Extracción crosslingüe de documentos usando mapas semánticos no-lineales [1]

Cross-Language Document Retrieval by using Non-linear Semantic Mapping

Rafael E. Banchs
Barcelona Media Innovation Centre
Av. Diagonal 177, planta 9
08010 Barcelona, Spain
+34-93-238-14-00
rafael.banchs@barcelonamedia.org

Marta R. Costa-jussà
Barcelona Media Innovation Centre
Av. Diagonal 177, planta 9
08010 Barcelona, Spain
+34-93-238-14-00
marta.ruiz@barcelonamedia.org

Resumen: Se propone un procedimiento no-lineal de mapeado semántico para extraer información multilingüe. El método consiste en utilizar una técnica de reducción de espacio no-lineal para agrupar colecciones de documentos multilingües. En el método propuesto, se construyen para cada lengua agrupaciones independientes de la colección multilingüe y se usan las similitudes de las expresiones semánticas para extraer documentos multilingües. Se implementan dos variantes del método y se comparan con técnicas de extracción de información multilingüe. El método propuesto, para unas tareas específicas, mejora el convencional.
Palabras clave: Extracción de información multilingüe, mapeado semántico

Abstract: A non-linear semantic mapping procedure is proposed for cross-language document retrieval. The method relays on a non-linear space reduction technique for constructing semantic embeddings of multilingual document collections. In the proposed method, an independent embedding is constructed for each language in the multilingual collection and the similarities among the resulting semantic representations are used for cross-language document retrieval. Two variants of the proposed method are implemented and compared with a state-of-the-art cross-language information retrieval technique. It is shown that, for some specific tasks, the proposed method outperforms the conventional one.
Keywords: Cross-language Information Retrieval, Semantic Mapping.

1 Introducción

La extracción de información multilingüe (EIM) se está haciendo más popular debido al aumento de información disponible en lenguas diferentes al inglés en internet. Actualmente, hay tres maneras de intentar afrontar la EIM: uso de cognados, técnicas basadas en traducción automática y técnicas basadas en interlingua [6].

En el primer caso, los cognados (palabras de diferentes lenguas con un origen etimológico parecido) se usan para identificar los contenidos textuales en diferentes lenguas que están relacionados con otros. En el segundo caso, las técnicas de traducción automática se usan para traducir o bien la consulta a buscar o bien la colección de documentos. De este modo, se puede realizar una extracción de información monolingüe. Finalmente, los métodos de interlingua se usan para asociar textos relacionados con contenidos de diferentes lenguas a través de representaciones semánticas.

Las técnicas convencionales EIM basadas en interlingua usan indexación semántica latente (LSI) para construir una representación vectorial de una colección paralela de documentos [3]. Una vez se tiene la representación vectorial, nuevos documentos y consultas se pueden proyectar y la tarea de extracción se lleva a cabo usando una métrica de similitud.

En este trabajo, se propone usar una técnica de mapeado semántico para implementar un

[1] Los autores quieren agradecer el soporte de los programas Ramón y Cajal y Juan de la Cierva, así como el permiso de publicación de la Fundación Barcelona Media

sistema EIM. Las técnicas de mapeado semántico se han usado tradicionalmente para asociar conceptos y términos relacionados con identificación de tareas [4,7], y se ha probado que el uso de métodos de proyecciones no-lineales en este contexto es más eficiente que los métodos lineales. En este trabajo, ilustramos como el uso de las técnicas semánticas no-lineales permite generar unas agrupaciones de espacio reducido para colecciones de documentos multilingües, que se pueden usar satisfactoriamente para aplicaciones EIM.

De manera diferente a las técnicas convencionales, en las cuales se construye una única representación vectorial multilingüe, en nuestro método, un mapa semántico se construye para cada lengua diferente y la tarea EIM se realiza explotando similitudes entre los diferentes mapas. Por otro lado, de manera similar a las técnicas convencionales, la construcción de los mapas requiere la disponibilidad de colecciones de documentos paralelos.

