Extracción de corpus paralelos de la Wikipedia basada en la obtención de alineamientos bilingües a nivel de frase*

Extracting Parallel Corpora from Wikipedia on the basis of Phrase Level Bilingual Alignment

Joan Albert Silvestre-Cerdà, Mercedes García-Martínez, Alberto Barrón-Cedeño, Jorge Civera y Paolo Rosso
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
mgarcia@iti.upv.es,{jsilvestre,lbarron,jcivera,prosso}@dsic.upv.es

Resumen: Este artículo presenta una nueva técnica de extracción de corpus paralelos de la Wikipedia mediante la aplicación de técnicas de traducción automática estadística. En concreto, se han utilizado los modelos de alineamiento basados en palabras de IBM para obtener alineamientos bilingües a nivel de frase entre pares de documentos. Para su evaluación se ha generado manualmente un conjunto de test formado por pares de documentos inglés-español, obteniéndose resultados prometedores.
Palabras clave: corpus comparables, extracción de oraciones paralelas, traducción automática estadística

Abstract: This paper presents a proposal for extracting parallel corpora from Wikipedia on the basis of statistical machine translation techniques. We have used word-level alignment models from IBM in order to obtain phrase-level bilingual alignments between documents pairs. We have manually annotated a set of test English-Spanish comparable documents in order to evaluate the model. The obtained results are encouraging.
Keywords: comparable corpora, parallel sentences extraction, statistical machine translation

1. Introducción

La extracción automática de corpus paralelos a partir de recursos textuales multilingües es, hoy por hoy, una tarea de especial interés debido al creciente auge de la traducción automática estadística. La web es una fuente inmensa de documentos en múltiples lenguas que tiene muchas posibilidades de explotación. No obstante, encontrar frases paralelas a nivel global en la web es una tarea muy dispersa y extremadamente difícil, aunque no imposible (Uszkoreit et al., 2010).

La Wikipedia es uno de los pocos recursos web que nos provee de forma explícita gran cantidad de textos multilingües comparables, pues sus contenidos se presentan como artículos en múltiples idiomas que describen un mismo concepto. El objetivo es, pues, explotar los contenidos comparables de dichos documentos con la finalidad de extraer frases paralelas que puedan ser utilizadas en el entrenamiento de sistemas de traducción automática.

En este trabajo se propone una aproximación heurística a la extracción de corpus paralelos de la Wikipedia basada en técnicas de Traducción Automática Estadística (TAE).

* Este trabajo se ha llevado a cabo en el marco del VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems, financiado parcialmente por parte de la EC (FEDER/FSE; WIQEI IRSES no. 269180 / FP 7 Marie Curie People), por el MICINN como parte del proyecto Text-Enterprise 2.0 (TIN2009-13391-C04-03) en el Plan I+D+i, y por la beca 192021 del CONACyT. También ha recibido apoyo por parte del EC (FEDER/FSE) y del MEC/MICINN bajo el programa MIPRCV "Consolider Ingenio 2010" (CSD2007-00018) y el proyecto iTrans2 (TIN2009-14511), por el MITyC en el marco del proyecto erudito.com (TSI-020110-2009-439), por la Generalitat Valenciana con las ayudas Prometeo/2009/014 y GV/2010/067, y por el "Vicerrectorado de Investigación de la UPV" con la ayuda 20091027.

En la siguiente sección analizaremos los trabajos previos que han servido de inspiración a este trabajo. Posteriormente, en la Sección 3 se describe ampliamente el sistema propuesto. La Sección 4 muestra los resultados experimentales y finalmente, una serie de conclusiones son expuestas en la Sección 5.

