Wikicardi: Hacia la extracción de oraciones paralelas de Wikipedia

Reporte técnico LSI-14-3-R

Josu Boldoba, Alberto Barrón-Cedeño y Cristina España-Bonet

Talp Research Center, Universitat Politècnica de Catalunya
[jboldoba,albarron,cristinae]@lsi.upc.edu

10 de marzo de 2014

Resumen

Uno de los objetivos del proyecto Tacardi (TIN2012-38523-C02-00) consiste en extraer oraciones paralelas de corpus comparables para enriquecer y adaptar traductores automáticos. Nosotros consideramos un subconjunto de Wikipedia como corpus comparable. En este reporte se describen nuestros avances con respecto a la extracción de fragmentos paralelos de Wikipedia. Primero, discutimos cómo hemos definido los tres dominios de interés —ciencia, informática y deporte—, en el marco de la enciclopedia y cómo hemos extraído los textos y demás datos necesarios para la caracterización de los artículos en las distintas lenguas. Después discutimos brevemente los modelos que usaremos para identificar oraciones paralelas y damos sólo una muestra de algunos resultados preliminares. Los datos obtenidos hasta ahora permiten vislumbrar que será posible extraer oraciones paralelas de los dominios de interés a corto plazo, si bien aún no contamos con una estimación del volumen de éstos.

1. Introducción

La traducción automática es una tarea dentro del procesamiento del lenguaje natural con gran impacto en el uso diario. Servicios web de traducción como el de Bing[1] o Google[2] cuentan con numerosos usuarios y son dos ejemplos de motores de traducción basados en traducción automática estadística.

El principal recurso necesario para el desarrollo de estos traductores es un corpus paralelo. Es decir, una colección de pares de oraciones que son traducciones una de la otra. Existen diversas colecciones de textos con estas características; por ejemplo el corpus del Parlamento Europeo, *Europarl* (Koehn, 2005), el de las Naciones Unidas, *UN* (Rafalovitch y Dale, 2009), o el *Opus* corpus (Tiedemann, 2012). Sin embargo, los textos en dichas colecciones suelen cubrir temas

[1] http://www.bing.com/translator
[2] http://translate.google.com/

Tabla 1: Número de artículos comparables entre ediciones de Wikipedia en distintas lenguas. La diagonal principal muestra la cantidad total de artículos en cada lengua.

	ca	**en**	**es**	**eu**
ca	$\sim 400K$	277 836	263 481	95 643
en		$\sim 4{,}5M$	631 710	124 401
es			$\sim 1{,}1M$	114 673
eu				$\sim 166K$

y géneros específicos (e.g., política, economía o noticias). Dicha desviación puede condicionar la traducción de un texto sobre temáticas distintas, por ejemplo deportes, usando vocabulario inadecuado adquirido en dominios muy distintos. Por esta razón es necesario recurrir a otros recursos para obtener más ejemplos paralelos.

En nuestra investigación hemos elegido Wikipedia como fuente de potenciales oraciones paralelas. Wikipedia, o al menos un subconjunto de ella, es uno de los mejores ejemplos de corpus comparable: una colección de documentos que incluye entradas sobre el mismo tema en distintos idiomas. Nuestro principal objetivo es, dados dos artículos en distintas lenguas sobre un tema común, extraer aquellos pares de oraciones que conformen un texto paralelo; es decir, que sean traducciones una de la otra.

Actualmente nos centramos en cuatro lenguas: castellano, catalán, inglés y euskera. Nos referiremos a ellas como *es*, *ca*, *en* y *eu* respectivamente. En cuanto al dominio, estamos interesados en tres: ciencia, informática y deporte (*sc*, *cs* y *sp*). En este documento nos centramos en los resultados obtenidos para el par en-es en los tres dominios mencionados.

El objetivo del trabajo es pues extraer oraciones paralelas de un conjunto de artículos de Wikipedia para enriquecer un traductor automático estadístico estándard y adaptarlo a los dominios considerados. Para ello, es necesario construir la colección de documentos comparables de Wikipedia en las categorías en cuestión y estimar la similitud entre sus frases.

El resto del reporte está distribuido de la siguiente manera. La sección 2 incluye una descripción del corpus que utilizamos. La sección 3 muestra un experimento real y una explicación de los resultados obtenidos. La sección 4 explica el trabajo que se está llevando a cabo actualmente y concluye el reporte.