El resto del artículo se estructura como sigue. En la sección 2, la idea principal de la metodología propuesta, que es la construcción de mapas semánticos, se describe y se ilustra sobre una colección de documentos trilingüe incluyendo chino, castellano e inglés. En la sección 3, se definen dos variantes del método EIM y se describe con detalle su implementación. En la sección 4 se ilustra la metodología propuesta mediante la realización de experimentos EIM en la tarea trilingüe mencionada anteriormente. También, en esta sección, la metodología se compara con la técnica de EIM convencional basada en interlingua y se muestra que para unas tareas específicas la técnica propuesta mejora. Finalmente, en la sección 5, se presentan las conclusiones más relevantes.

2 Mapeado semántico

La idea principal del método EIM propuesto gira en torno el mapeado semántico. En la técnica de EIM basada en interlingua, se usa el modelo de espacio vectorial para representar una colección de documentos determinada. Esta representación se sabe que es ruidosa, por lo tanto, para obtener una representación vectorial más eficiente, se aplica una técnica de reducción de espacio como la indexación semántica latente [2] y el análisis semántico latente estocástico [5]. Se supone que las nuevas representaciones en dimensiones más reducidas son capaces de capturar relaciones semánticas entre palabras y documentos. Se ha demostrado que es posible hacer una reducción de espacio mayor al mismo tiempo que se preservan las características estructurales de los datos a través de métodos no-lineales de proyección y que las representaciones de menor dimensión se pueden usar para realizar asociación de conceptos y aplicación de identificación de términos [7].

En este trabajo, intentamos usar técnicas no-lineales de proyección, conocidas como escalado multidimensional, para construir unas agrupaciones semánticas de documentos en lugar de términos o conceptos. Si las representaciones obtenidas realmente responden a relaciones semánticas entre documentos, entonces podemos esperar que para una colección paralela de documentos, se obtengan agrupaciones similares para lenguas independientes. Esta sección explora e ilustra esta idea en profundidad.

2.1 Escalado multidimensional

El escalado multidimensional (MDS) constituye un método no-lineal para visualizar datos que se puede usar también como una técnica de reducción de espacio [1]. Dado un conjunto de relaciones entre los elementos de la colección, el objetivo del MDS es encontrar una representación de menor dimensión para la colección de manera que las relaciones entre los elementos se preserven lo máximo posible. Lo interesante sobre MDS es que las relaciones entre el conjunto original de elementos puede ser de naturaleza cuantitativa (las similitudes se basan en un distancia métrica) o de naturaleza de cualitativa (como relaciones ordinales o rankings).

En la práctica, el MDS se implementa como un problema de optimización, donde se define una función objetivo. Esta función objetivo, a la cual nos referiremos como función estrés, se define típicamente en términos del error cuadrático medio entre las distancias de los elementos en la representación resultante de dimensión reducida y las relaciones entre los elementos originales, los cuales se expresan generalmente en términos de una matriz de similitud. De acuerdo con esto, en general, la forma de la función estrés es como sigue:

$$\text{Stress function} = \sqrt{\frac{\sum \sum (f(x_{ij}) - d_{ij})^2}{\text{Scaling factor}}} \quad (1)$$

donde x_{ij} se refiere a las similitudes entre los elementos originales, d_{ij} se refiere a las distancias entre los elementos en la representación de dimensión reducida y f representa una transformación monótona de la función.

El mínimo valor de la función estrés para una dimensión dada, ofrece una medida de la cantidad de distorsión, o pérdida estructural, dada la mejor representación posible de la colección original de elementos para aquella dimensión específica.

2.2 Ejemplo ilustrativo

En esta sección, ilustramos la generación de mapas semánticos para una colección paralela trilingüe que se extrae de las versiones china, castellana e inglesa de la Sagrada Biblia. Las características básicas de esta colección se describen en la tabla 1.