2. Trabajos relacionados

Debido a su creciente necesidad e importancia, la extracción automática de corpus paralelos es una tarea bastante explorada en la actualidad, aunque los primeros trabajos se realizaron hace ya más de dos décadas (Brown, Lai, y Mercer, 1991; Gale y Church, 1991), si bien éstos se ceñían a encontrar alineamientos entre frases en textos paralelos. Estos trabajos proponen métodos de alineamiento muy rápidos pero poco precisos, pues para detectar relaciones entre frases utilizaban únicamente la información de longitud de las oraciones. Posteriormente, Chen propuso utilizar información léxica mediante un sencillo modelo de traducción estadístico basado en palabras, demostrando una mejora significativa de la calidad de los alineamientos extraídos (Chen, 1993), y unos años más tarde, Moore combinó ambas aproximaciones (Moore, 2002). Más recientemente, González propuso un modelo de alineamiento entre frases y palabras inspirado en el modelo 1 de IBM (González-Rubio et al., 2008).

Con el problema de alinear frases en textos paralelos bien estudiado, y ante la creciente demanda de corpus paralelos para TAE, los principales esfuerzos se centraron en la extracción de corpus paralelos (Eisele y Xu, 2010; Uszkoreit et al., 2010; Varga et al., 2005), en incluso monolingües (Barzilay y Elhadad, 2003; Quirk, Brockett, y Dolan, 2004), a partir de la web. En éste ámbito, la Wikipedia ha sido un recurso bastante explotado, presentándose una gran variedad de aproximaciones, desde métodos heurísticos (Adafre y de Rijke, 2006; Mohammadi y GhasemAghaee, 2010) hasta aproximaciones basadas en clasificación estadística utilizando combinaciones lineales de características (Smith, Quirk, y Toutanova, 2010; Tomás et al., 2008). También se han llevado a cabo algunos trabajos en la vertiente monolingüe (Yasuda y Sumita, 2008). Ahora bien, ninguno de los trabajos previos ha explorado la utilización de modelos de traducción estadísticos como sistemas de evaluación de alineamientos en recursos comparables como la Wikipedia, y es precisamente este vacío experimental el que se pretende cubrir en este trabajo.

3. Descripción del sistema

Para la tarea de extracción de corpus paralelos de la Wikipedia consideraremos pares de documentos de Wikipedia $X = (x_1, \ldots, x_j, \ldots, x_{|X|}) \in \mathcal{X}^*$ e $Y = (y_1, \ldots, y_i, \ldots, y_{|Y|}) \in \mathcal{Y}^*$ que representen un mismo concepto, siendo x_j la j-ésima frase del documento X, y_i la i-ésima frase del documento Y, y $\mathcal{X}$ e $\mathcal{Y}$ los vocabularios de los lenguajes en los que se encuentran los respectivos documentos. Definimos (x_j, y_i) como un alineamiento entre la j-ésima frase del documento X y la i-ésima frase del documento Y, y A un conjunto finito de alineamientos.

Inicialmente asumiremos que $A = (X \times Y)$, es decir, el conjunto A contiene todo alineamiento posible entre las frases de X y de Y. La probabilidad de cada alineamiento $(x_j, y_i) \in A$ se calcula de acuerdo con el modelo 4 de IBM (Brown y others, 1993), que es un modelo de alineamiento a nivel de palabra ampliamente utilizado en Traducción Automática Estadística. Un alineamiento recibirá una probabilidad alta si el grado de coocurrencia de las palabras que componen las frases es alto, pero por contra recibirá una probabilidad baja si las palabras involucradas tienen poca o ninguna correlación. Cabe decir que las puntuaciones otorgadas por los modelos de IBM provienen de una serie de productos de probabilidades, tantos como el número de palabras que conforman la frase de destino y_i, por lo que dicha puntuación debe ser normalizada convenientemente para que no sea dependiente de la longitud. De no ser así, los alineamientos con frases destino y_i de menor número de palabras tenderían a ser más probables, pudiendo darse casos de alineamientos (x_j, y_i) con altos valores de probabilidad con $|x_j| = 8$ e $|y_i| = 1$, por ejemplo.