2. Corpus

Acceder a los artículos de Wikipedia es relativamente sencillo, gracias a la disponibilidad de los respaldos realizados periódicamente por la fundación Wikimedia[3]. La tabla 1 incluye algunas estadísticas para los pares de lenguas que nos atañen. Únicamente consideramos aquellos artículos que pertenecen al espacio principal de nombres de Wikipedia; es decir, artículos, anexos y listas (hemos descartado las páginas de desambiguación). Actualmente estamos trabajando con un conjunto de respaldos de julio de 2013.

[3] `http://dumps.wikimedia.org/`

2.1. Definición de dominios

Dado que nuestra intención es trabajar con los subconjuntos de artículos de cada dominio por separado, es necesario discriminarlos previamente. Sin embargo, la definición de dominios en Wikipedia no es una tarea trivial. La taxonomía de áreas y dominios de esta enciclopedia se basa principalmente en una colección de categorías que, al igual que los propios contenidos de los artículos, pueden ser creadas y manipuladas por los propios editores voluntarios. Si bien estas categorías permiten crear una taxonomía bastante rica en la cual distribuir los más diversos artículos, ha provocado que la tarea de obtener todos los artículos de un cierto dominio sea complicada[4].

Formalmente, la decisión de que un artículo pertenezca a un dominio o no puede definirse de la siguiente manera. Sea a un artículo de Wikipedia. Sea C_a el conjunto de categorías asociadas a a. Sea d el dominio de interés. Consideramos que a pertenece al dominio d si al menos una de las categorías en C_a pertenece a dicho dominio. Para determinar si una categoría c pertenece a un dominio d, hemos explorado la taxonomía completa de categorías de Wikipedia: un grafo dirigido relativamente denso y con ciclos[5].

El planteamiento consiste en explorar el grafo de categorías a partir de una raíz determinada —las propias categorías ciencia, informática y deporte. Nuestro método de exploración está inspirado en el de Cui *et al.* (2008), el cual propone una búsqueda en anchura con memorización de nodos visitados en la que un artículo se puntúa para cuantificar cuál es su relación con el dominio de interés. Sin embargo, nuestro objetivo es definir las categorías que pertenecen a un dominio, y no directamente los artículos.

Nuestra búsqueda en anchura de las categorías de un dominio no ha generado los resultados deseados. En el caso particular del inglés, las colecciones de categorías obtenidas a partir de las tres distintas raíces son idénticas (915 619). Es decir, el grafo era explorado por completo. Para la edición en castellano las categorías relacionadas con informática (206 546) son un subconjunto de las de ciencia (206 550), mientras que las de deporte está claramente separado (37 319). Esto nos ha llevado a definir dos estrategias:

1. Explorar el grafo completo y seleccionar únicamente aquellas categorías cuyos nombres incluyan *al menos* una palabra del vocabulario del dominio.

2. Explorar el grafo hasta que un porcentaje mínimo de las categorías en cierto nivel incluyan al menos una palabra del vocabulario del dominio y seleccionar todas las categorías desde dicho nivel hasta la raíz.

Para definir el vocabulario de un dominio, hemos procesado todos aquellos artículos que pertenecían a la categoría raíz del mismo. La tabla 2 muestra la cantidad de artículos asociados a cada categoría raíz[6]. Hemos descartado las

[4] Por ejemplo, las categorías "Futbolistas del Fútbol Club Barcelona 'C'" y "Pato" tendrían que ser consideradas dentro del dominio *deporte*. La consideración de la segunda resulta claramente más difícil.

[5] Una taxonomía para este tipo de material no debería tener ciclos. Los propios editores de la Wikipedia en castellano los consideran una consecuencia de que, a menudo, un editor confunde los conceptos de categoría y etiqueta e incluyen categorías generales como subcategorías de otras más específicas. Llaman a este fenómeno "categorización circular" (Wikipedia, 2014c).

