and Ureña, A. (2002). SINAI on CLEF 2002: Experiments with merging strategies. In *Proceedings of CLEF 2002*.

Mayfield, J., McNamee, P., Costello, C., Piatko, C., and Banerjee, A. (2001). JHU/APL at TREC 2001: Experiments in Filtering and in Arabic, Video, and Web Retrieval. In *Proceedings of TREC10*. NIST, Gaithesburg, MD.

McCarley, J. S. (1999). Should we Translate the Documents or the Queries in Cross-language Information Retrieval? In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 208–214. Association for Computational Linguistics.

McNamee, P. and Mayfield, J. (2001). A Language-Independent Approach to European Text Retrieval. In Peters, C., editor, *Cross-Language Information Retrieval and Evaluation, Workshop, CLEF 2000*, volume 2069 of *LNCS*, pages 129–139. Springer.

McNamee, P. and Mayfield, J. (2002). JHU/APL Experiments at CLEF: Translation Resources and Score Normalization. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *LNCS*, pages 193–208. Springer.

Miller, G. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography 3(4)*.

Moffat, A. and Zobel, J. (1995). Information Retrieval Systems for Large Document Collections,. In *Proceedings of TREC3*, pages 85–93. NIST, Gaithesburg, MD.

Monz, C. and de Rijke, M. (2002). Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German and Italian. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *LNCS*, pages 262–277. Springer.

National Library of Medicine (1997). Unified Medical Language System (UMLS) Knowledge Sources, 6th experimental edition.

Nie, J.-Y. (1999). TREC-7 CLIR using a Probabilistic Translation Mode. In *Proceedings of*

TREC7, pages 547–554. NIST, Gaithesburg, MD.

Nie, J.-Y. (2002). Towards a Unified Approach to CLIR and Multilingual IR. In *Proceedings of Workshop on Cross-Language Information Retrieval: A Research RoadMap. SIGIR 2002*.

Nie, J.-Y. and Jin, F. (2002). Merging Different Languages in a Single Document Collection. In *Proceedings of CLEF 2002*.

Nie, J.-Y., Simard, M., and Foster, G. (2001). Multilingual Information Retrieval Based on Parallel Texts from the Web. In Peters, C., editor, *Cross-Language Information Retrieval and Evaluation, Workshop, CLEF 2000*, volume 2069 of *LNCS*, pages 188–201. Springer.

Nirenburg, S., Beale, S., and Domashnev, C. (1994). A Full-Text Experiment in Example-Based Machine Translation. In *Proceedings of the International conference on New Methods in Language Processing*, pages 78–87.

Oard, D. W. (1998). A comparative study of query and document translation for cross-language information retrieval. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas*.

Oard, D. W. (2001). Evaluating Interactive Cross-Language Information Retrieval: Document Selection. In Peters, C., editor, *Cross-Language Information Retrieval and Evaluation, Workshop, CLEF 2000*, volume 2069 of *LNCS*, pages 57–71. Springer.

Oard, D. W. (2002). When You Come to a Fork in the Road, Take It: Multiple Futures for CLIR Research. In *Proceedings of Workshop on Cross-Language Information Retrieval: A Research RoadMap. SIGIR 2002*.

Oard, D. W. and Gonzalo, J. (2002). The CLEF 2001 interactive track. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *LNCS*, pages 308–319. Springer.

Oard, D. W., Levow, G.-A., and Cabezas, C. I. (2001). CLEF Experiments at Maryland: Statistical stemming and backoff translation. In Peters, C., editor, *Cross-Language Information Retrieval and Evaluation, Workshop, CLEF 2000*, volume 2069 of *LNCS*, pages 176–187. Springer.

Oard, D. W. and Resnik, P. (1999). Support for interactive document selection in cross-language information retrieval. *Information Processing and Management*, 35(3):363–379.

Ogden, W., Cowie, J., Davis, M., Ludovic, E., Molina-Salgado, H., and Shin, H. (1999a). Getting Information from Documents You Cannot Read: An interactive Cross-Language Text Retrieval and Summarization System. In *Joint ACM DL/SIGIR Workshop on Multilingual Information Discovery and Access*.

Ogden, W., Cowie, J., Davis, M., Ludovic, E., Nirenburg, S., Molina-Salgado, H., and Sharples, N. (1999b). Keizai: An Interactive Cross-Language Text Retrieval System. In *Proceeding of the MT SUMMIT VII Workshop on Machine Translation for Cross Language Information Retrieval*.

Ogden, W., Cowie, J., Ludovik, E., Molina-Salgado, H., Niremburg, S., Sharples, N., and Sheremtyeva, S. (2000). CRL's TREC-8 Systems Cross-Lingual IR, and Q&A. In *Proceedings of TREC8*, pages 513–522. NIST, Gaithesburg, MD.

Ogden, W., Davis, M., and Rice, S. (1999c). Document thumbnail visualizations for rapid relevance judgements: When do they pay off? In *Proceedings of TREC7*. NIST, Gaithesburg, MD.

Ozawa, T., Yamamoto, M., Umemura, K., and Church, K. (1999). Japanese word segmentation using similarity measure for IR. In *Proceedings of the First NTCIR Workshop*, pages 89–96.

Peñas, A. (2002). *Website Term Browser: Un sistema interactivo y multilingüe de búsqueda textual basado en técnicas lingüísticas* . PhD thesis, Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia.