Una vez se han evaluado todos los alineamientos del conjunto A, se obtiene el conjunto de alineamientos más probables $B \subseteq A$ mediante la siguiente maximización:

$$(x_j, y_i) \in B \,/\, p_{IBM}(x_j \mid y_i) > p_{IBM}(x_j \mid y_{i'}) \quad (1)$$
$$\forall i' = 1 \ldots |Y| \quad \forall j = 1 \ldots |X|$$

Es decir, para cada frase x_j del documento X, conservaremos el alineamiento (x_j, y_i)

que maximice la probabilidad del modelo 4 de IBM para toda posible frase y_i. Esto implica añadir una restricción importante en el proceso de alineamiento, pero que no obstante nos permite definir un sistema base o inicial que tenemos previsto mejorar en el futuro mediante el cálculo y la posterior combinación de los alineamientos en ambas direcciones.

Por último, se genera el conjunto final de alineamientos filtrados $C \subseteq B$, formado por aquellos alineamientos cuya puntuación supere un cierto umbral α, es decir:

$$(x_j, y_i) \in C \,/\, p_{IBM}(x_j \mid y_i) > \alpha \qquad (2)$$

El umbral α puede interpretarse como un parámetro que afecta a la calidad de los alineamientos extraídos, ya que cuanto mayor es el umbral, mayor es nuestra exigencia sobre el sistema, extrayéndose en consecuencia un menor número de alineamientos. En la Sección 4 estudiaremos la influencia de este parámetro en las prestaciones de nuestro sistema.

4. Experimentación

Con el objetivo de evaluar las prestaciones que ofrece nuestro método de extracción de corpus paralelos de la Wikipedia, hemos realizado un estudio experimental en el que se evalúa la calidad de los pares de frases extraídos automáticamente por nuestro sistema a partir de un conjunto de prueba que tuvimos que generar de forma manual, debido a la inexistencia de corpus adecuadamente etiquetados para esta tarea. La generación de dicho conjunto, formado por pares de documentos de la Wikipedia en inglés y español, es detallada en las Secciones 4.1 y 4.2.

El modelo 4 de IBM fue entrenado con MGIZA, un software basado en el popular GIZA++ que nos ofrece la posibilidad de evaluar un conjunto de prueba con los modelos ya entrenados, además de que permite realizar un entrenamiento paralelo de los mismos. Con el fin de minimizar los problemas relacionados con las palabras fuera de vocabulario y generalizar el dominio del sistema, los modelos de IBM se entrenaron con un subconjunto de pares de frases, definido en (Sanchis-Trilles et al., 2010), de tres corpus de referencia en el área de la Traducción Automática Estadística: Europarl-v5 (Koehn, 2005), News-Commentary y United Nations (Rafalovitch y Dale, 2009). Las estadísticas de este subconjunto pueden ser consultadas en la Tabla 1. Cabe destacar la gran cantidad de pares de frases empleados para el entrenamiento de los modelos, así como el considerable tamaño de los vocabularios de cada una de las lenguas.

Tabla 1: Estadísticas básicas del corpus empleado para el entrenamiento de los modelos IBM.

	Entrenamiento	
Idioma	En	Es
Número de frases	2.8M	
Tamaño Vocabulario	118K	164K
Número Total Palabras	54M	58M

El resto de esta sección se estructura como sigue: la Sección 4.1 muestra el procedimiento de extracción de documentos y su preproceso. Posteriormente, las Secciones 4.2 y 4.3 presentan la metodología de etiquetado y las métricas de evaluación empleadas, respectivamente. Finalmente, la Sección 4.4 expone los resultados obtenidos al evaluar el conjunto de entrenamiento generado manualmente.

4.1. Selección de documentos y preproceso

La Wikipedia alberga miles de artículos disponibles en inglés y español, y abarcan un dominio extremadamente amplio. Por ese motivo, y con el objetivo de realizar una prueba optimista con el sistema, se realizó una selección de pares de documentos cuyos dominios se asemejaran al dominio del corpus empleado en el entrenamiento del modelo de alineamiento. En concreto, se seleccionaron un total de 15 pares de documentos inglés-español relacionados con la economía y procesos administrativos de la Unión Europea. De dichos documentos se extrajo el texto plano, que posteriormente fue sometido a un preproceso consistente en la separación de frases en líneas (sentence-splitting), aislamiento de palabras y signos de puntuación (tokenizing) y conversión a minúsculas (lowercasing). Las estadísticas de dicho corpus después de ser sometido a este preproceso se muestran en la Tabla 2.

