Uso de mapas semánticos para la búsqueda crosslingüe de oraciones paralelas

Rafael E. Banchs and Marta R. Costa-jussà
Barcelona Media Innovation Center
Av Diagonal 177, 9th floor, 08018 Barcelona
{rafael.banchs,marta.ruiz}@barcelonamedia.org

31 de mayo de 2010

Resumen

Este trabajo presenta el uso de una técnica de recuperación de información crosslingüe basada en escalamiento multidimensional para la identificación de oraciones paralelas entre lenguas diferentes. El método propuesto permite hacer una reducción no-lineal del espacio de representación de las oraciones que se puede aprovechar para identificar similitudes semánticas entre conjuntos de oraciones en distintas lenguas. La técnica se ilustra con una colección pentalingüe extraída de la Constitución Española, la cual está disponible en las cuatro lenguas oficiales del Estado e inglés. Presentamos una evaluación comparativa entre nuestro método y un sistema de búsqueda crosslingüe basado en la traducción automática de las consultas. Los resultados muestran que nuestro sistema mejora consistentemente en las 20 direcciones experimentales de búsqueda crosslingüe que permite nuestra colección de datos.

1. Introducción

La recuperación de información crosslingüe permite obtener resultados en una lengua destino a partir de una búsqueda o consulta realizada en una lengua fuente diferente. Debido al aumento de información multilingüe en Internet, la necesidad de avanzar en esta área ha tomado especial relevancia recientemente. En particular, la investigación en recuperación de información crosslingüe se ha visto apoyada por eventos como el Cross-Language Evaluation Forum (CLEF) y el NTCIR Asian Language Evaluation.

Básicamente, hay dos maneras fundamentales de afrontar la recuperación de información crosslingüe: mediante técnicas basadas en traducción automática, y mediante técnicas basadas en interlingua (Kishida, 2005).

Por un lado, las técnicas basadas en traducción automática se usan para traducir, o bien la consulta, o bien la colección de documentos. En cuanquiera de estos dos casos, la recuperación de información crosslingüe se reduce a un problema de recuperación de información monolingüe. No obstante, debido a la enorme extensión que comúnmente presentan las colecciones multilingües, resulta más práctico realizar la traducción de la consulta (Chen and Bao, 2009).

Por otro lado, los métodos basados en interlingua se usan para asociar textos relacionados con contenidos en diferentes lenguas a través de representaciones semánticas que son independientes de la lengua. Algunas técnicas convencionales de recuperación de información crosslingüe basadas en interlingua usan indexación semántica latente (LSI) para construir una representación vectorial multilingüe de una colección paralela de documentos (Dumais et al., 1996). Una vez construída esta representación vectorial, se pueden proyectar en ella nuevos documentos y consultas, y la tarea de recuperación se lleva a cabo usando una métrica de similaridad o distancia.

Este trabajo, como extensión de (Banchs and Costa-jussa, 2009), propone usar una técnica de mapeado semántico para implementar un sistema de recuperación de información crosslingüe. Concretamente, la tarea a ser considerada consiste en la identificación crosslingüe de oraciones paralelas. Es decir, dada una oración en la lengua L_A, el objetivo es encontrar su equivalente en cualquier otra lengua diferente L_B. Esta tarea es importante para ciertas aplicaciones como, por ejemplo, la recopilación de texto paralelo a nivel de oración (Utiyama and Tanimura, 2007) o la detección de plagio (Potthast et al., 2009).

Las técnicas de mapeado semántico se han usado tradicionalmente para asociar conceptos y términos relacionados (Evans et al., 1998), y se ha verificado que el uso de proyecciones no-lineales en este contexto constituye una aproximación más eficiente que el uso de métodos lineales. La idea fundamental de la metodología propuesta es construir un mapa semántico para cada lengua disponible en la colección y aproximar el problema de recuperación de información crosslingüe mediante el aprovechamiento de las similaridades entre los diferentes mapas. La construcción de estos mapas, no obstante, requiere la disponibilidad de colecciones de documentos paralelos.

