Problemas de la Traducción de la Consulta en la Búsqueda de Información Multilingüe

Problems in Query Translation in Multilingual Information Retrieval

Claudia Deco, Cristina Bender, Mario Chiari
Departamento de Investigación Institucional, Facultad de Química e Ingeniería,
Universidad Católica Argentina, Rosario, Argentina
cbender@uca.edu.ar

Resumen

En una búsqueda multilingüe de información, los idiomas de la consulta y de los documentos son distintos. Por lo tanto, si se desean recuperar documentos en otro idioma, es necesario efectuar una traducción de la consulta para realizar la búsqueda en dicho idioma. La Recuperación de Información Multilingüe trata el problema de encontrar documentos que están escritos en otros idiomas, distintos al idioma de la consulta. Este proceso no es simple debido a la complejidad semántica del vocabulario. La necesidad de realizar búsquedas multilingües es un hecho, y la demanda de este tipo de búsquedas aumentará en los próximos años con el crecimiento de la Web.

En este trabajo se presenta el problema de la búsqueda de información multilingüe, con especial atención a los distintos recursos lingüísticos que se pueden utilizar, y los problemas que se presentan en la traducción de la consulta. Se describen algunas técnicas utilizadas en la recuperación de información, y se presenta la expansión de la consulta como un método para mejorar esta recuperación. Además, se presentan los resultados de la experimentación realizada para evaluar algunos diccionarios multilingües disponibles en línea, para las traducciones entre los idiomas español, inglés y francés

Palabras claves: Recuperación de información multilingüe, recursos lingüísticos, traducción de la consulta.

Abstract

The problem in multilingual information retrieval is that the language of the query and the languages of the documents could be different. Because of this, it is necessary to carry out a query translation in order to retrieve documents in other languages. This is not a simple process due to the semantic complexity of vocabulary. The necessity to make multilingual searches is a fact, and the demand of this type of searches will increase with the growth of the web in the next years.

This work analyzes multilingual information retrieval, specifically linguistic resources that can be used in query translation, and the problems encountered in query translation. In addition, some information retrieval techniques are described and query expansion is proposed as a method to

improve such retrieval. Finally, we present the results of the experimentation conducted to evaluate some multilingual dictionaries available online, for Spanish, English and French translation.

Keywords: Multilingual information retrieval, linguistic resources, query translation

1. INTRODUCCION

La Búsqueda ó Recuperación de Información es el proceso en el que, dadas una consulta y una colección de documentos, se devuelve una lista ordenada de documentos relevantes para la consulta. El objetivo principal de la Recuperación de Información es satisfacer la necesidad de información planteada por un usuario en una consulta en lenguaje natural, especificada a través de un conjunto de palabras claves. Un motor de búsqueda ideal recuperaría todos y sólo aquellos documentos que son relevantes a la consulta del usuario. Recuperar todos los documentos relevantes implica tener una cobertura completa, y recuperar sólo los documentos relevantes implica tener una precisión perfecta.

En general, este proceso hacia la recuperación de documentos textuales relevantes a la consulta presentada no es un proceso simple debido a la complejidad semántica del vocabulario. El problema central es establecer una correspondencia entre el lenguaje de la consulta y el lenguaje del documento. Esto se debe a que los autores de los documentos y los usuarios frecuentemente utilizan diferentes palabras ó expresiones cuando se refieren a un mismo concepto. Por ejemplo, en medicina, "cáncer" puede también ser expresado como "neoplasma". Si en un documento, en lugar del término "cáncer" apareciera la palabra "neoplasma", este documento no se recuperaría. Esto se soluciona utilizando sinónimos.

Por otro lado, algunos términos pueden tener significados diferentes. Por ejemplo, la palabra "cáncer" puede referirse a una enfermedad en medicina, a un signo zodiacal en astrología ó a una constelación de estrellas en astronomía. Esto se soluciona desambiguando el término. Esta desambiguación se puede hacer agregando otros términos específicos relacionados con la acepción de interés; por ejemplo, utilizar ("cáncer" y "terapia") en lugar de usar sólo el término "cáncer", si interesa la acepción médica.