[6] Resulta interesante observar que tanto ciencia como informática son consideradas "contenedor general, utilizado para organizar categorías más precisas" (Wikipedia, 2014a,b). Ello

Tabla 2: Número de artículos pertenecientes a la categoría raíz para cada uno de los dominios en inglés y castellano en julio de 2013.

	sc	**cs**	**sp**
en	29	4	3
es	3	130	10

palabras de paro, números y signos de puntuación de los textos. Además, hemos aplicado un proceso de *stemming* (Porter, 1980). Finalmente, hemos seleccionado sólo aquellos tokens con al menos dos caracteres y los hemos ordenado con base en su frecuencia (el conocido modelo de frecuencia de término, tf). El 10 % de los tokens con mayor frecuencia ha sido seleccionado para representar el vocabulario del dominio.

Una exploración visual nos permitió ver que la primera estrategia ignoraba muchas categorías relacionadas con el dominio a la vez que incluía muchas categorías que no lo estaban. Por ejemplo, la categoría `Pato`[7] es ignorada erróneamente a pesar de que en la Wikipedia en castellano se refiere en realidad a un deporte y no a un animal. Por otro lado, la categoría `Sistema circulatorio`[8] se incluía dentro del dominio de informática, al igual que otras estructuras anatómicas, también llamadas sistemas, debido a que sistemas es considerada un término de este dominio.

Por lo tanto, nos hemos decantado por la segunda estrategia, considerando dos umbrales de decisión: 50 % y 60 %. De nuevo, una exploración visual nos permitió observar que tanto la estrategia como los umbrales resultaban en una selección de categorías relativamente satisfactoria. La figura 1 resume los números obtenidos con los dos umbrales en las dos lenguas. La relajación en un 10 % del umbral de aceptación supone ampliar considerablemente el número de artículos recuperados sin recuperar demasiados textos no relacionados (si bien para el castellano no existe diferencia para las categorías ciencia e informática). Así, con el afán de perder la menor cantidad de artículos relevantes, hemos optado por utilizar el umbral más flexible: 50 %.

Los pares de artículos obtenidos de esta manera son: 161 130 para ciencia, 18 168 para informática y 72 315 para deporte.

2.2. Extracción de contenidos

Los modelos de similitud a considerar requieren distintas caracterizaciones. Por ejemplo, el modelo basado en n-gramas (Mcnamee y Mayfield, 2004) requiere texto plano (cf. sección 3), mientras que un método basado en enlaces que probaremos en el futuro (Adafre y de Rijke, 2006) requiere *wikitexto*. Por lo tanto, hemos extraído las oraciones de los artículos en esos dos formatos. El procedimiento de extracción de contenidos de un artículo se resume de la siguiente forma:

1. Extracción del texto completo del artículo, ya sea en formato plano o wikitexto, por medio de la biblioteca JWPL (Zesch *et al.*, 2008).

implica que se sugiere no asignar artículos a ellas, sino usar alguna subcategoría. Claramente esta recomendación no se ha seguido en el caso de la categoría informática.

[7] `http://es.wikipedia.org/wiki/Categora:Pato`

[8] `http://es.wikipedia.org/wiki/Categora:Sistema_circulatorio`

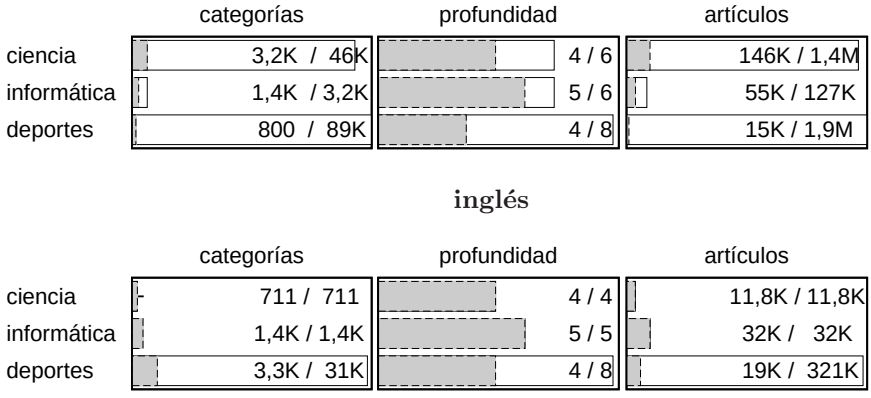

Figura 1: Visión general de la selección de categorías y artículos del dominio. Incluye la cantidad de categorías, profundidad del árbol y cantidad de artículos con umbral de 60 %/50 % (barra obscura y clara, respectivamente). Observe que las barras de *deporte* correspondientes a categorías y artículos en inglés con umbral=50 % van más allá de la gráfica.