El presente artículo se organiza de la siguiente manera. En la sección 2 se describe la metodología de recuperación de información crosslingüe propuesta. En la sección 3 se ilustra dicha metodología mediante la realización de experimentos sobre la colección pentalingüe de la Constitución Española. También, en esta sección, se compara la metodología propuesta con la técnica de recuperación de informa-

ción crosslingüe basada en traducción automática de consultas, mostrando que la técnica propuesta supera a la de referencia. Finalmente, en la sección 4 se presentan las conclusiones más relevantes y los próximos pasos de nuestra investigación.

2. Mapeado semántico

La idea principal del método de recuperación de información crosslingüe propuesto se centra en el problema de mapeado semántico. En este trabajo, proponemos usar técnicas no-lineales de proyección, conocidas como escamiento multidimensional (MDS), para construir mapas semánticos de documentos[1] 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 mapas similares para las distintas lenguas. Esta sección explora e ilustra esta idea en profundidad.

El MDS constituye un método de proyección no-lineal para la visualización de datos que se puede usar también como una técnica de reducción de espacio (Cox and Cox, 2001). Dado un conjunto de relaciones entre los elementos de una 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 de la mejor manera posible. Lo interesante sobre MDS es que las relaciones entre los elementos de la colección en cuestión pueden ser tanto de naturaleza cuantitativa (similaridades basadas en una métrica de distancia) como de naturaleza de cualitativa (relaciones ordinales o jerárquicas).

Dada una colección monolingüe, el proceso de cómputo de un mapa semántico mediante MDS se puede definir en tres pasos:

- obtener una representación vectorial para la colección mediante el estándard TF-IDF (Salton and Buckley, 1988);
- construir una matriz de similaridad para los documentos usando las distancias coseno (o alguna otra métrica similar) entre sus correspondientes vectores; y
- construir un mapa semántico de dimensión reducida para los documentos en la colección usando MDS.

[1] Tal y como se indicó en la introducción, la tarea a ser considerada en el presente trabajo es la identificación crosslingüe de oraciones paralelas; en este sentido, cuando hablamos de documentos en la presente sección y subsiguientes secciones del trabajo nos estaremos refiriendo realmente a oraciones.

Por otro lado, el procedimiento propuesto para la implementación de un sistema de recuperación de información crosslingüe mediante el uso de mapas semánticos se puede resumir en los siguientes tres pasos (Banchs and Costa-jussa, 2009):

- seleccionar con una colección multilingüe de documentos (conjunto de entrenamiento) y construir el mapa de recuperación usando MDS a partir de una de las lenguas;
- proyectar nuevos documentos y consultas desde cualquier lengua fuente en el mapa de recuperación usando proyecciones lineales diseñadas para tal fin; y
- extraer documentos relevantes para una consulta del mapa de recuperación usando una métrica de distancia.

La figura 1 muestra una representación esquemática de este procedimiento.