Este modelo tradicional de búsqueda de información supone que la consulta y los documentos están escritos en el mismo idioma. La mayoría de los motores de búsqueda tienen la limitación de encontrar documentos sólo en el idioma en el que se escribe la consulta. La Recuperación de Información Multilingüe trata el problema de encontrar documentos que están escritos en otros idiomas, distintos al idioma de la consulta.

En este trabajo se analizan algunas técnicas y recursos que pueden utilizarse en una búsqueda de información multilingüe. En la Sección 2 se describen algunas técnicas utilizadas en la recuperación de información, y la expansión de la consulta como un método para mejorar la recuperación. En la Sección 3 se describen recursos que pueden utilizarse en la traducción de la consulta. En la Sección 4 se analizan los problemas que se presentan en la traducción de la consulta, y se describe la experimentación realizada. Finalmente, en la Sección 5 se presentan las conclusiones.

2. ALGUNAS TÉCNICAS UTILIZADAS EN LA RECUPERACIÓN DE INFORMACIÓN

Se han desarrollado muchas técnicas ó herramientas para mejorar la recuperación de información. Una de ellas es el *stemming*. La técnica de stemming consiste en obtener la raíz de las palabras, de forma que el proceso de búsqueda se realice sobre las raíces y no sobre las palabras originales. Esta

técnica permite a un sistema de recuperación de información relacionar términos presentes en la consulta con los que se encuentren en los documentos y que aparezcan en alguna de sus variantes morfológicas. Para esto se supone que dos palabras que tengan la misma raíz representan el mismo concepto.

Los primeros algoritmos de stemming se desarrollaron para el idioma inglés. Pero esta técnica necesita ser adaptada para idiomas que presentan características distintas al inglés, como ser idiomas más flexivos, tal como el español. Uno de los algoritmos más utilizados para el inglés, es el de Porter [1]. También existen algoritmos para otros idiomas tales como el francés [2], el español [3], el holandés [4], el griego [5] y el latín [6]. En general, estos algoritmos se basan en un conjunto sencillo de reglas que truncan las palabras hasta obtener una raíz común.

En idiomas aglutinativos, como el alemán y el holandés, en los cuales se unen palabras para formar otras más largas, otra técnica que se puede aplicar es la *segmentación de palabras compuestas* [7]. Por ejemplo, la palabra alemana "Fachinformationszentrum", está compuesta por "Fach" (especialidad), "Information" (información) y "Zentrum" (centro), y se traduce como "centro de información especializada". 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.

Por otro lado, en el entorno de búsqueda tradicional, el usuario debe dividir su interés de búsqueda en distintos conceptos. No siempre un término representa en forma adecuada un concepto. Utilizar otros términos equivalentes ó más adecuados para expresar un concepto es realizar una *expansión de consulta* [8]. Esta situación requiere un cambio en el pensamiento del proceso para elegir los términos de búsqueda. Podría ser necesario consultar recursos lingüísticos, tales como un tesauro o un diccionario, para incorporar nuevos términos. La expansión de consultas es el proceso de suplementar la consulta original con términos adicionales, y es un método para mejorar el desempeño de la recuperación. La expansión de consultas puede ser desarrollada manual, automática o interactivamente.

Para realizar la expansión, en [9] se propone un refinamiento semántico de la consulta para la recuperación de información monolingüe. Este refinamiento consiste en guiar al usuario para desambiguar los conceptos ingresados por él, permitirle seleccionar conceptos jerárquicamente relacionados a fin de precisar los documentos a recuperar, y *expandir* semánticamente los conceptos a fin de aumentar la cantidad de documentos a recuperar. El esfuerzo inicial que se pretende por parte del usuario en la desambiguación de términos y en la selección de conceptos relacionados sugeridos por el sistema, es recompensado evitándole a posteriori la lectura y el descarte de los documentos que no sean de su interés. La cantidad de documentos recuperados aumenta mediante el agregado de sinónimos y palabras relacionadas. La mejora en la precisión de los resultados se logra presentándole al usuario una estructura jerárquica de conceptos que le permite hacer un recorrido conceptual de su consulta. Es decir, moverse por jerarquías conceptuales, subiendo ó bajando de nivel conceptual, y seleccionando un término más cercano a su necesidad de información. Los resultados generales obtenidos en dicho trabajo muestran que tanto el promedio de la cantidad de documentos recuperados como la precisión de las búsquedas se incrementan en cerca de un 20% al realizar el refinamiento semántico.