4.2. Metodología de etiquetado

A continuación se describe la metodología seguida para generar el conjunto de evalua-

Tabla 2: Estadísticas básicas del conjunto de evaluación construido de forma manual.

	Evaluación	
Idioma	En	Es
Número de documentos	15	
Número de frases	661	341
Alineamientos posibles	22680	
Tamaño Vocabulario	3,4K	2,8K
Número Total Palabras	24,5K	16,2K

ción etiquetado, partiendo de un conjunto de pares de documentos previamente preprocesados. Esta metodología está inspirada en (Och y Ney, 2003), pero tomando alineamientos entre frases en lugar de alineamientos entre palabras.

Dos personas se encargaron de etiquetar manualmente y independientemente todo el conjunto de pares de documentos. Se les pidió que anotaran aquellos alineamientos, de entre todos los posibles para cada par de documentos, que guardaran una relación de paralelismo.

Adicionalmente, los etiquetadores fueron instruidos para que asignaran cada uno de los alineamientos a uno de los siguientes dos conjuntos:

- P: Conjunto de alineamientos probables. Definen alineamientos entre frases que conforman traducciones similares, aunque no exactas, en las que se expresa la misma idea semántica, o bien para indicar que un determinado alineamiento forma parte de una relación 1-a-muchos o muchos-a-1.
- S: Conjunto de alineamientos seguros, siendo $S \subseteq P$. Define alineamientos entre frases que son traducciones exactas o casi exactas (paralelas).

En este contexto, el etiquetador 1 genera los conjuntos S_1 y P_1, mientras que el etiquetador 2 genera S_2 y P_2. Entonces, los conjuntos S_1, P_1, e S_2, P_2 se combinan en S y P de la siguiente forma:

$$S = S_1 \cap S_2$$
$$P = P_1 \cup P_2$$

El conjunto P (que incluye S) representa los pares de frases que deberían ser extraídos por el sistema, y por tanto son tomados como referencia para la tarea. Para el caso concreto de este corpus, el conjunto S está formado por 10 alineamientos, mientras que el conjunto P engloba un total de 115 alineamientos.

4.3. Medidas de Evaluación

La evaluación de la calidad del conjunto filtrado de alineamientos C obtenido de forma automática mediante nuestro sistema se ha realizado mediante la métrica Sentence Alignment Error Rate, claramente inspirada en la presentada en (Och y Ney, 2003).

Dado un par de documentos X e Y, los conjuntos de alineamientos entre ambos documentos S y P etiquetados manualmente, y el conjunto filtrado de alineamientos C, se define la métrica Sentence Alignment Error Rate (SAER) como sigue:

$$SAER(S, P, C) = 1 - \frac{|C \cap S| + |C \cap P|}{|C| + |S|} \quad (3)$$

Al igual que (Och y Ney, 2003), también hemos empleado las medidas de cobertura y precisión para obtener más información acerca de las prestaciones del sistema:

$$\text{Cobertura} = \frac{|C \cap S|}{|S|}, \ \text{Precisión} = \frac{|C \cap P|}{|C|} \quad (4)$$

4.4. Resultados

En la presente sección se presentan los resultados de las pruebas experimentales llevadas a cabo con nuestro sistema, utilizando el conjunto de evaluación generado de forma manual. En la Sección 3 hemos resaltado la necesidad de estudiar la influencia del parámetro α, puesto que radica directamente en la calidad de la frases extraídas. Un valor alto para dicho umbral puede conllevar a que el sistema no sea capaz de extraer ningún alineamiento. Por contra, un valor pequeño de α se traduciría en la extracción de un gran número de pares de frases, e idealmente en un aumento del número de alineamientos correctos (Verdaderos Positivos, VP), aunque hay que tener en cuenta que el número de casos de Falsos Positivos (FP), es decir, alineamientos que no existen en la referencia, aumenta generalmente en mayor proporción que los VPs. La clave está pues en encontrar un valor de α que garantice la obtención de la mayor proporción posible de Verdaderos Positivos (VPR) y que minimice el ratio de Falsos Positivos (FPR). Ambas proporciones se calculan de la siguiente forma:

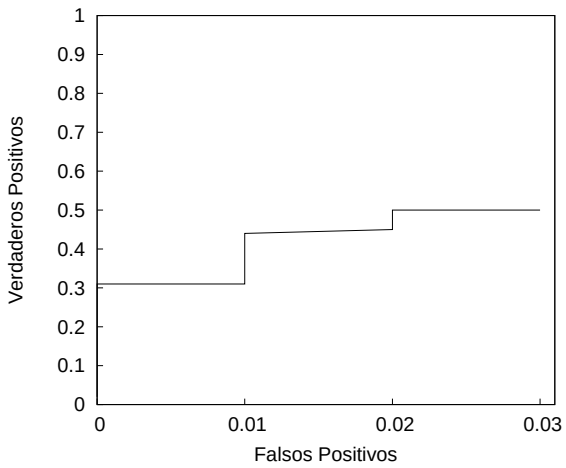

Figura 1: Curva ROC para constatar la relación entre Verdaderos Positivos y Falsos positivos en función del parámetro α.

$$VPR = \frac{VP}{P} = \frac{VP}{VP + FN} \quad (5)$$

$$FPR = \frac{FP}{N} = \frac{FP}{FP + VN} \quad (6)$$

donde P representa el número de muestras positivas, que es igual al número de casos de Verdaderos Positivos (VP) más el número de casos de Falsos Negativos (FN), mientras que N representa el número de muestras negativas, que es igual al número de casos de Falsos Positivos (FP) más el número de casos de Verdaderos Negativos (VN).

Con esta finalidad, hemos realizado una exploración exhaustiva del parámetro α, y posteriormente hemos dibujado una curva ROC, mostrada en la Figura 1, en la que se observa la relación entre los Verdaderos Positivos (VPR, eje vertical) y los Falsos Positivos (FPR, eje horizontal) en función del umbral α, cuyo valor es inversamente proporcional al desplazamiento de ambos ejes. Cabe decir que dicha exploración debería de haberse llevado a cabo mediante un conjunto de desarrollo, pero debido a la ausencia del mismo tuvimos que emplear el conjunto de evaluación. En el futuro planeamos ampliar dicho corpus para poder generar un conjunto de desarrollo.

De la Figura 1 cabe destacar varias cosas. En primer lugar, la gráfica tiene un aspecto degenerado debido a que la proporción relativa de Falsos Positivos nunca podrá llegar a valer 1, puesto que está acotada superiormente por $FP/(FP + VN)$ teniendo en cuenta que $FP \leq |X|$ (como máximo se darán lugar tantos FPs como número de frases del documento de entrada) y que $VN \leq |X \times Y|$ (el sistema puede llegar a descartar el conjunto de todos los posibles alineamientos), por lo que el valor del cociente será muy pequeño. En segundo lugar, podemos observar que para valores más altos del umbral α la relación de Falsos Positivos llega a ser casi cero para un ratio del 0.3 de Verdaderos Positivos, mientras que para valores de α más pequeños podemos llegar a conseguir un 0.5 de VPR con un ratio del 0.02 de FPR. En términos relativos, este segundo punto parece ser el óptimo, pero si tomamos en cuenta los valores absolutos, nos encontramos con diferencias del orden de centenares de FPs. Es por este motivo por el cual nos decantaremos por el primer de ellos, con $\alpha = 1{,}1 \cdot 10^{-3}$.

En la Tabla 3 se muestran los valores de las métricas, presentadas en la Sección 4.3, tras la evaluación del conjunto de prueba, además de otras estadísticas de interés, para el valor del umbral que hemos considerado como óptimo ($\alpha = 1{,}1 \cdot 10^{-3}$) y para dos casos extremos, con el objetivo de apreciar más notoriamente la influencia de dicho parámetro en las prestaciones del sistema. La primera fila muestra el tamaño del conjunto de alineamientos filtrados C, mientras que las cuatro filas siguientes muestran el número de muestras clasificadas como Verdaderos Positivos (VP), Verdaderos Negativos (VN), Falsos Positivos (FP) y Falsos Negativos (FN). Por último, se muestran los valores de las tres métricas empleadas para evaluar las prestaciones del sistema: cobertura, precisión y SAER.