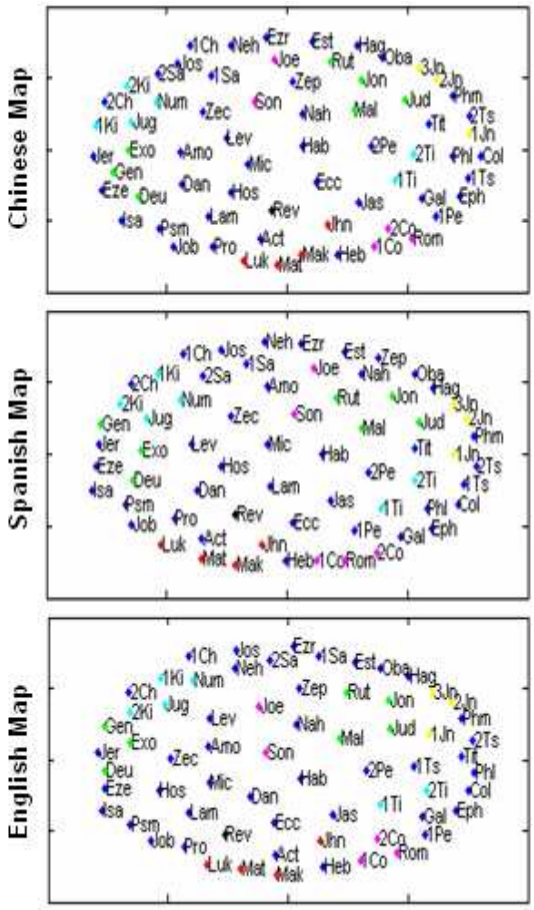
Tabla 1. Características básicas de los datos experimentales.

Idioma	Documentos	Vocabulario	No Singletons
Chino	**66**	**12,670**	**6,286**
Castellano	**66**	**26,251**	**13,632**
Inglés	**66**	**13,216**	**7,265**

El proceso de cómputo de la agrupación semántica para una colección monolingüe dada mediante MDS se puede definir como sigue: (1) obtener una representación vectorial para la colección mediante el estándar TF-IDF; (2) construir una matriz de similitud para los documentos usando las distancias coseno con respecto a sus correspondientes vectores; (3) construir un mapa semántico de dimensión reducida para los documentos en la colección usando MDS.

Siguiendo estos 3 pasos, construimos mapas de dos dimensiones para los 66 documentos descritos en la tabla 1 en cada idioma. Para cada uno de las tres representaciones vectoriales originales, se eliminaron las palabras que aparecían una única vez antes de calcular las similitudes entre documentos y realizar las proyecciones. Los mapas resultantes se presentan en la figura 1. Se puede observar, que aunque los mapas se han obtenido para cada idioma independientemente de los otros dos idiomas, las similitudes se reflejan claramente en el hecho que por encima del idioma, los documentos en la colección se relacionan con los otros, a través de la semántica contenida y estas relaciones semánticas han sido plasmadas en los mapas construidos.

Figura 1: Mapas semánticos obtenidos para una colección de documentos trilingüe



Una pregunta obvia que se deriva de esta observación tiene que ver con la posibilidad de explotar las similitudes observadas para realizar tareas de EIM. En la próxima sección describimos la metodología para hacerlo.

3 Metodología propuesta para EIM

Como se ha mostrado en la sección anterior, es posible construir representaciones semánticas para una colección de documentos utilizando técnicas de proyección no lineales. Además, si se dispone de una colección de documentos

paralelos, se puede calcular un mapa semántico independiente para cada idioma y los mapas resultantes muestran un alto grado de similitud entre ellos. Nuestra propuesta en esta sección es que las similitudes se pueden explotar para propósitos de EIM.

3.1 Descripción del procedimiento

Las similitudes que se observan en las representaciones semánticas presentadas en la figura 1 sugieren la posibilidad de mover documentos (o consultas) de diferentes idiomas de un mapa a otro. En este sentido, por ejemplo, las coordenadas para un documento en chino en un mapa de castellano o de un mapa en inglés se pueden estimar conociendo las coordenadas en el mapa chino.

De acuerdo con esto, si podemos colocar nuevos documentos en el correspondiente mapa del idioma del documento, deberíamos ser capaces de realizar una EIM implementando un sistema EI monolingüe en cualquiera de los mapas disponibles. El principal problema con esta aproximación es que la MDS es no-lineal. Esto significa que después de generar los mapas, no hay un operador que permita situar nuevos documentos en los mapas construidos. La única manera trivial de incluir un nuevo documento en un mapa es añadiendo el documento en la colección original y realizar la MDS otra vez. Pero esto no es una opción porque la influencia del nuevo documento, podría influir en que el nuevo mapa que sea diferente que el mapa original.