3. TRADUCCIÓN DE LA CONSULTA

El problema en una búsqueda multilingüe de información es que los idiomas de la consulta y de los documentos son distintos. Por lo tanto es necesario efectuar una traducción para poder realizar una búsqueda en la que tanto la consulta como los documentos se encuentren en el mismo idioma.

Salton [10] planteó por primera vez el problema de encontrar documentos escritos en un idioma diferente al de la consulta. Propuso la utilización de un tesauro bilingüe alemán-inglés. Los resultados obtenidos fueron similares a los de una búsqueda monolingüe, debido a que el tesauro utilizado había sido construido manualmente. De esta forma la correspondencia entre los términos entre ambos idiomas era perfecta y no existía ambigüedad en los términos de búsqueda.

En el problema de la recuperación de información multilingüe, la traducción de la consulta es la opción más frecuente, porque su costo computacional es menor al costo de traducir los documentos.

Los tres problemas principales para automatizar la traducción de la consulta, según Grefenstette [11], son: saber cómo un término escrito en un idioma puede ser expresado en otro idioma; decidir cuáles de las posibles traducciones de cada término son las adecuadas en un contexto dado; y saber cómo medir la importancia de las diferentes traducciones que se consideran adecuadas. Estos problemas son compartidos por los sistemas de traducción automática y los sistemas de recuperación de información multilingüe.

Para realizar la traducción automática de la consulta se pueden utilizar recursos tales como diccionarios multilingües y tesauros multilingües.

Un *diccionario* indica las distintas acepciones de un término y permite su expansión con sinónimos. Algunos de los diccionarios permiten además la expansión con otros términos relacionados jerárquica y/o semánticamente a cada acepción del término, como ser merónimos, hipónimos e hiperónimos. Un diccionario muy utilizado, en sistemas automatizados, es WordNet [12], que es un sistema de referencia léxica online, cuyo diseño está inspirado en teorías psicolingüísticas actuales. Los sustantivos, verbos, adjetivos y adverbios están organizados en conjuntos de sinónimos, cada uno de los cuales representa un concepto subyacente. Estos conjuntos de sinónimos además se relacionan jerárquicamente. Este sistema provee las distintas acepciones de un concepto, permitiendo además la expansión de éste con sinónimos, merónimos, hipónimos y otros tipos de términos relacionados a la acepción elegida.

Para ampliar cada concepto de la consulta a otros idiomas, pueden utilizarse *diccionarios multilingües* generales ó especializados. Un ejemplo de diccionario multilingüe general es EuroWordNet [13], que es una base de datos multilingüe con redes de palabras para varios de los idiomas europeos: holandés, italiano, español, alemán, francés, checo y estonio. Está basado en el diccionario WordNet. Los idiomas están interconectados de forma que se puede ir de palabras en un idioma a sus palabras equivalentes en cualquiera de los otros idiomas.

Otra posibilidad es el uso de *programas de traducción automática*. En consultas formadas por frases, el uso de estos programas produce una mejora en la desambiguación, frente al uso de diccionarios que traducen palabras aisladas. Esto se debe a que los sistemas de traducción automática consideran la estructura sintáctica del texto.

Un *tesauro* es un vocabulario controlado y dinámico de términos relacionados semántica y genéricamente, los cuales cubren un dominio específico del conocimiento. En el lenguaje natural, existen sinónimos, es decir grupos de palabras que representan el mismo concepto, por ejemplo “cáncer” y “neoplasma”; y homónimos, que son palabras que representan más de un concepto, por ejemplo “banco”, que puede referirse al mueble ó a la institución financiera. El control de vocabulario implica la selección de un término preferido, también conocido como descriptor ó palabra clave, entre un grupo de sinónimos; y la calificación de homónimos para diferenciar su significado, eligiendo un significado preferido para cada término.

El tesauro está estructurado formalmente con el objeto de hacer explícitas las relaciones entre los conceptos. Estas relaciones pueden ser: jerárquicas, de afinidad, y preferenciales. Las relaciones jerárquicas indican términos más amplios ó más específicos de cada concepto. Las relaciones de afinidad muestran términos relacionados conceptualmente, pero que no están ni jerárquica ni

preferencialmente relacionados. Las relaciones preferenciales se utilizan para indicar cuál es el término preferido en el caso de sinónimos; y para indicar un término alternativo en el caso de homónimos.