En ella se puede ver como, a pesar de la simplicidad de nuestro planteamiento, se obtienen unos resultados bastante aceptables para el valor óptimo de α, con una tasa del 0.36 de error de alineamiento, un 0.59 de grado de precisión, y sobretodo un 0.9 de cobertura, aunque cabe decir que esta última no es una medida fiable dado que en el corpus sólo existen 10 alineamientos etiquetados como seguros. A continuación se muestran algunos ejemplos de los pares de frases extraídos por nuestro sistema:

En: On 20 april 2005, the European Commission adopted the communication on Kosovo to the council “a european futu-

Tabla 3: Resultados del sistema para el conjunto de test generado manualmente, con $\alpha = \{1 \cdot 10^{-4}, 1{,}1 \cdot 10^{-3}, 5 \cdot 10^{-2}\}$.

	$\alpha = 1 \cdot 10^{-4}$	$\alpha = 1{,}1 \cdot 10^{-3}$	$\alpha = 5 \cdot 10^{-2}$
$\|C\|$	656	59	4
VP	58	35	2
VN	21967	22541	22563
FP	598	24	2
FN	57	80	113
Cobertura	1,00	0,90	0,1
Precisión	0,09	0,59	0,50
SAER	0,90	0,36	0,79

re for Kosovo" which reinforces the commission's commitment to Kosovo.

Es: El 20 de abril de 2005, la Comisión Europea adoptó la comunicación sobre kosovo en el consejo "un futuro europeo para Kosovo" que refuerza el compromiso de la comisión con Kosovo.

En: He added that the decisive factor would be the future and the size of the eurozone, especially whether Denmark, Sweden and the UK would have adopted the euro or not.

Es: Añadió que el factor decisivo será el futuro y el tamaño de la zona del euro, especialmente si Dinamarca, Suecia y el Reino Unido se unen al euro o no.

En: Montenegro officially applied to join the EU on 15 december 2008.

Es: Oficialmente, Montenegro pidió el acceso a la UE el 15 de diciembre de 2008.

Si observamos nuevamente la Tabla 3 y nos fijamos en las diferencias existentes entre el caso óptimo y los casos extremos, se pueden extraer algunas conclusiones interesantes. Para $\alpha = 1 \cdot 10^{-4}$ no se filtra ningún alineamiento, esto es, $C = B$, y por tanto nos damos cuenta que nuestro sistema nunca será capaz de encontrar 57 alineamientos que sí están en la referencia. Para evitar esta severa limitación tenemos previsto obtener los alineamientos entre frases en ambos sentidos (X a Y, e Y a X), y posteriormente aplicar un algoritmo heurístico inspirado en (Och y Ney, 2003) que los combine, partiendo de la intersección entre ambos alineamientos y añadiendo alineamientos adicionales. Esto nos llevará, en primer lugar, a obtener alineamientos más robustos, y en segundo lugar, a capturar relaciones entre frases de muchas-a-1, 1-a-muchas, e incluso muchas-a-muchas.

5. Conclusiones y Trabajo Futuro

En este trabajo hemos presentado una aproximación heurística alternativa a las ya existentes para la extracción automática de corpus paralelos a partir de los contenidos multilingües comparables que ofrece la Wikipedia. La evaluación experimental ha mostrado unos resultados francamente prometedores para nuestro sistema inicial. Como extensión de este trabajo planeamos obtener de forma heurística los alineamientos entre frases en ambas direcciones con el objetivo de mejorar la calidad del sistema, una mejora que creemos que será sustancial. Otra alternativa de cara al futuro sería emplear la variante del modelo 1 de IBM presentada en (González-Rubio et al., 2008) en esta tarea, ya que nos permitiría obtener los alineamientos bidireccionales de forma no heurística mediante un entrenamiento Expectation-Maximization (Dempster, Laird, y Rubin, 1977). Con la implementación de estas mejoras, realizaremos un estudio comparativo de nuestro sistema con otros sistemas del estado del arte.