2. Eliminación de metadatos de Wikipedia incluyendo *infoboxes* (tablas con datos puntuales del artículo en cuestión), citas y enlaces a artículos en otras lenguas.

3. Separación de oraciones por medio de los métodos disponibles en la biblioteca *OpenNLP*[9].

Este procedimiento genera la entrada a los procesos de cálculo de similitudes y selección de potenciales oraciones paralelas.

3. Experimentos preliminares

Hemos seleccionado una pequeña muestra de artículos comparables para realizar algunos experimentos preliminares basados en la similitud del coseno con dos caracterizaciones: 3-gramas de caracteres (Mcnamee y Mayfield, 2004) y pseudo-cognados (Simard *et al.*, 1992).

La caracterización basada en 3-gramas de caracteres consiste en eliminar signos de puntuación y diacríticos para luego romper los textos en secuencias solapadas de 3 caracteres. Por ejemplo, los 3-gramas de carácter de "*Esta pequeña frase.*" son `est, sta, ta_ , a_p, _pe, peq, equ, que, uen, ena, na_ , a_f, _fr, fra, ras` y `ase` (hemos sustituido los espacios por guiones bajos a efectos de visualización). Como puede observarse, los caracteres son convertidos a minúscula.

En la caracterización basada en pseudo-cognados se descartan de nuevo los signos de puntuación y los diacríticos y, a grandes rasgos, los tokens se conservan dentro de la caracterización si (i) su longitud es mayor o igual a 4 caracteres (si

[9] `http://opennlp.apache.org/`

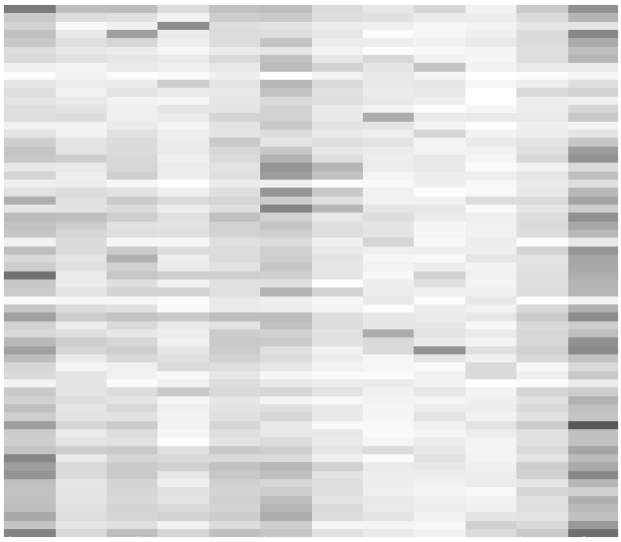

Figura 2: Mapa de calor que representa la matriz de similitudes entre oraciones del artículo en inglés *GameSpy Industries* (eje y) y su comparable en castellano *GameSpy* (eje x); caracterización basada en n-gramas de carácter. El intervalo de colores va de blanco ($sim = 0$ o nula) a través de la escala de grises hasta negro ($sim = 1$ o identidad).

es mayor, sólo se conservan los cuatro primeros), (*ii*) es un número, o (*iii*) es una combinación de letras y números. Por ejemplo, la representación de "*del virus H1N1*" es `viru` y `h1n1`. Una vez más, los caracteres son convertidos a minúscula.

Dado nuestro afán es extraer toda oración paralela dentro del par de documentos paralelos $\{a_{en}, a_{es}\}$, calculamos una matriz de similitud de $n \times m$, donde n (m) es el número de oraciones en a_{en} (a_{es}). Suponemos que aquellos pares con la mayor similitud son aquellos que son buenos candidatos a representar traducciones. La figura 2 muestra una de las matrices de similitud resultantes en forma de mapa de calor. Las frases correspondientes a la primer celda, cuya intensidad es relativamente alta, son:

en GameSpy Industries, Inc., known simply as GameSpy, was a former division of IGN Entertainment, which operates a network of game websites and provides online video game-related services and software.

es GameSpy Industries, Inc., más conocida simplemente como GameSpy, es una división de IGN Entertainment.

las cuales pueden considerarse oraciones paralelas con ruido.

