Procesamiento del lenguaje Natural, nº 41 (2008), pp. 267-272 recibido 7-05-2008; aceptado 16-06-2008

Generación de múltiples hipótesis ponderadas de reordenamiento para un sistema de traducción automática estadística*

Generating multiple weighted reordering hypotheses for an SMT system

Marta R. Costa-jussà
Universitat Politècnica de Catalunya
Campus Nord 08034 Barcelona
mruiz@gps.tsc.upc.edu

José A. R. Fonollosa
Universitat Politècnica de Catalunya
Campus Nord 08034 Barcelona
adrian@gps.tsc.upc.edu

Resumen: Los errores debidos al cambio de orden de las palabras son uno de los principales retos en los sistemas de traducción automática estadística (TAE). Esta comunicación propone la estrategia estadística de reordenamiento automático estadístico (RAE) para afrontar el reordenamiento. El método propuesto aprovecha la poderosas técnicas de aprendizaje estadístico desarrolladas en traducción estadística para traducir la lengua fuente (*S*) a una lengua fuente reordenada (*S'*), que nos permita mejorar la traducción final a la lengua destino (*T*). Esta técnica permite extraer un grafo de hipótesis ponderadas de reordenamiento que se utiliza como entrada al sistema TAE. Además, el uso de clases de palabras en la estrategia RAE ayuda a generalizar reordenamientos. En este artículo se presentan resultados en la tarea EPPS en la dirección inglés a español y se muestra una mejora de 2.4 puntos BLEU en la calidad de la traducción.
Palabras clave: traducción automática estadística, grafo de reordenamiento, tuplas

Abstract: Reordering is one of the most important challenges in Statistical Machine Translation (SMT) systems. This paper describes a novel strategy to face it: Statistical Machine Reordering (SMR). It consists in using the powerful techniques developed for Statistical Machine Translation (SMT) in order to translate the source language (*S*) into a reordered source language (*S'*), which allows for an improved translation into the target language (*T*). This technique allows to extract a weighted reordering graph which is used as SMT input. In addition, the use of classes in SMR helps to generalize word reorderings. Experiments are reported in the EPPS task in the direction English to Spanish showing a 2.4 point BLEU improvement in translation quality.
Keywords: statistical machine translation, reordering graph, tuples

1. Introducción

La traducción automática estadística (TAE) considera que una oración *s* de una lengua fuente puede ser traducida en cualquier oración *t* de la lengua destino con cierta probabilidad. La traducción consiste precisamente en determinar la oración con mayor probabilidad de constituir una traducción para la oración fuente. Estas probabilidades se aprenden principalmente a partir textos paralelos bilingües.

Los sistemas TAE tienden a utilizar secuencias de palabras, denominadas sintagmas (Koehn, Och, y Marcu, 2003), como unidades básicas del modelo de traducción, con el objetivo de introducir el contexto en dicho modelo. En paralelo, al modelo de sintagmas, también se ha propuesto el uso de *n*-gramas de tuplas bilingües (Casacuberta, Vidal, y Vilar, 2002; Mariño et al., 2006) (o unidades de traducción) como una alternativa para tener en cuenta el contexto con unidades bilingües más pequeñas. Ambos sistemas llevan a cabo la traducción mediante una búsqueda que maximiza una combinación de las probabilidades asignadas a la traducción por el modelo de traducción en sí y otras funciones de traducción (Och y Ney, 2002). La Ecuación 1 muestra la combinación donde h_m son las funciones de traducción y λ_m los pesos que se asigna a cada una de ellas.

* Este trabajo ha sido parcialmente subvencionado por el gobierno español (beca FPU), el proyecto AVIVAVOZ y TECNOPARLA

ISSN: 1135-5948

$$\tilde{t} = \text{argmax}_t \left\{ \sum_{m=1}^{M} \lambda_m h_m(t, s) \right\} \qquad (1)$$

Tanto en los sistemas TAE basados en sintagmas como el los basados en n-gramas, la introducción de reordenamientos es crucial. La técnica más directa consiste en permitir que las palabras traducidas no sigan el orden de la lengua fuente cuando se traduce. De esta manera el decodificador permite reordenamientos según los criterios de diversos modelos estadísticos, pero con el inconveniente de incrementar sensiblemente el coste computacional (Knight, 1999).