A diferencia de un diccionario, donde todos los sinónimos de un concepto son representativos y tratados por igual, en un tesauro se tiene una palabra clave preferida y representativa del conjunto de sinónimos para cada concepto.

Un *tesauro multilingüe* sobre un área del conocimiento permite la traducción de términos específicos de ese dominio que quizá no puedan encontrarse en un diccionario. Los tesauros multilingües son recursos diseñados específicamente para la recuperación multilingüe de información. Un ejemplo de este tipo de tesauro sobre el dominio médico es UMLS (Unified Medical Language System), que es el Sistema Unificado de Terminología Médica de la Biblioteca Nacional de Medicina de Estados Unidos [14]. Otro ejemplo, de tesauro multilingüe general es EuroVoc [15], de la Comunidad Europea, que abarca nueve idiomas.

El refinamiento propuesto en [9] puede ser extendido a la recuperación de información multilingüe si en la etapa de *expansión*, se utilizan recursos multilingües para traducir los términos originales a otros idiomas, realizando así una expansión multilingüe de la consulta.

4. PROBLEMAS QUE SE PRESENTAN EN LA TRADUCCIÓN DE LA CONSULTA

El uso de un diccionario como recurso en la traducción automática de la consulta presenta problemas, tales como los siguientes:

- Los términos específicos ó técnicos, propios de un área del conocimiento, pueden no existir en un diccionario de uso general. Para dominios específicos del conocimiento se logran mejores resultados si se utilizan diccionarios especializados.
- En un diccionario pueden no estar todas las variantes morfológicas de una palabra. Este problema se soluciona utilizando la técnica de stemming, llevando la palabra no encontrada a su forma raíz y buscando ésta en el diccionario.
- Muchos diccionarios no tienen traducciones para los sustantivos propios.
- Una palabra en un idioma, puede tener varias traducciones distintas en otro idioma. Para decidir cuál es la traducción adecuada, debe contemplarse el contexto. Este es un problema complejo, ya que se debe automatizar la desambiguación de la traducción.
- Muchos diccionarios no tienen traducciones para conceptos formados por varias palabras, es decir por frases. La traducción de cada término por separado puede llevar a un error en la traducción del concepto.

4.1. Experimentación

El objetivo de las experiencias fue evaluar algunos diccionarios multilingües, disponibles en línea, para las traducciones entre los idiomas español, inglés y francés. Para esto, se utilizaron los siguientes recursos:

- Systran (tr.voila.fr).
- Reverso (www.elmundo.es/traductor/): traductor del diario El Mundo de España.
- El servicio de SDL internacional (www.freetranslation.com/). Este servicio no ofrece la traducción del español al francés.

- Wordlingo (www.worldlingo.com/en/products_services/worldlingo_translator.html).

Los resultados de estas experiencias se encuentran en las tablas 1 y 2. En la Tabla 1 se muestran las traducciones del español al inglés. En la Tabla 2 se presentan las traducciones del español al francés. En ambos casos se utilizó el mismo grupo de términos.

Tabla 1: Traducciones entre el español y el inglés.

Término en Español	Traducción de Systran	Traducción de Reverso	Traducción de SDL	Traducción de Wordlingo
Alemania	Germany	Germany	Germany	Germany
Almohada	Pillow	Pillow	Pillow	Pillow
Anglosajón	Anglo-saxon	Anglo-saxon	Anglo-saxon	Anglo-saxon
Arreglo	Adjustement	Arrangement	I arrange	Adjustment
Ayuda	Aid	Help	It helps	Aid
Bandeja	Tray	Tray	Tray	Tray
Base De Datos	Data base	Base of information	Database	Data base
Basto	Coarse	Pack-saddle	I suffice	Coarse
Bujía	Spark plug	Candlestick / Spark plug	Sparkplug	Spark plug
Callo	Callus	Corn	I silence	Callus
Camboya	Cambodia	Cambodia	Cambodia	Cambodia
Cisne	Swan	Swan	Swan	Swan
Comida	Food	Food	Food	Food
Erizo	Sprocket wheel	Hedgehog	I bristle	Sprocket wheel
Falta	Lack	Lack / Mistake	It lacks	Lack
Ginebra	Geneva	Geneva	Geneva	Geneva
Guardarropa	Wardrobe	Wardrobe	Coat room	Wardrobe
Hamaca	It swings	Hammock	Hammock	Hammock
Lamento	Moan	Lament	Lament	Moan
Loco	Crazy person	Madman	Crazy	Crazy person
Loza	Stoneware	Crockery	China	Stoneware
Matriz	Matrix	Counterfoil	Headquarters	Matrix
Mesa	It pulls	Table	Table	Table
Móvil	Movable	Mobile	Mobile	Moving body
Pánico	Panic	Panic	Panic	Panic
Pekin	The beijing	Pekin	Pekin	The beijing
Remera	Rower	Remere	Oarswoman	Rower
Ruido	Noise	Noise	Noise	Noise
Telaraña	Spiderweb	Spiderweb	Web	Spiderweb
Tocino	Bacon	Bacon	Bacon	Bacon
Ultra Rápido	Extreme express	Ultra rapid	Right-wing fast	Extreme express
Uso	Use	Use	Use	Use
Zorra	Vixen	Fox	Foxy	Vixen