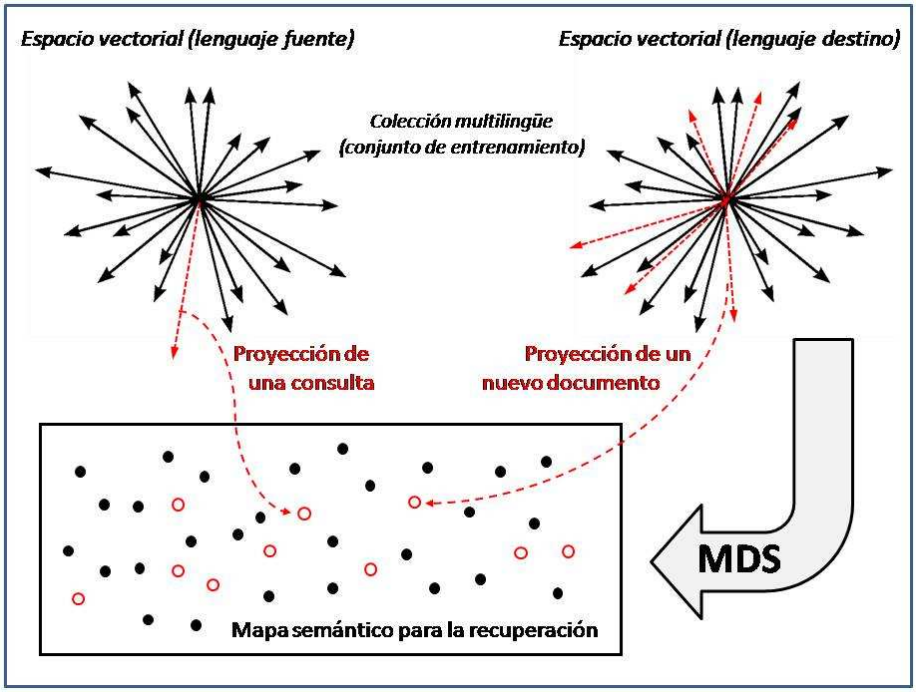

Figura 1: Representación esquemática del método de recuperación de información crosslingüe basado en mapas semánticos

Una vez construido el mapa semántico de recuperación, podemos inferir una transformación lineal T para proyectar en dicho mapa documentos o consultas desde un espacio original en cualquier lengua, como sigue:

$$M = TD \longrightarrow T = MD^{-1} \tag{1}$$

donde D es una matriz cuadrada NxN que contiene las distancias entre los documentos de entrenamiento en el espacio original (matriz de similaridad de documentos) y M es una matriz KxN que contiene las coordenadas de los documentos de entrenamiento una vez proyectados en el mapa semántico de dimensión reducida, siendo N el número total de documentos contenido en el conjunto de entrenamiento y K la dimensionalidad del mapa semántico construido. Tal y como se explica en (Banchs and Costa-jussa, 2009), la transformación lineal descrita en (1) puede ser de naturaleza crosslingüe, para lo cual M se debe calcular en la lengua de recuperación y D se debe calcular en la lengua original del documento o consulta que se desea proyectar en el mapa. De esta forma, cualquier nuevo documento o consulta se puede situar en el mapa de recuperación utilizando la siguiente expresión:

$$m = Td \tag{2}$$

donde d representa un vector que contiene las distancias entre el documento (o consulta) a proyectar y el conjunto de documentos de entrenamiento en la dimensión original, T es la matriz de transformación definida en (1), y m es el vector que contiene el resultado de las coordenadas del documento (o consulta) a proyectar en el mapa semántico de dimensión reducida.

Finalmente, como se pueden generar tantos mapas de recuperación como lenguas diferentes hay en la colección multilingüe, proponemos una nueva variante de nuestro método que combina diferentes mapas usando una votación. De acuerdo con esta estrategia, se construye un mapa de recuperación para cada lengua de la colección. A continuación, cada uno de los documentos y consultas del conjunto de evaluación son proyectados sobre cada uno de los mapas disponibles y se calculan las similaridades entre consultas y documentos. Como resultado se obtiene un ordenamiento para cada consulta, y por cada mapa de recuperación disponible, de todos los documentos que constituyen el conjuto de evaluación. Finalmente, se construye un ordenamiento global para cada consulta mediante la votación mayoritaría de cada uno de los ordenamientos obtenidos para dicha consulta en los distintos mapas.

3. Experimentos

Como mencionamos en la introducción, en este trabajo nos centramos en la tarea de la identificación crosslingüe de oraciones paralelas. En esta tarea en particular, se usa como consulta una oración en una lengua fuente, y se pretende recuperar esta misma oración en una o varias lenguas destino diferentes.

3.1. Colección de datos

La colección de datos utilizada es un conjunto pentalingüe de oraciones que se extrajo de la Constitución Española [2]. Las cinco lenguas en las que está disponible esta colección son: castellano, catalán, gallego, euskera e inglés.

Los textos de la Constitución están organizados en 169 artículos más algunas disposiciones reguladoras adicionales. Para la tarea, segmentamos todos los textos en oraciones. Se aplicó un filtro por longitud para eliminar todas aquellas oraciones de menos de 5 palabras. Esto se hizo con el fin de eliminar títulos y cualquier otro tipo de información no relevante. Posteriormente se llevó a cabo una randomización de las oraciones resultantes y se seleccionaron 200 oraciones (con sus respectivas paralelas en todas la lenguas) como conjunto de evaluación. Las tablas 1 y 2 resumen respectivamente las principales estadísticas de la colección completa y del conjunto de evaluación seleccionado. Finalmente, la tabla 3 muestra un ejemplo específico de oración extraído de la colección multilingüe.