Recientemente han apareciendo diversas estrategias de reordenamiento de palabras que intentan modificar el orden de la oración fuente para que se corresponda con el orden de la oración destino (Kanthak et al., 2005; Crego y Mariño, 2007; Zhang, Zens, y Ney, 2007). En el primer caso, se limitan los reordenamientos posibles utilizando diversos criterios como la técnica IBM. En el segundo caso, se extraen reglas de reordenamiento utilizando información morfológica directamente del corpus paralelo y se seleccionan las menos dispersas por frecuencia relativa. En el tercer caso, estas reglas se aprenden utilizando información sintáctica. En todos los casos se genera un grafo de hipótesis de reordenamiento que se utiliza como entrada al sistema TAE. Las hipótesis de reordenamiento no tienen ninguna probabilidad asignada.

La aproximación propuesta en esta comunicación (RAE) para el reordenamiento de las palabras se basa en los mismos principios que la traducción automática estadística (TAE) y comparte el mismo tipo de decodificador. El reordenamiento se trata como una traducción estadística de la lengua fuente (S) a la lengua fuente reordenada (S') y se entrena a partir de la información de alineado. Una vez estimadas las probabilidades de reordenado, el sistema RAE reordena la oración fuente y esta se pasa al sistema TAE (Costa-jussà y Fonollosa, 2006). En esta comunicación especialmente introducimos un nuevo acoplamiento entre el sistema RAE y el TAE. El sistema RAE puede computar una única hipótesis de reordenamiento o bien un grafo ponderado de hipótesis. Este grafo de hipótesis se utiliza como entrada al sistema TAE y la decisión final de reordenamiento se produce juntamente a la decisión de traducción. Para mejorar la capacidad de generalización del sistema propuesto, se usarán clases de palabras en vez de palabras como entrada al sistema RAE.

La comunicación se organiza de la siguiente manera. La Sección 2 describe brevemente el sistema de referencia. La Sección 3 describe con detalle la estrategia de reordenamiento propuesta. La Sección 4 presenta y discute los resultados, y finalmente la Sección 5 concluye.

2. Sistema de referencia TAE basado en n-gramas

El modelo de traducción puede entenderse como un modelo de lenguaje de unidades bilingües (llamadas tuplas). Dichas tuplas, definen una segmentación monótona de los pares de oraciones utilizadas en el entrenamiento del sistema (s_1^J, t_1^I), en K unidades $(u_1, ..., u_K)$.

En la extracción de las unidades bilingües, cada par de oraciones da lugar a una secuencia de tuplas que sólo depende de los alineamientos internos entre las palabras de la oración.

La Figura 1 muestra un ejemplo de extracción de tuplas.

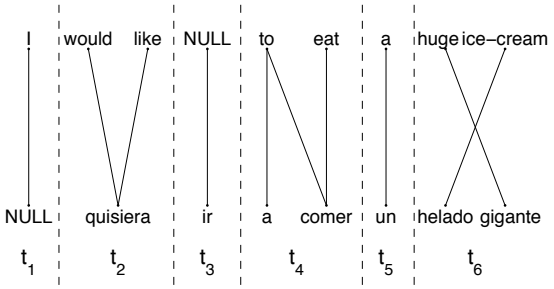

Figura 1: *Extracción de tuplas a partir de un par de oraciones alineadas palabra a palabra.*

En la traducción de una oración de entrada, el decodificador debe encontrar la secuencia de tuplas asociada a una segmentación de la oración de entrada que produzca probabilidad máxima. Tal probabilidad máxima, se calcula como combinación lineal de los modelos utilizados en el sistema de traducción.

El modelo de traducción se ha implementado utilizando un modelo de lenguaje bilingüe basado en n-gramas, (con $N = 4$).

El decodificador utiliza la combinación de las cuatro funciones de traducción (definidas como probabilidades):

- Un **modelo de lenguaje** basado en n-gramas del idioma destino (LM).
- Una **bonificación** basada en el número de palabras de la traducción, usada para compensar la preferencia del decodificador por las traducciones cortas (WB).

- Un **modelo lexicalizado de traducción** calculado utilizando las probabilidades **léxicas** del modelo IBM1, para ambas direcciones (fuente-destino y viceversa).