En estas tablas, se observa que el término *Basto*, que puede corresponder a un sustantivo ó a un verbo conjugado, es traducido por Systran como sustantivo y como verbo, para las traducciones al francés. Pero Reverso lo traduce como sustantivo solamente y Wordlingo lo traduce como verbo solamente. En las traducciones al inglés, sólo SDL lo traduce como verbo, el resto lo traduce como sustantivo.

Los términos *Hamaca*, *Ayuda*, *Arreglo*, *Falta*, *Uso* y *Callo*, que pueden corresponder tanto a un verbo conjugado como a un sustantivo, son traducidos por todos los traductores al francés como sustantivo. Sin embargo, en el caso de *Lamento*, la mayoría de los traductores analizados lo traducen como verbo. En sus traducciones al inglés, Systran es el único que considera a *Hamaca* como verbo; y SDL es el único que considera a *Ayuda*, *Arreglo*, *Falta*, y *Callo* como verbos

conjugados.

Respecto a los sustantivos propios, Reverso en su traducción al inglés los interpreta como tales si están escritos en mayúsculas. Así, *Ginebra* lo traduce como *Geneva*, pero *ginebra* lo traduce como *gin*. Si un sustantivo propio se ingresa en minúsculas, y no corresponde a un sustantivo común, ni Reverso ni Systran los traducen.

Tabla 2: Traducciones entre el español y el francés.

Término en Español	Traducción de Systran	Traducción de Reverso	Traducción de Wordlingo
Alemania	L'Allemagne	L'Allemagne	L'Allemagne
Almohada	Oreiller	Oreiller	Oreiller
Anglosajón	Anglo-saxon	Anglo-saxon	Anglo-saxon
Arreglo	Ajustement	Entente	Ajustement
Ayuda	Aide	Aide	Aide
Bandeja	Plateau	Plateau	Plateau
Base de datos	Base de données	Base de données	Base de données
Basto	Brut / Je suffis	Bât	Je suffis
Bujía	Bougie	Chandelier / Bougie	Bougie
Callo	Calus	Grain / Maïs	Calus
Camboya	Le Cambodge	Le Cambodge	Le Cambodge
Cisne	Cygne	Cygne	Cygne
Comida	Repas	Alimentation	Repas
Erizo	Hérisson	Hérisson	Hérisson
Falta	Manque	Manque / Erreur	Manque
Ginebra	Genève	Genève	Genève
Guardarropa	Guardarropa	Garde-robe	Guardarropa
Hamaca	Hamac	Hamac	Hamac
Lamento	Je regrette	Lamenter	Je regrette
Loco	Fou	Fou	Fou
Loza	Faïence	Poterie	Faïence
Matriz	Matrice	Souche	Matrice
Mesa	Table	Table	Table
Móvil	Mobile	Portable	Raison
Pánico	Panique	Panique	Panique
Pekin	Pekin	Pékin	Pekin
Remera	Rémige	Resimple	Rémige
Ruido	Bruit	Bruit	Bruit
Telaraña	Toile d'araignée	Spiderweb	Toile d'araignée
Tocino	Lard	Bacon	Lard
Ultra rápido	Ultra rapide	Ultra rapide	Ultra rapide
Uso	Utilisation	Utilisation	Utilisation
Zorra	Renard	Renard	Renard

En sus traducciones del español al francés del término *Pekín*, tanto Systran como Wordlingo, omiten la acentuación de la letra "e", lo que es un error.