Cabe destacar, además, que en este trabajo hemos adaptado una metodología existente para la evaluación de alineamientos a nivel de frase. Para ello, hemos definido una metodología de etiquetado adecuada para generar un conjunto de evaluación, así como una serie de métricas para cuantificar las prestaciones del sistema. Como trabajo futuro pre-

tendemos aumentar el tamaño del corpus y el número de anotadores, con el fin de hacer más robusto el proceso de etiquetado manual de los alineamientos.

Bibliografía

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

Barzilay, Regina y Noemie Elhadad. 2003. Sentence Alignment for Monolingual Comparable Corpora. En *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, páginas 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Brown, P. F. y others. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Brown, Peter F., Jennifer C. Lai, y Robert L. Mercer. 1991. Aligning Sentences in Parallel Corpora. En *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, páginas 169–176, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chen, Stanley F. 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information. En *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, ACL '93, páginas 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dempster, A. P., N. M. Laird, y D. B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Roy. Statistical Society. Series B*, 39(1):1–38.

Eisele, Andreas y Jia Xu. 2010. Improving Machine Translation Performance using Comparable Corpora. En *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora LREC 2010*, páginas 35–41. ELRA.

Gale, William A. y Kenneth W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpora. En *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, páginas 177–184, Stroudsburg, PA, USA. Association for Computational Linguistics.

González-Rubio, Jesús, Germán Sanchis-Trilles, Alfons Juan, y Francisco Casacuberta. 2008. A Novel Alignment Model Inspired on IBM Model 1. En *Proceedings of the 12th conference of the European Association for Machine Translation*, páginas 47–56.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. En *Proc. of the MT Summit X*, páginas 79–86, September.

Mohammadi, M. y N. GhasemAghaee. 2010. Building Bilingual Parallel Corpora Based on Wikipedia. En *Computer Engineering and Applications (ICCEA), 2010 Second International Conference on*, volumen 2, páginas 264 –268, march.

Moore, Robert C. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. En *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, AMTA '02, páginas 135–144, London, UK, UK. Springer-Verlag.

Och, Franz Josef y Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51, March.

Quirk, Chris, Chris Brockett, y William Dolan. 2004. Monolingual Machine Translation for Paraphrase Generation. En *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, páginas 142–149.

Rafalovitch, Alexandre y Robert Dale. 2009. United Nations General Assembly Resolutions: A Six-Language Parallel Corpus.

Sanchis-Trilles, Germán, Jesús Andrés-Ferrer, Guillem Gascó, Jesús González-Rubio, Pascual Martínez-Gómez, Martha-Alicia Rocha, Joan-Andreu Sánchez, y Francisco Casacuberta. 2010. UPV-PRHLT English–Spanish System for WMT10. En *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, páginas

172–176, Uppsala, Sweden, July. Association for Computational Linguistics.

Smith, Jason R., Chris Quirk, y Kristina Toutanova. 2010. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. En *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, páginas 403–411, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tomás, Jesús, Jordi Bataller, Francisco Casacuberta, y Jaime Lloret. 2008. Mining Wikipedia as a Parallel and Comparable Corpus. *LANGUAGE FORUM*, 34(1). Article presented at CICLing-2008, 9th International Conference on Intelligent Text Processing and Computational Linguistics, February 17 to 23, 2008, Haifa, Israel.

Uszkoreit, Jakob, Jay M. Ponte, Ashok C. Popat, y Moshe Dubiner. 2010. Large Scale Parallel Document Mining for Machine Translation. En *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, páginas 1101–1109, Stroudsburg, PA, USA. Association for Computational Linguistics.

Varga, Dániel, László Németh, Péter Halácsy, András Kornai, Viktor Trón, y Viktor Nagy. 2005. Parallel Corpora for Medium Density Languages. En *Proceedings of the RANLP 2005*, páginas 590–596.

Yasuda, Keiji y Eiichiro Sumita. 2008. Method for Building Sentence-Aligned Corpus from Wikipedia. En *Proceedings of the 33th AAAI workshop on Artificial Intelligence (AAAI-08)*.