3. Sistema de reordenamiento estadístico (RAE)

3.1. Concepto

El sistema de reordenamiento automático estadístico (RAE) se basa en utilizar un sistema de TAE para solventar el reto del reordenamiento. Por lo tanto, un sistema RAE se puede ver como un sistema TAE que traduce de una lengua fuente (*S*) a una lengua fuente modificada (*S'*), dado una lengua destino (*T*). Con lo cual la estrategia de reordenamiento se enfoca como una tarea de traducción *SS'* (fuente-a-fuente reordenado). Y en consecuencia, la tarea de traducción en sí cambia de *ST* (fuente-a-destino) a *S'T* (fuente reordenado-a-destino). El sistema RAE utiliza clases estadísticas de palabras en lugar de clases para poder generalizar los reordenamientos que aprende.

3.2. Descripción

La Figura 2 muestra un diagrama de bloques de descripción del sistema RAE. La entrada es una oración fuente (*S*) y la salida el una oración fuente reordenada (*S'*). El sistema RAE se basa en tres bloques: (1) el extractor de clases; (2) el decodificador que requiere un RAE-LM, es decir, un modelo de traducción; y, (3) el bloque que reordena la oración original usando los índices a la salida del decodificador (postprocesado).

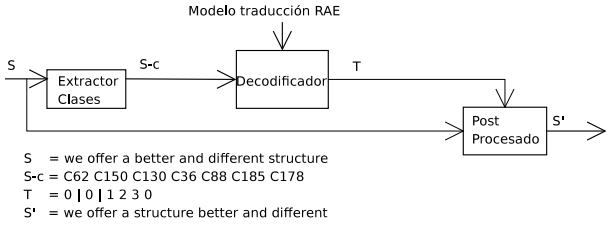

Figura 2: *Diagrama de bloques del sistema RAE*

3.3. Entrenamiento

Para traducir de *S* a *S'* utilizamos un sistema TAE basado en *n*-gramas de tuplas, considerando únicamente el modelo básico de traducción. El entrenamiento de este sistema consta de los siguiente pasos:

1. Obtener clases de palabras de la lengua fuente y de la destino.
2. Alinear a nivel de palabra, en general, obtenemos un alineado de múltiples-a-múltiples palabras.
3. Extraer tuplas de reordenamiento.

   a) Partiendo del alineamiento unión, extraer tuplas bilingües *ST* (es decir, fragmentos fuente y destino) manteniendo la información de alineado. Un ejemplo de tupla bilingüe *ST* es: *only possible compromise # compromiso sólo podría # 0-1 1-1 1-2 2-0*, donde los diferentes campos están separados por # y se corresponden a: (1) fragmento destino; (2) fragmento fuente; y (3) alineado de palabras (aquí, los campos están separados por - y se corresponden a la palabra destino y fuente, respectivamente).

   b) Pasar de un alineado de múltiples-a-múltiples palabras dentro de cada tupla a un alineado de múltiples-a-una palabra. Si una palabra fuente está alineada con dos o más palabras destino, se escoge el vínculo más probable según el modelo IBM 1, y los otros vínculos se omiten (por ejemplo, el número de palabras fuente se mantiene antes y después de la traducción de reordenamiento). En el ejemplo anterior, la tupla cambiará a: *only possible compromise # compromiso sólo podría # 0-1 1-2 2-0*, porque P_{IBM1}(only, sólo) es mayor que P_{IBM1}(possible, sólo).

   c) A partir de las *ST* tuplas bilingües (con el alineado de palabras múltiples-a-una), extraer *SS'* tuplas bilingües (fragmento fuente y su reordenamiento). Siguiendo el ejemplo: *compromiso sólo podría # 1 2 0*, donde el primer campo es el fragmento fuente, y el segundo el reordenamiento de estas palabras fuente.

   d) Eliminar aquellas tuplas cuyo fragmento fuente es la palabra NULL.

   e) Sustituir las palabras de cada fragmento fuente por las clases del paso 1.
4. Calcular el modelo de lenguaje de la secuencia de tuplas bilingües *SS'* compuestas por el fragmento fuente (en clases) y su reordenamiento.

3.4. Uso de la técnica RAE para mejorar el entrenamiento TAE

El corpus fuente original *S* se traduce al corpus fuente reordenado *S′* con el sistema RAE.

El sistema TAE aquí se construye sobre la tarea *S′T* en lugar de sobre la tarea original *ST*.