Colección completa	Inglés	Castellano	Catalán	Euskera	Gallego
Num. oraciones	611	611	611	611	611
Num. de palabras	15285	14807	15423	10483	13760
Vocabulario	2080	2516	2523	3633	2667
Long. media oración	25.01	24.23	25.24	17.16	22.52

Tabla 1: Estadísticas básicas del conjunto de datos completo.

3.2. Evaluación comparativa

En esta subsección, comparamos la metodología propuesta con el método de recuperación de información crosslingüe basado en la traducción automática de las consultas (Chen and Bao, 2009). Como ya se mencionó, la tarea consiste en recuperar una oración en cualquiera de las 5 lenguas disponibles usando la misma oración en cualquiera de las otras 4 lenguas como consulta. La calidad de los

[2] www.la-moncloa.es

Colección de test	Inglés	Castellano	Catalán	Euskera	Gallego
Num. oraciones	200	200	200	200	200
Num. de palabras	4667	4492	4669	3163	4175
Vocabulario	1136	1256	1273	1618	1316
Long. media oración	23.34	22.46	23.34	15.82	20.88

Tabla 2: Estadísticas básicas del conjunto de datos de evaluación.

Lengua	Ejemplo de oración
Inglés	The capital of the State is the city of Madrid.
Castellano	La capital del Estado es la villa de Madrid.
Catalán	La capital de l'Estat és la vila de Madrid.
Euskera	Estatu hiriburua Madrid hiria da.
Gallego	A capital do Estado é a vila de Madrid

Tabla 3: Ejemplo de oración de la Constitución Española extraído de la colección multilingüe.

sistemas se evalúa en términos de los aciertos en la primera posición (top-1) y las primeras cinco posiciones (top-5) de la lista de oraciones recuperadas para cada consulta, usando el conjunto de evaluación descrito en la tabla 2.

Para construir el sistema de recuperación de información crosslingüe basado en MDS, usamos 400 oraciones paralelas seleccionadas al azar de las 411 que quedaron al extraer las 200 correspondientes al conjunto de evaluación. Esta subcolección de 400 oraciones se usó para construir los mapas. Se construyó un mapa para cada una de las 5 lenguas disponibles en la colección, fijando la dimensión de los mapas en 350, lo cual implica una reducción de dimensionalidad entre un 83 % (en el caso del inglés) y un 90 % (en el caso del euskera). Siguiendo (1) y (2), se construyeron las matrices de transformación y se ubicaron todas las oraciones de evaluación en las 5 lenguas en cada uno de los mapas. Finalmente, se extrajeron las oraciones de evaluación más similares entre sí usando la distancia coseno como métrica de similaridad, obteniendo un ordenamiento individual para cada mapa. Dados estos ordenamientos individuales, se implementó la combinación por votación mayoritaria y se obtuvo un ordenamiento global. La tabla 4 muestra los resultados para este ordenamiento global.

Como sistema de referencia usamos el sistema basado en traducción de consultas (Chen and Bao, 2009). Este sistema implementa la recuperación de información crosslingüe concatenando un sistema de traducción automática con un sistema de búsqueda monolingüe. En nuestro caso, para la traducción de las consultas usamos

la plataforma $Opentrad$[3] (Ramírez-Sánchez et al., 2006). Este sistema proporciona traducciones en el estado-del-arte entre las lenguas oficiales del estado español, el inglés y otras lenguas adicionales. Hay que tener en cuenta que este sistema no proporciona traducciones directamente desde cualquier lengua a cualquier otra. Así pues, en los casos de gallego-catalán tuvimos que hacer una traducción en cascada pasando por el castellano. Adicionalmente, en el caso del euskera, el sistema proporciona traducciones pero sólo a nivel de oración, no pudiéndose automatizar para la traducción de de conjuntos de oraciones. Por esta razón, las 200 oraciones del euskera no fueron consideradas para los experimentos con este método. Por otro lado, el sistema recuperación de información monolingüe se implementó usando $SOLR$, que es una plataforma de código abierto basado en la librería $Lucene$[4].

La tabla 4 resume los resultados de los dos sistemas evaluados, tanto el propuesto (MDS) como el basado en traducción de consultas (MT-IR), para las 25 direcciones posibles y para cada una de las métricas de evaluación usadas (top-1 y top-5).