También hay dos alternativas más, la primera sería confiar en una proyección lineal para colocar los documentos nuevos en los mapas ya construidos; la segunda sería implementar una nueva optimización para el procedimiento de encontrar la mejor situación del nuevo documento de acuerdo con la similitud al resto de documentos.

Hemos realizado una experimentación en la dirección de la segunda alternativa. Pero los resultados obtenidos, que no se reportan en este trabajo, demuestran que la optimización presenta bastantes inconvenientes debido a motivos de convergencia. Por este motivo nos centramos en la primera propuesta.

En resumen, el procedimiento genérico para EIM mediante mapeado semántico se puede reducir a tres pasos: (1) empezar con una colección multilingüe de documentos y construir el mapa de extracción; (2) proyectar nuevos documentos y consultas de cualquier idioma fuente en el mapa de extracción usando proyecciones lineales (se describe en la próxima sección); (3) extraer documentos del mapa de extracción usando una métrica de distancia.

3.2 Proyección lineal de nuevos documentos

Una transformación lineal con el operador **T** para proyectar documentos o consultas del espacio original en un mapa de dimensión reducida se puede inferir de un conjunto de documentos como sigue:

$$\mathbf{M=TD} \quad => \quad \mathbf{T=MD}^{-1} \qquad (2)$$

donde **D** es una matriz cuadrada *NxN* que contiene las distancias entre documentos en el espacio original (la matriz de documentos de similitud), y **M** es la matriz *KxN* que contiene las coordenadas de los documentos en el mapa semántico de dimensión reducida.

Nótese que se pueden calcular dos variantes de la matriz de proyección lineal **T**:

1. Una matriz de proyección monolingüe, por la cual **M** y **D** se calculan en el idioma de extracción.

2. Una matriz de proyección multilingüe, por la cual **M** se calcula en el idioma de extracción y **D** se calcula en el idioma original.

3. Finalmente, cualquier nuevo documento o consulta se puede situar en el mapa de extracción utilizando la siguiente expresión:

$$m = T\,d \qquad (3)$$

donde **d** representa un vector que contiene las distancias entre el documento a añadir (o consulta) y el conjunto de documentos en la dimensión original,, **T** es la matriz de transformación definida en (2), y **m** es el vector que contiene el resultado de las coordenadas del documento (o consulta) a añadir en el mapa de menor dimensión.

4 Resultados Experimentales

4.1 Comparación del método propuesto

En primer lugar comparamos la metodología propuesta con el método interlingua descrito en [3]. Consideramos todas las posibles combinaciones multilingües entre los idiomas disponibles.

4.1.1 Detalles del experimento

La tarea que se considera en este primer experimento es extraer un libro usando el mismo libro en otro idioma diferente como consulta.

Los detalles del experimento se resumen como sigue: (1) se consideran 30 libros para entrenamiento; (2) se consideran 30 libros como test; (3) la dimensión del espacio de extracción es 30; (4) se consideran todas las posibles combinaciones entre idiomas; (5) se selecciona en idioma destino para devolver el libro extraído; (6) se hacen 500 simulaciones independientes para cada sistema; (7) se hace una selección aleatoria de los conjuntos de entrenamiento y test a cada iteración (no se permite solapamiento entre ellos).

4.1.2 Resultados experimentales y discusión

La tabla 2 resume los resultados para la todas las tareas de extracción consideradas en este experimento. Las filas de la tabla indican el idioma fuente y las columnas indican el idioma destino. Los valores resultantes corresponden a la media de valores de las 500 simulaciones independientes para cada experimento de extracción. Los valores correspondientes de desviación estándar también se proporcionan.

Cada entrada de la tabla 2 contiene tres valores: el primero (en cursiva) se refiere a los resultados del sistema de referencia basado en interlingua; el segundo se refiere a la metodología propuesta implementando una matriz de proyección monolingüe; y el tercero se refiere al resultado de la variante implementando una matriz de proyección multilingüe.