Reverso traduce *Telaraña* al francés como *Spiderweb*. Esto es llamativo, porque ninguna de las dos componentes de esta palabra (*spider* y *web*) son de origen francés. Sin embargo, con Reverso, *Spiderweb* no es traducida al español ni al inglés.

Se ha observado además que en algunos casos, que se detallan a continuación, se presenta el problema de que la traducción no es bidireccional. En las traducciones entre el español y el inglés realizadas por Reverso, se advirtió que:

- *Matriz* lo traduce al inglés como *Counterfoil*. *Counterfoil* lo traduce al español como *Talón*.

Talón lo traduce al inglés como *Heel*. Sin embargo, el término *Matrix* lo traduce al español como *Matriz*.

- *Callo* lo traduce al inglés como *Corn*. *Corn* lo traduce al español como *Grano*. *Grano* lo traduce al inglés como *Grain*. *Callus* lo traduce al español como *Callo*
- *Basto* es traducido al inglés como *Pack-saddle*. *Pack-saddle* es traducido al español como *Albarda*.

Y en las traducciones entre el español y el francés realizadas por Reverso, se observó que:

- *Arreglo* lo traduce al francés como *Entente*. *Entente* lo traduce al español como *Armonía*. *Armonía* lo traduce al francés como *Harmonie*.
- *Loza* lo traduce al francés como *Poterie*. *Poterie* lo traduce al español como *Alfarería*. Sin embargo, *Faïence* lo traduce al español como *Loza*.
- *Callo* lo traduce al francés como *Grain* (Maïs). *Grain* lo traduce al español como *Grano*. *Cal*, traducida al español, da como resultado *Callo*. *Durillon*, traducido al español, da como resultado *Callosidad*.
- *Basto* es traducido al francés como *Bât*. Pero *Bât* no es reconocido para traducirlo al español.
- *Móvil* es traducido al francés como *Portable*. *Portable* es traducido de idéntica forma al español. Sin embargo, *Mobile* es también traducido al español como *Móvil*.

En las traducciones entre el español y el inglés realizadas por Wordlingo, se observó que:

- *Bujía* es traducida al inglés como *Spark plug*. *Spark plug* es traducido al español como *Chispa Enchufe*. *Chispa* es traducido al inglés como *Spark*. *Enchufe* es traducido como *Fit*. *Fit* es traducido al inglés como *Ajuste*. Sin embargo, *Sparkplug* (todo junto) sí es traducido al español como *Bujía*.
- *Comida* es traducido al inglés como *Food*. *Food* es traducido al español como *Alimento*. *Meal* es traducido al español como *Comida*.
- *Lamento* es traducido al inglés como *Moan*. *Moan* es traducido al español como *Quejido*. *Quejido* es traducido al inglés como *Complaint*. *Complaint* es traducido al español como *Queja*.
- *Almohada* es traducido al inglés como *Pillow*. *Pillow* es traducido al español como *Almohadilla*. *Almohadilla* es traducido al inglés como *Pad*. *Pad* es traducido al español como *Cojín*. *Cojín* es traducido al inglés como *Cushion*. *Cushion* es traducido al español como *Amortiguador*.

Con sustantivos compuestos, también se presenta el problema de la traducción bidireccional. En este sentido, se observó que:

- SDL traduce *Guardarropas* como *Coat room*. *Coat room* es traducida al español como *Revista el espacio*. Sin embargo, *Wardrobe* es traducida al español como *Guardarropa.*
- SDL traduce *Telaraña* como *Web*. *Spiderweb* también es traducido al español como *Telaraña*. Sin embargo, Systran y Wordlingo traducen *Web* como *Tela*. Reverso no traduce *Web* al español.
- Reverso traduce *Base de datos* como *Base of information*. *Base of information* es traducida al español como *Base de información*. Sin embargo, Reverso traduce al español la palabra inglesa *Database* como *Base de datos*.
- Systran traduce *Ultra rápido* al inglés como *Extreme express*. Pero, *Extreme express* lo traduce al español como *Extremo expreso*. Sin embargo, traduce la palabra inglesa *Ultrarapid* al español

como *Ultrarrápido*. Esta última palabra no es de existencia reconocida por la Real Academia Española.