Lengua fuente	Sistema	Lengua destino									
		Inglés		Castellano		Catalán		Euskera		Gallego	
		top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
Inglés	MT-IR	100	100	95.0	99.5	92.0	96.0	-	-	93.0	96.0
	MDS	100	100	**96.0**	**99.5**	**96.5**	**100**	**75.5**	**91.0**	**95.5**	**99.0**
Castellano	MT-IR	96.0	99.0	100	100	100	100	-	-	99.0	100
	MDS	**97.5**	100	100	100	100	100	**78.0**	**94.5**	**99.5**	100
Catalán	MT-IR	95.5	99.0	100	100	100	100	-	-	93.5	97.0
	MDS	**96.5**	**99.5**	100	100	100	100	**74.0**	**93.5**	**99.5**	**99.5**
Euskera	MT-IR	-	-	-	-	-	-	100	100	**79.5**	**93.5**
	MDS	**76.0**	**92.5**	**81.5**	**94.5**	**79.0**	**94.5**	100	100	**79.5**	**93.5**
Gallego	MT-IR	93.5	97.5	99.5	99.5	83.5	90.5	-	-	100	100
	MDS	**94.5**	**99.5**	**99.5**	**99.5**	**99.5**	**99.5**	**77.0**	**95.0**	100	100

Tabla 4: Resultados comparativos entre el sistema propuesto (MDS) y el basado en traducción de consultas (MT-IR).

De la tabla 4 se pueden extraer las siguientes observaciones:

- Usando las dos métricas propuestas (top-1 y top-5), el método propuesto (MDS) mejora la técnica de referencia en todos los casos.
- Se observa que el método de traducción de consultas se ve perjudicado en los casos en que se usa una concatenación de sistemas de traducción (específicamente para el par catalán-gallego). La traducción en cascada generalmente

[3] www.opentrad.com

[4] http://lucene.apache.org/solr/tutorial.html

concatena errores con lo que se obtiene una caída en la calidad de traducción. Así pues, en estos experimentos vemos como el comportamiento de este método depende de la calidad de la traducción.

- Se observa que, en general, los sistemas de recuperación de información crosslingüe obtienen mejores resultados cuanto más semejanza tienen las lenguas (ver el caso de las lenguas romances castellano-catalán-gallego respecto a los pares de lenguas más distantes castellano-euskera o gallego-inglés).

Después de hacer un análisis de aquellas oraciones que no recuperan correctamente su oración paralela para cada uno de los métodos utilizados, hemos descubierto que ambos métodos se equivocan siempre en oraciones diferentes. Esto se puede justificar por la naturaleza diferente de ambos métodos. De hecho, un análisis más detallado revela que la naturaleza de los errores cometidos por cada método son muy diferentes entre sí.

La tabla 5 presenta un par de ejemplos de salidas erróneas del sistema de referencia (MT-IR) y del sistema propuesto (MDS), para el caso en el que el castellano era la lengua fuente y el inglés era la lengua destino. En el primer caso, la consulta es la salida del traductor. En el segundo caso, la consulta es la oración en castellano de la Constitución. Podemos observar que en el caso del MT-IR, los errores parecieran derivarse de la existencia de palabras poco frecuentes que tiene un gran peso (en términos de TF-IDF) y que coinciden en la consulta y la oración recuperada. En cambio, en el caso del MDS, se puede observar una mayor semejanza entre la oración que devuelve el sistema y la que debería haber devuelto tanto en términos de palabras coincidentes, como del tema central de la oración y su semántica.

4. Conclusiones y trabajo futuro

Se ha presentado un procedimiento de mapeado semántico para identificar oraciones paralelas en distintas lenguas. El método se basa en el uso de una técnica no-lineal de reducción de espacio (escalamiento multidimensional) para identificar relaciones semánticas entre oraciones dada una colección multilingüe. Adicionalmente, el método propuesto implementa una combinación de los resultados obtenidos con los distintos mapas construídos, la cual permite un mejor aprovechamiento de la información multilingüe de la que se dispone. Se ha evaluado y comparado nuestro método con una estrategia estándar de recuperación de información crosslingüe, la basada en traducción de consultas, sobre una colección pentalingüe extraída de la Constitución Española. Los resultados demuestran que la técnica propuesta mejora consistentemente la técnica de referencia.