La tabla 3 muestra los pares de oraciones más similares entre los artículos sobre "Edsger Wybe Dijkstra". La caracterización en este ejemplo está también

Tabla 3: Pares de oraciones con mayor similitud sobre la caracterización basada en n-gramas de los artículos sobre *Edsger Wybe Dijkstra* en inglés y castellano. La tabla muestra la similitud entre las oraciones en las dos lenguas, en las que los fragmentos paralelos se destacan en negritas.

0.58	**He retired in 2000.**
	Se retiró en 2000.
0.47	**One starts with a mathematical specification of what a program is supposed to do and applies mathematical transformations to the specification until it is turned into a program that can be executed.**
	Uno comienza con una especificación matemática del programa que se supone va a hacer y aplica transformaciones matemáticas a la especificación hasta que se transforma en un programa que pueda ser ejecutado.
0.45	**Among his contributions to computer science are a shortest path algorithm, known as Dijkstra's algorithm; the Shunting yard algorithm; the THE multiprogramming system**, an important early example of structuring a system as a set of layers; **the Banker's algorithm; and the semaphore construct for coordinating multiple processors and programs.**
	Entre sus contribuciones a las ciencias de la computación está la solución del problema del camino más corto, también conocido como el algoritmo de Dijkstra, la notación polaca inversa y **el relacionado algoritmo shunting yard, THE multiprogramming system, el algoritmo del banquero y la construcción del semáforo para coordinar múltiples procesadores y programas.**

basada en n-gramas. Éste es un ejemplo en el que una caracterización sencilla ha permitido extraer buenos candidatos: dos oraciones paralelas y una comparable. Tenemos planeado estudiar cuál puede ser el impacto de considerar pares como el tercero para entrenar un traductor. En caso de tener un número elevado de frases comparables no paralelas los aliniamentos automáticos entre fragmentos más pequeños se podrían ver dañados, obteniendo así ruido en los modelos de traducción. Utilizar distintos valores límite para seleccionar frases paralelas permitirá incluir más o menos pares comparables en el corpus de entrenamiento y así se podrá ver qué efecto tienen en la traducción final.

Como ejemplo de un caso de frases paralelas, el primer par de oraciones, vemos las caracterizaciones en el vector de 3-gramas:

en { `he_, e_r, _re, ret, eti, tir, ire, red, ed_, d_i, _in, in_, n_2, _20, 200, 000` }

es { `se_ e_r, _re, ret, eti, tir, iro, ro_, o_e, _en, en_, n_2, _20, 200, 000` }

La intersección es significativa: {`e_r, _re, ret, eti, tir, n_2, _20, 200, 000`}; motivo por el cual la similitud es alta.

La tabla 4 muestra los pares de oraciones más similares entre los artículos comparables sobre "António Carlos Silva", del dominio deporte. Esta vez la caracterización se basa en el modelo de pseudo-cognados. Si bien de nueva cuenta los pares tienden al paralelismo, el segundo incluye algo de ruido en el fragmento en inglés. Éste es un caso particular en el que el problema no recae en la medida de similitud, sino en el preproceso: el fragmento en inglés está erroneamente compuesto por dos oraciones.

Tabla 4: Pares de oraciones con mayor similitud sobre la caracterización basada en pseudo-cognados los artículo sobre *Antônio Carlos Silva* en inglés y castellano. La tabla muestra la similitud entre las oraciones en las dos lenguas, en las que los fragmentos paralelos se destacan en negritas.

0.49	**Silva next faced Alistair Overeem on February 2, 2013 at UFC 156.**
	Silva se enfretaría ante Alistair Overeem el 2 de febrero de 2013 en UFC 156.
0.42	Cormier knocked Silva out with standing punches at 3:56 of round 1. **Ultimate Fighting Championship On January 7, 2012, Antônio Rodrigo Nogueira told "Portal do Vale Tudo"Silva had signed a UFC contract.**
	Ultimate Fighting Championship El 7 de enero de 2012, Antonio Rodrigo Nogueira dijo en el "Portal do Vale todo"que Silva había firmado un contrato con UFC.
0.38	**Silva faced Travis Browne on October 5, 2012 at UFC on FX 5. Early in the fight Browne injured his hamstring, limiting his movement.**
	Silva enfrentaría a Travis Browne el 5 de octubre de 2012 en UFC on FX 5. A principios de la lucha, Browne se lesionó el tendón, lo que limitó en su movimiento.