Peters, C. (2001). Introduction. In Peters, C., editor, *Cross-Language Information Retrieval and Evaluation, Workshop, CLEF 2000*, volume 2069 of *LNCS*, pages 1–6. Springer.

Peters, C. (2002a). Introduction. In *Proceedings of CLEF 2002*.

Peters, C. (2002b). Introduction. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *LNCS*, pages 1–8. Springer.

Peters, C. and Picchi, E. (1997). Using linguistic tools and resources in cross-language retrieval.

Petrelli, D., Beaulieu, M., Sanderson, M., Demetriou, G., and Herring, P. (2002). Is Query Translation a Distinct Task from Search? In *Proceedings of CLEF 2002*.

Pirkola, A. (1998). The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval. In *Proceedings of SIGIR'98*, pages 55–63.

Porter, M. (1980). An Algorithm for Suffix Stripping. *Program*, 14:130–137.

Porter, M. (2001). Snowball: A language for stemming algorithms. `http://snowball.sourceforge.net`.

Powell, A., French, J., Callan, J., Connell, M., and C.L., V. (2000). The impact of database selection on distributed searching. In *Proceedings of SIGIR'2000*, pages 232–239.

Qiu, Y. (1995). *Automatic query expansion based on a similarity Thesaurus*. PhD thesis, Swiss Federal Institute of Technology.

Qiu, Y. and Frei, H.-P. (1993). Concept-based query expansion. In *Proceedings of SIGIR'93*, pages 160–169, Pittsburgh, US.

Resnik, P. (1997). Evaluating Multilingual Gisting of Web Pages. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence.

Resnik, P. (1998). Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text. In *AMTA*, pages 72–82.

Ruiz, M., Diekema, A., and Sheridan, P. (2000). CINDOR Conceptual INterlingua DOcument Retrieval: TREC-8 Evaluation. In *Proceedings of TREC8*, pages 597–606. NIST, Gaithesburg, MD.

Salton, G. (1970). Automatic Processing of Foreign Language Documents. *Journal of American Society for Information Sciences*, 21:187–194.

Salton, G. (1989). *Automatic Text Processing - The Analysis, Transformation and Retrieval of Information by Computer*. Addison-Wesley.

Salton, G. and Buckley, C. (1983). *Introduction to Modern Information Retrieval*. Mc-Graw Hill.

Savoy, J. (1999). A Stemming Procedure and Stopword List for General French Corpora. *Journal of the American Society for Information Science*, 50:944–952.

Schinke, R., Robertson, A., Willet, P., and Greengrass, M. (1996). A stemming algorithm for Latin text databases. *Journal of Documentation*, 52:172–187.

Senseval-2 (2001). *Proceedings of Senseval-2, Second International Workshop on Evaluating Word Sense Disambiguation Systems*. Association for Computational Linguistics.

Sheridan, P., Braschler, M., and Schäuble, P. (1997). Cross-language information retrieval in a multi-lingual legal domain. In Peters, C. and Thanos, C., editors, *Proceedings of ECDL-97, 1st European Conference on Research and Advanced Technology for Digital Libraries*, pages 253–268, Pisa, IT.

Simard, M., G.F., F., and Isabelle, P. (1992). Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81.

Soergel, D. (1997). Multilingual thesauri in cross-languate text and speech retrieval. In Hull, D. and Oard, D., editors, *AAAI Symposium on Cross-Language Text and Speech Retrieval*.

Sperer, R. and Oard, D. W. (2000). Structured Translation for Cross-Language Information Retrieval. In *Proceedings of SIGIR'2000*, pages 120–127.

Srinivasan, P. (1996). Optimal document-indexing vocabullary for MEDLINE. *Information Processing & Management*, 32(5):503–514.

Suzuki, Inoue, N., and Hashimoto, K. (2001). A Method for Supporting Document Selection in Cross-Language Information Retrieval and its Evaluation. *Computers and the Humanities*, 35(4):421–438.

Taylor, K. and White, J. (1998). Predicting what MT is good for: User judgments and task performance. In Farwell, D., Gerber, L., and Hovy, E., editors, *Third conference of the Association for Machine Translation in the Americas*, Lecture Notes in Artificial Intelligence, pages 364–373. Springer.

Volk, M., Ripplinger, B., Vintar, S., Buitelaar, P., raileanu, D., and Sacaleanu, B. (2002). Semantic annotation for concept-based cross-language medical information retrieval. *International Journal of Medical Informatics*, 67(1-3).

Voorhees, E., Gupta, N., and Johnson-Laird, B. (1995). The Collection Fusion Problem. In *Proceedings of TREC3*, pages 95–104. NIST, Gaithesburg, MD.

Vossen, P. (1998). Introduction to EuroWordNet. *Computers and the Humanities, Special Issue on EuroWordNet*.

Wang, J. and Oard, D. W. (2002). iCLEF 2001 at Maryland: Comparing Term-for-Term Gloss and MT. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *LNCS*, pages 336–354. Springer.

Wong, S., Ziarko, W., and Wong, P. (1985). Generalizad vector space model in information retrieval. In *Proceedings of SIGIR'85*, pages 18–25.

Yang, Y., Carbonell, J. G., Brown, R. D., and Frederking, R. E. (1998). Translingual Information Retrieval: Learning from Bilingual Corpora. *Artificial Intelligence*, 103(1-2):323–345.