MT-IR	
CONSULTA:	3.ª Ordenación del territorio, urbanismo y vivienda.
SALIDA MT:	3.ª Ordenación Of the **territory**, urbanismo and house.
SALIDA IR:	The Constitutional Court has jurisdiction over the whole Spanish **territory** and is entitled to hear:
REFERENCIA:	3.ª Town and country planning and housing.
CONSULTA:	La ley electoral determinará las causas de inelegibilidad e incompatibilidad de los Diputados y Senadores, que comprenderán en todo caso:
SALIDA MT:	The electoral **law will** determine the causes of inelegibilidad and incompatibility of the Deputies and Senators, that **will** comprise, anyway:
SALIDA IR:	Once the Statute is approved, the King **will** sanction it and promulgate it as **law**
REFERENCIA:	The Electoral Act shall establish grounds for ineligibility and incompatibility for Members of Congress and Senators, which shall in any case include those who are:
MDS	
CONSULTA:	La declaración de inconstitucionalidad de una norma jurídica con rango de ley, interpretada por la jurisprudencia, afectará a ésta, si bien la sentencia o sentencias recaídas no perderán el valor de cosa juzgada.
SALIDA:	**Those declaring the unconstitutionality of** an act or of a statute **with the force of an act and** all those which are **not limited to** the acknowledgment of an individual right, shall be fully binding on all persons.
REFERENCIA:	**A declaration of unconstitutionality of** a legal provision **having the force of an act and** that has already been applied by the Courts, shall also affect the case-law doctrine built up by the latter, but the decisions handed down **shall not lose** their status of res judicata.
CONSULTA:	Las Comunidades Autónomas designarán además un Senador y otro más por cada millón de habitantes de su respectivo territorio.
SALIDA:	Under no circumstances **shall** a federation of **Self-governing Communities** be allowed
REFERENCIA:	The **Self-governing Communities shall**, in addition, appoint one Senator and a further Senator for every million inhabitants in their respective territories.

Tabla 5: Análisis de las salidas de los métodos de referencia (MT-IR) y propuesto (MDS).

Como trabajo futuro proponemos seguir explorando la utilidad y beneficios de la técnica propuesta en las siguientes direcciones:

- la evaluación y prueba de su rendimiento en un escenario de entrenamiento en el que se utilice una colección multilingüe de documentos comparables, en lugar de paralelos;
- el estudio y evaluación de otras técnicas alternativas para la proyección de consultas y nuevos documentos en los mapas semánticos previamente construídos; y
- la evaluación del rendimiento del método propuesto, tanto en términos de calidad como de eficiencia, al utilizar colecciones paralelas de datos de mayor tamaño.

Bibliografía

R. E. Banchs and M. R. Costa-jussa. 2009. Extracción crosslingüe de documentos usando mapas semánticos no-lineales. *Revista del Procesamiento del Lenguaje Natural, SEPLN*, 43:169–176.

J. Chen and Y. Bao. 2009. Cross-language search: The case of google language tools. *First Monday*, 14(3-2).

M.F. Cox and M.A.A. Cox. 2001. *Multidimensional Scaling*. Chapman and Hall.

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

D.A. Evans, S.K. Handerson, I.A. Monarch, J. Pereiro, L. Delon, and W.R. Hersh. 1998. Mapping vocabularies using latent semantics. *G. Grefenstette (ed.) Cross-Language Information Retrieval*, pages 63–80.

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

M. Potthast, B. Stein, A. Eiselt, A. Barrón, and P. Rosso. 2009. Overview of the 1st international competition on plagiarism detection. In *Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse*.

G. Ramírez-Sánchez, F. Sánchez-Martínez, S. Ortiz-Rojas, J. A. Pérez-Ortiz, and M. L. Forcada. 2006. Opentrad apertium open-source machine translation system: an opportunity for business and research. In *Proceeding of Translating and the Computer 28 Conference*, November.

G. Salton and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5).

M. Utiyama and M. Tanimura. 2007. Automatic construction technology for parallel corpora. *Journal of the National Institute of Information and Communications Technology*, 54(3):25–31.