Miremos de nuevo la intersección entre las caracterizaciones del primer par como ilustración. Las caracterizaciones basadas en pseudo-cognados son:

en { `silv, next, face, alis, over, febr, 2, 2013, 156` }

es { `silv, enfr, ante, alis, over, 2, febr, 2013, 156` }

cuya intersección está compuesta por {`silv, alis, over, febr, 2, 2013, 156` }.

Esperamos que la combinación de estos y otros modelos permita filtrar mejor los fragmentos paralelos de los que no lo son.

4. Observaciones finales

En este reporte hemos descrito nuestro trabajo actual avocado a la extracción automática de oraciones paralelas de Wikipedia para el enriquecimiento de un traductor automático estadístico. Dado que estamos interesados en tres dominios —ciencia, informática y deporte—, hemos discutido un método para definir dominios dentro de la taxonomía categórica de Wikipedia. Acto seguido, hemos descrito el proceso de extracción de textos, así como algunos resultados preliminares relacionados con la estimación translingüe de similitudes entre oraciones con caracterizaciones independientes de la lengua. Los resultados preliminares son prometedores y muestran la factibilidad de extraer oraciones paralelas desde la enciclopedia con una calidad aceptable.

Actualmente nuestros esfuerzos están enfocados en la aplicación de medidas de similitud de carácter semántico, léxico y estructural que permitan identificar con certeza mayor las oraciones paralelas. Planteamos realizar una combinación de las diversas medidas basada en una combinación lineal o en un modelo de clasificación. Para estimar los coeficientes para dicha combinación, requeriremos

de un conjunto de pares de artículos anotados manualmente. Para ello hemos desarrollado ya una interfaz gráfica para facilitar la tarea.

La evaluación global de nuestros desarrollos se realizará de manera indirecta. Nuestro objetivo será determinar si los candidatos a pares paralelos mejoran la calidad de la traducción de un sistema estadístico entrenado previamente con otro corpus paralelo. Dicho traductor estadístico será un sistema MOSES (Koehn *et al.*, 2007)[10]. El traductor inicial será entrenado con el corpus *Europarl* y luego será enriquecido con las oraciones extraídas de Wikipedia. Nuestra hipótesis es que la calidad del traductor enriquecido será mayor, en particular cuando tenga que procesar textos de alguno de nuestros dominios de interés.

Agradecimientos

Esta investigación se realiza en el marco del proyecto Tacardi (TIN2012-38523-C02-00).

Referencias

Adafre, S. y de Rijke, M. (2006). Finding Similar Sentences across Multiple Languages in Wikipedia. En *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 62–69.

Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., y Tapias, D., editores (2008). *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Cui, G., Lu, Q., Li, W., y Chen, Y. (2008). Corpus Exploitation from Wikipedia for Ontology Construction. En Calzolari *et al.* (2008), páginas 2126–2128.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. En *Proceedings of the Machine Translation Summit X*, páginas 79–86. AAMT, AAMT.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., y Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. En *ACL*. The Association for Computer Linguistics.

Mcnamee, P. y Mayfield, J. (2004). Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval*, **7**(1-2), 73–97.

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

Rafalovitch, A. y Dale, R. (2009). United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. En *Proceedings of the Machine Translation Summit XII*, páginas 292–299. International Association of Machine Translation.

[10] `http://www.statmt.org/moses/`

Simard, M., Foster, G. F., y Isabelle, P. (1992). Using Cognates to Align Sentences in Bilingual Corpora. En *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. En N. Calzolari, K. Choukri, T. Declerck, M. Dogan, B. Maegaard, J. Mariani, J. Odijk, y S. Piperidis, editores, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Wikipedia (2014a). Categoría:Ciencia. [http://es.wikipedia.org/wiki/Categoría:Ciencia]. Accedido 6/mar/2014.

Wikipedia (2014b). Categoría:Informática. [http://es.wikipedia.org/wiki/Categoría:Informática]. Accedido 6/mar/2014.

Wikipedia (2014c). Wikipedia: Categorización. [http://es.wikipedia.org/wiki/Wikipedia:Categorización]. Accedido 6/mar/2014.

Zesch, T., Müller, C., y Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wikictionary. En Calzolari *et al.* (2008).