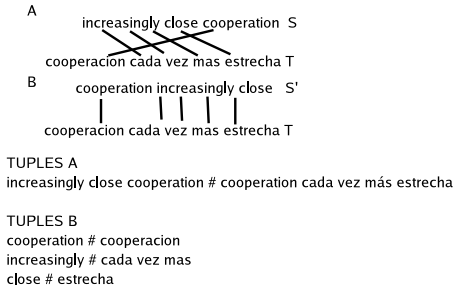

Figura 3: *Extracción de tuplas*

La Figura 3 (B) y (A) muestran respectivamente un ejemplo de extracción de unidades construido con las mismas correspondencias de alineado pero con la oración fuente ordenada diferentemente (este entrenamiento proviene de la salida del sistema RAE). A pesar que la calidad en alineado es la misma, las tuplas que se extraen son diferentes (notar que la extracción de tuplas sigue el criterio de monotonicidad). Se pueden extraer tuplas de longitud menor, con lo cual se reduce la dispersión del vocabulario de traducción.

3.5. Uso de la técnica RAE para generar múltiples hipótesis ponderadas de reordenamiento

El sistema RAE puede generar o bien una única salida (RAE_{1mejor}) o bien un grafo de salidas (RAE_{grafo}) que hemos denominado grafo ponderado de reordenamientos. El diagrama de bloques se muestra en la Figura 4 (A) y (B), respectivamente. El grafo ponderado de reordenamientos permite extender la búsqueda del sistema TAE. Este grafo ponderado de reordenamientos contiene múltiples caminos y cada camino tiene su propio peso. Este peso se incorpora como función adicional en la combinación del sistema TAE.

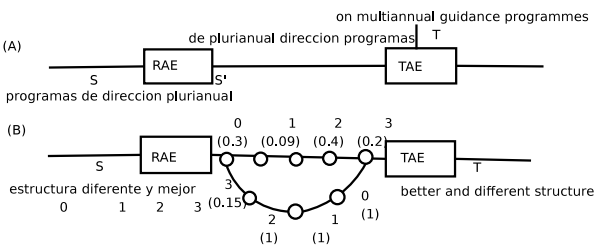

Figura 4: *Concatenación del sistema RAE y TAE: (A) mediante 1mejor (B) mediante un grafo ponderado.*

4. Experimentos

4.1. Corpus

Los experimentos se efectuaron utilizando el corpus proporcionado en la segunda evaluación del proyecto Tc-Star [1] en la tarea inglés a español. El objetivo de este proyecto era construir un sistema de traducción voz-a-voz que pudiera trabajar en tiempo real. Los corpus consisten en la version oficial de los discursos efectuados en las Sesiones Plenarias del Parlamento Europeo (EPPS). El Cuadro 1 muestra las estadísticas básicas de dicho corpus, es decir, número de oraciones, palabras y vocabulario.

		Español	Inglés
Ent.	Oraciones	1,2M	
	Palabras	32M	31M
	Vocabulario	159k	111k
Dev	Oraciones	-	1122
	Palabras	-	26k
Test	Oraciones	-	894
	Palabras	-	26k
	Palabras fv	-	150

Cuadro 1: *Corpus EPPS (Segunda Evaluación del Tc-Star). Ent significa entrenamiento y fv significa fuera de vocabulario.*

4.2. Descripción del sistema

Construimos el sistema TAE basado en *n*-gramas que se ha explicado brevemente en la Sección 2.

Se segmentaron los corpus con herramientas estándard, los detalles se pueden encontrar en (Costa-jussà y R. Fonollosa, 2007) La herramienta GIZA++ (Och y Ney, 2003) se utilizó para alinear en las direcciones fuente-destino y destino-fuente y el alineado final se calculó a partir de la unión de ambos. Las clases de palabras se entrenaron usando la herramienta *mkcls* [2]. Los modelos de lenguaje se entrenaron con el SRILM (Stolcke, 2002). El sistema RAE utiliza un modelo 5-grama y se utilizan 200 clases estadísticas entrenadas sobre el corpus de entrenamiento. El sistema TAE utiliza un modelo de traducción 4-grama y un modelo de lenguaje destino 3-grama, con suavizado Kneser-Ney. Por último, la herramienta MARIE [3] se utilizó como decodificador.