Tabla 2. Evaluación comparativa entre el sistema básico EIM (en cursiva) y las dos metodologías propuestas.

	Chino	**Castellano**	**Inglés**
Chino	*100% ±0%* **100% ± 0%** **100% ± 0%**	*82% ±7%* **82% ± 7%** **92% ± 5%**	*73% ±7%* **73% ± 7%** **84% ± 6%**
Castellano	*82% ±6%* **82% ± 7%** **91% ± 5%**	*100% ±0%* **100% ± 0%** **100% ± 0%**	*76% ±7%* **68% ± 9%** **83% ± 7%**
Inglés	*75% ±8%* **77% ± 7%** **86% ± 6%**	*74% ±8%* **77% ± 7%** **88% ± 6%**	*100% ±0%* **100% ± 0%** **100% ± 0%**

Como se puede ver en la tabla, los resultados de la metodología propuesta implementada usando la matriz de proyección multilingüe mejora los otros dos sistemas.

4.2 Dimensión semántica del mapa

En este experimento, se evalúan los efectos de la dimensión del mapa semántico. De nuevo, se compara el sistema de referencia con las dos metodologías propuestas, pero en este caso para un par específico de idiomas.

4.2.1 Detalles del experimento

En este experimento, la tarea continua siendo de extracción de un libro usando el mismo libro en un idioma diferente como consulta.

Los detalles del experimento se resumen de la siguiente manera: (1) se consideran 30 libros para entrenamiento; (2) se consideran 30 libros como test; (3) la dimensión del espacio de extracción es variable de 2 a 30; (4) se considera el inglés como idioma fuente y el chino como destino; (5) se selecciona el chino como idioma destino para devolver el libro extraído; (6) se hacen 100 simulaciones independientes para cada sistema; (7) se hace una selección aleatoria de los conjuntos de entrenamiento y test a cada iteración (no se permite solapamiento entre ellos).

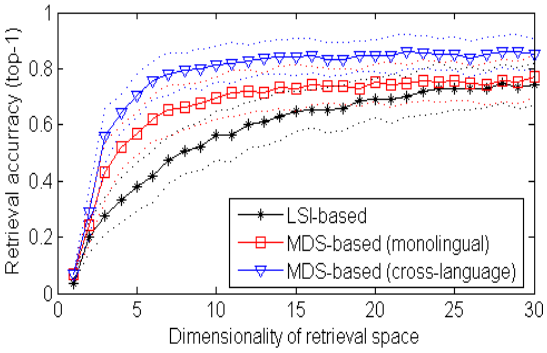
4.2.2 Resultados del experimento y discusión

La figura 2 presenta la media de la precisión (líneas continuas) ± su correspondiente intervalo de desviación estándar (líneas puntuadas) para la tarea de extracción que se considera. Se usa un sistema de referencia interlingua (asteriscos) y las dos variantes de la técnica propuesta: la implementada con una

matriz de proyección monolingüe (cuadros) y la implementada con una matriz de proyección multilingüe (triángulos).

Para todos los experimentos definidos en la figura 2, la colección de documentos multilingüe de tamaño 30 documentos se usó para todos los experimentos, mientras que la reducción de dimensión de espacio se varió de 2 a 30 (representada en los ejes horizontales).

Figura 2: Precisión en la extracción vs. dimensión del mapeado semántico



Como se puede observar en la figura, la variante que implementa una matriz de proyección multilingüe es consistentemente mejor que las otras dos aproximaciones.

4.3 Efectos en el tamaño del documento

En este experimento, se evalúa el efecto del tamaño de los documentos en el resultado del sistema. En esta sección se usan capítulos de libros para realizar la tarea de extracción. De acuerdo con esto, la colección original trilingüe de 66 libros se divide en capítulos y la nueva colección de documentos tiene un total de 1200 capítulos.

El objetivo de este test es doble: primero, ver como la metodología propuesta se comporta con documentos de menor tamaño, y segundo, evaluar como se comporta el sistema con una colección de documentos mayor.