Por todos estos problemas, la utilización de un diccionario como único recurso de traducción reduce la efectividad de las búsquedas multilingües.

Diversos trabajos, como los de Hull [16] y Ballesteros [17], comprueban que si se sustituye cada término de la consulta por todas las traducciones ofrecidas por el diccionario, la efectividad se reduce entre un 40 y un 60%, respecto de la misma búsqueda realizada en un contexto monolingüe.

Con respecto a la polisemia, Davis [18] propone utilizar la categoría gramatical de las palabras de la consulta para elegir entre las posibles traducciones de los términos. 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 sustituir cada término por todas las traducciones ofrecidas por el diccionario.

Ballesteros y Croft [19] intentan mejorar la efectividad de las traducciones utilizando traductores de expresiones multipalabra. Con este tipo de recurso, las búsquedas fueron aproximadamente 150% más eficientes que aquellas en las que se tradujo cada palabra por separado.

Pirkola [20] concluye que la traducción de la consulta escrita en lenguaje natural provee una mayor precisión que si la consulta está expresada con palabras aisladas y se traduce cada palabra por separado. Además, para la traducción experimentó varias formas de combinar dos diccionarios bilingües: uno de propósito general y otro específico del dominio. Comprueba así que los mejores resultados se obtenían al utilizar todas las distintas traducciones proporcionadas por ambos diccionarios.

Boughanem [21] realiza una selección de las traducciones empleando las traducciones inversas, seleccionando sólo aquellas que pueden volver a traducirse al término de partida. Los resultados obtenidos en este trabajo muestran que esta estrategia puede ser más efectiva que otras más complejas, como la desambiguación de traducciones.

La interacción con el usuario es fundamental para solucionar estos problemas. Un sistema de búsqueda de información debe proporcionar al usuario la capacidad de expresar su necesidad de información en su propio idioma y ayudarlo a traducirla al idioma en el cual se encuentran los documentos. Para esto, el sistema puede utilizar un diccionario para traducir cada término de la consulta, permitiéndole al usuario, en el caso de términos ambiguos, seleccionar la traducción adecuada. A partir de esta selección, el sistema de búsqueda de información puede realizar una búsqueda automática.

En el caso de la traducción de frases pueden ocurrir que traducciones correctas no arrojen resultados. Por ejemplo, en el caso de *Enfermedad de Munchausen*, la traducción al inglés realizada por Systran es *Disease of Munchausen*. Cuando se busca esta frase en Google, la búsqueda arrojó cero resultados. En cambio, si se utilizan *Muchausen disease* ó *Munchausen's disease*, traducciones provistas por un usuario especialista en temas médicos, se obtuvieron 286 resultados y 150 resultados respectivamente. Con esto se ve la importancia de utilizar recursos especializados en cada área del conocimiento y no diccionarios ó recursos generales, en el caso de búsquedas especializadas.

La frase *Polimialgia reumática* es traducida por Systran como *Rheumatic polimialgia*, que buscada en Google arroja 3 resultados. Sin embargo, la traducción correcta es *Rheumatic polymyalgia*, que buscada en Google arroja 472 resultados. Una mala traducción, cuyo error pase inadvertido, puede llevar a obtener malas conclusiones, puesto que aun frases incorrectas arrojan algún tipo de resultados, lo que puede inducir a pensar que la traducción fue acertada y que en realidad no hay información abundante sobre eso en la Web.

Un enfoque distinto al presentado hasta aquí de traducción de la consulta, es la traducción de los documentos al idioma utilizado en la escritura de la consulta. Según Dumais [22] y Oard [23], este enfoque brinda traducciones más precisas porque se cuenta con información del contexto en el que se utilizan las palabras. Pero el problema que se presenta en este caso, es que el tiempo que lleva traducir los documentos es mucho mayor que el necesario para traducir la consulta.