4.3.1 Detalles del experimento

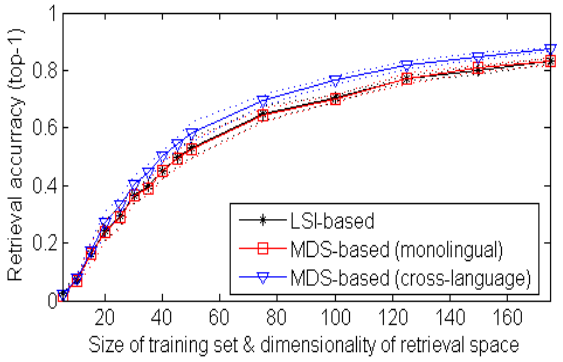
Los detalles del experimento se resumen como sigue: (1) se considera un tamaño variable de documentos entre 5 y 175 capítulos para entrenamiento; (2) se consideran 1000 capítulos como test; (3) la dimensión del espacio de extracción es igual a la de entrenamiento; (4) se consideran como idioma fuente el inglés e idioma destino el chino; (5) se selecciona el idioma destino (chino) para devolver el libro extraído; (6) se hacen 100 simulaciones independientes para cada sistema; (7) se hace una selección aleatoria de los conjuntos de entrenamiento y test a cada iteración (no se permite solapamiento entre ellos).

4.3.2 Resultados del experimento y discusión

La figura 3 presenta la media de la precisión (líneas continuas) ± su correspondiente intervalo de desviación estándar (líneas puntuadas) para la tarea de extracción que se considera. Se usa un sistema de referencia interlingua (asteriscos) y las dos variantes de la técnica propuesta: la implementada con una matriz de proyección monolingüe (cuadros) y la implementada con una matriz de proyección multilingüe (triángulos).

Para los experimentos que se muestran en la figura 4, el tamaño del conjunto de entrenamiento se varió de 5 a 175 capítulos. En todos los casos, la dimensión del espacio reducido era igual al tamaño del conjunto de entrenamiento.

Figura 3: Precisión de extracción vs. tamaño de la colección multilingüe de documentos para capítulos



Nótese que en la figura se ve que de nuevo la variante que implementa una matriz de proyección multilingüe es consistentemente mejor que las otras dos aproximaciones.

4.4 Efectos en el tamaño de la consulta

En este experimento final, se evalúa el efecto del tamaño de la consulta, considerándola de menor tamaño que los documentos.

4.4.1 Detalles del experimento

Los detalles del experimento se resumen como sigue: (1) se considera un tamaño variable de documentos entre 2 y 30 libros para entrenamiento; (2) se consideran 30 libros como test; (3) la dimensión del espacio de extracción es igual a la de entrenamiento; (4) se consideran como idioma fuente el inglés e idioma destino el chino; (5) se selecciona el idioma destino (chino) para devolver el libro extraído; (6) se hacen 100 simulaciones independientes para cada sistema; (7) se hace una selección aleatoria de los conjuntos de entrenamiento y de test.

4.4.2 Resultados del experimento y discusión

La figura 4 presenta la media de la precisión (líneas continuas) ± su correspondiente intervalo de desviación estándar (líneas puntuadas) para la tarea de extracción que se considera. Se usa un sistema de referencia interlingua (asteriscos) y las dos variantes de la técnica propuesta: la implementada con una matriz de proyección monolingüe (cuadros) y la implementada con una matriz de proyección multilingüe (triángulos).

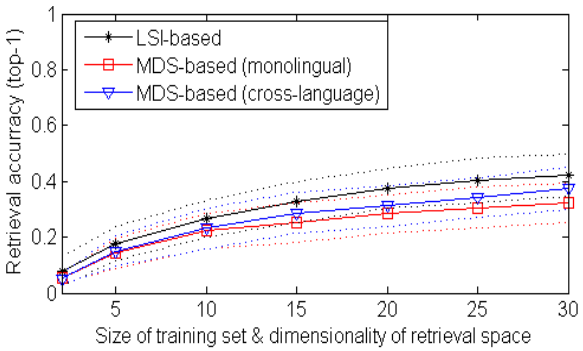
Para los experimentos que se muestran en la figura 4, el tamaño del conjunto de entrenamiento variaba entre 2 y 30 libros y la dimensión del espacio reducido era igual al tamaño de entrenamiento en todos los casos. En este último experimento, se usaron como consultas capítulos de libros en inglés seleccionados aleatoriamente para extraer libros completos (en chino).

Nótese que en la figura se observa que la aproximación de referencia mejora las dos variantes propuestas de nuestra metodología. Pero la diferencia entre sistemas es muy pequeña para considerarla estadísticamente significante ya que cae en medio del intervalo de la desviación estándar.

5 Conclusiones

Se ha propuesto un procedimiento de mapeado semántico no-lineal para implementar aplicaciones de extracción de documentos multilingües. El método se basa en usar una técnica de reducción de espacio no-lineal (escalado multidimensional) para construir agrupaciones multilingües de colecciones de documentos. En el método propuesto, se construye una agrupación independiente para cada idioma de la colección multilingüe y las similitudes entre las representaciones semánticas se aprovechan para realizar experimentos de extracción de documentos multilingüe.

Figura 4: Precisión de extracción vs el tamaño del set de entrenamiento, usando capítulos de libro como consultas para extraer libros.



Se han propuesto dos variantes del método, que se han implementado y evaluado comparando su resultado con una técnica clásica de extracción de información multilingüe. Todos los experimentos de extracción de información se han desarrollado utilizando una colección paralela trilingüe que incluye chino, castellano e inglés.

Las principales conclusiones del trabajo se pueden resumir como sigue:

1. Se han presentado resultados preliminares que muestran que las técnicas de mapeado semántico se pueden explotar para realizar extracción de información multilingüe.
2. La variante del método implementado que considera una matriz de proyección multilingüe presenta

mejores resultados que utilizando una matriz de proyección monolingüe.

3. Asimismo, la variante que considera una matriz de proyección multilingüe presenta mejores resultados que la técnica de referencia cuando se utilizan documentos enteros para la extracción. Una clara aplicación es cuando se utilizan los documentos como consulta en la búsqueda de patentes multilingüe o en la identificación de corpus comparables o paralelos.

4. Por último, las dos variantes de la metodología propuesta se comportan peor que el sistema de referencia cuando se utilizan consultas que son más cortas que el documento a extraer.

Con el propósito de continuar estudiando y explorando la potencialidad del método propuesto, se plantean las siguientes tres líneas de investigación:

1. Entender porqué la precisión del método se deteriora respecto al sistema de referencia cuando se usan textos parciales como consultas para la extracción.

2. Implementar un método no-lineal robusto de optimización para situar el documento y la consulta en los mapas de dimensión reducida ya calculados.

3. Explorar las diferentes alternativas para mejorar la precisión de la extracción mediante la combinación de mapas de idiomas diferentes.

Bibliografía

[1] Cox, M. F., Cox, M. A. A., 2001, Multidimensional Scaling, Chapman & Hall, UK.

[2] Deerwester, S., Dumais, S., Furnas, G. W., Landauer, T. K., Harshman, R., 1990, Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science, 41 (6): 391-407.

[3] Dumais, S. T., Landauer, T. K., Littman, M. L., 1996, Automatic cross-linguistic information retrieval using latent semantic indexing, in SIGIR96 Workshop on Cross-Linguistic Information Retrieval.

[4] Evans, D. A., Handerson, S. K., Monarch, I. A., Pereiro, J., Delon, L., Hersh, W. R., 1998, Mapping Vocabularies Using Latent Semantics, in Grefenstette, G. (ed.), Cross-Language Information Retrieval, Kluwer Academic Publishers, 63-80.

[5] Hofmann, T., 1999, Probabilistic Latent Semantic Analysis, in Proceedings of Uncertainty in Artificial Intelligence, UAI99, 289-296.

[6] Kishida, K., 2005, Technical issues of cross-language information retrieval: a review, Information Processing and Management, 41 (3): 433-455.

[7] Van Eck, N., Waltman, L., van den Berg, J., 2005, A novel algorithm for visualizing concept associations, in Proceedings of the 16th International Workshop on Database and Expert System Applications, 405-409.