5. CONCLUSIONES

En el globalizado mundo actual, la tecnología pone a disposición de quienes pueden acceder a ella una gran masa de documentos de infinitud de temáticas y entre los cuales se encuentran textos de altísimo valor. Estos textos pueden estar en un idioma distinto al utilizado para la consulta. La necesidad de realizar búsquedas multilingües es un hecho, y la demanda de este tipo de búsquedas aumentará en los próximos años con el crecimiento de la Web. La Recuperación de Información Multilingüe trata el problema de encontrar documentos que están escritos en otros idiomas, distintos al idioma de la consulta. Este proceso no es simple debido a la complejidad semántica del vocabulario.

En este trabajo, se presentó el problema de la búsqueda de información multilingüe, con especial atención a distintos recursos lingüísticos que pueden utilizarse, y los problemas que se presentan en la traducción de la consulta. Se describieron algunas técnicas utilizadas en la recuperación de información; y se presentó la expansión de la consulta como un método para mejorar la recuperación.

En una búsqueda multilingüe de información, los idiomas de la consulta y de los documentos son distintos. Por lo tanto, es necesario efectuar una traducción para poder realizar una búsqueda en la que tanto la consulta como los documentos se encuentren en el mismo idioma. La traducción de la consulta es la opción más frecuente, porque su costo computacional es menor al costo de traducir los documentos. La traducción será de gran ayuda, a condición de que se trate de un trabajo de gran precisión y realizado con todo el respeto que la lengua de origen amerita. En este punto, y a pesar de los ingentes esfuerzos de los profesionales informáticos y lingüistas, es irrebatible que los progresos logrados en la traducción automática de textos no logran poner a la misma en un pie de igualdad con la traducción humana, que sigue siendo, con mucho, más exacta y comprensible.

Para realizar la traducción automática se pueden utilizar recursos tales como diccionarios multilingües y tesauros multilingües. Otra posibilidad es el uso de programas de traducción automática. En consultas formadas por frases, el uso de estos programas produce una mejora en la desambiguación, frente al uso de diccionarios que traducen palabras aisladas. Esto se debe a que los sistemas de traducción automática consideran la estructura sintáctica del texto. Una tercera posibilidad es trabajar directamente con la consulta expresada en lenguaje natural. La traducción en este caso, provee una mayor precisión que si la consulta está expresada con palabras aisladas y se traduce cada palabra por separado.

Las experiencias realizadas en este trabajo, tuvieron como objetivo evaluar algunos diccionarios multilingües, disponibles en línea, para las traducciones entre los idiomas español, inglés y francés. Los diccionarios utilizados fueron: Systran, Reverso, SDL y Wordlingo. De estas experiencias se ha observado que en algunos casos la traducción no es bidireccional. Otros problemas que se presentan son que muchos diccionarios no tienen traducciones para conceptos formados por varias palabras, ni para los sustantivos propios, ni para términos específicos ó técnicos. Además, una palabra puede tener varias traducciones distintas. En este caso, para decidir cuál es la traducción adecuada, debe contemplarse el contexto. Por todos estos problemas, la utilización de un diccionario como único recurso de traducción reduce la efectividad de las búsquedas multilingües.

La interacción con el usuario es fundamental para solucionar estos problemas. El sistema puede

utilizar un diccionario para traducir cada término de la consulta, permitiéndole al usuario, en el caso de términos ambiguos, seleccionar la traducción adecuada, y a partir de esta selección el sistema puede realizar una búsqueda automática.

Referencias

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

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

[3] Figuerola, C. G., Gomez, R., Rodrıguez, A. F. Z., 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.

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

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

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

[7] Monz, C., de Rijke, M. (2001) 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.

[8] Efthimiadis E.N. (1996) Query Expansion. In Annual Review of Information Systems and Technology (ARIST), v31, pp 121-187.

[9] Deco, C., Bender, C., Saer, J., Chiari, M., Motz, R. (2005). Semantic refinement for web information retrieval. In Proceedings of the 3rd Latin American Web Congress. IEEE Press. pp 106-110.

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

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

[12] Miller, G. (1995). WordNet: An on-line lexical database. International Journal of Lexicography 3(4).

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

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

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

[16] Hull, D. A. & 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.

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

[18] Davis, M. (1997). New Experiments in CrossLanguage Text Retrieval at NMSU's Computing Research Lab. In Proceedings of TREC5, pages 447–454. NIST, Gaithesburg, MD.

[19] Ballesteros, L. & 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.

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

[21] Boughanem, M., Chrisment, C., 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. Springer.

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

[23] 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.