[1] www.tc-star.org

[2] http://www.fjoch.com/mkcls.html

[3] http://gps-tsc.upc.es/veu/soft/soft/marie/

En todos los experimentos, el algoritmo Simplex (Nelder y Mead, 1965) se usó para optimizar los pesos de la combinación de funciones de traducción, con la medida BLEU (Papineni et al., 2002) como función objetivo.

4.3. Resultados

En este apartado se exponen los experimentos realizados y los resultados obtenidos al evaluar la influencia de la estrategia de reordenamiento propuesta (RAE) en la calidad de la traducción.

Hemos estudiado la influencia del reordenamiento RAE propuesto en el sistema de traducción basado en n-gramas que incluye además las cuatro funciones de traducción descritas en las Sección 2: el modelo de lenguaje, la bonificación de palabras, y los modelos de lexicón IBM1 en ambos sentidos. Cuando se introduce el sistema RAE en la configuración $\text{RAE}_{\text{grafo}}$, además se añade como función de traducción la ponderación de las hipótesis de reordenamiento. Todas las funciones de traducción se optimizan conjuntamente.

El Cuadro 2 muestra los resultados en el corpus de test. Podemos observar que la técnica RAE en su versión $1mejor$ obtiene una mejora de 0.7 puntos BLEU. En la versión $grafo$ se obtiene una mejora adicional de 1.7 puntos BLEU dando lugar a una mejora total de 2.4 puntos BLEU respecto a la configuración de referencia.

Sistema	$BLEU_{\text{test}}$
NB	49.12
$\text{RAE}_{\text{1mejor}}$ + NB	49.83
$\text{RAE}_{\text{grafo}}$ + NB	**51.53**

Cuadro 2: *Resultados de traducción. NB es el sistema de referencia.*

4.4. Discusión

La mejora en la calidad de traducción se observa en los resultados BLEU y adicionalmente la Figura 5 muestra mediante unos ejemplos de traducción cómo cambia la traducción final. La mejora en la calidad del sistema puede provenir principalmente por dos motivos. En primer lugar, se monotoniza la tarea y como es sabido las tareas de traducción más monótonas ofrecen una traducción más fiable puesto que los alineaminetos a nivel de palabra son más fáciles de aprender y las unidades de traducción tienden a ser más cortas y menos dispersas (ver Figura 3).

En segundo lugar, las hipótesis de reordenamiento se han aprendido con las potentes técnicas de traducción utilizando clases de palabras que permiten generalizar. Cuando utilizamos la configuración $\text{RAE}_{\text{1mejor}}$ el reordenamiento ofrece un único reorden posible y el sistema TAE no interviene en la decisión final de reordenamiento. Por el contrario, cuando utilizamos la configuración $\text{RAE}_{\text{grafo}}$ el reordenamiento ofrece varias alternativas de reordenamiento y el sistema TAE puede intervenir en la decisión final de reordenamiento. Los pesos de las funciones de traducción del sistema TAE (especialmente el modelo de lenguaje destino) combinadas con el peso de la función de reordenamiento que contiene el grafo se optimizan utilizando el BLEU.

5. Conclusiones

En esta comunicación hemos propuesto una solución para el reto del reordenamiento de palabras en un sistema de traducción automática estadística (TAE). La técnica propuesta ha sido descrita y probada en un sistema de TAE actual basado en n-gramas de tuplas, y se puede aplicar de manera similar a un sistema de TAE basado en sintagmas.

El sistema de reordenamiento automático estadístico (RAE) se aplica previamente al sistema de traducción estadística. Ambos sistemas, RAE y TAE se basan en los mismos principios y comparten el mismo tipo de decodificador. Cuando se extraen las unidades bilingües de traducción, el cambio de orden que se realiza en la oración fuente permite mejorar su modelado, dado que las unidades de traducción son ahora más cortas.

Por otro lado, el hecho de aprender el reordenamiento como un preproceso y de manera independiente al sistema propiamente dicho de traducción permite obtener un sistema final y una traducción eficientes. Además, la estrategia propuesta permite utilizar clases de palabras en el reordenamiento para inferir reordenamientos no vistos durante el entrenamiento del sistema. El hecho de modelar el reordenamiento como una traducción de una lengua fuente a una lengua fuente monotonizada permite que podamos extraer un único reordenamiento (configuración $1mejor$) o bien un grafo ponderado de reordernamientos (configuración $grafo$). En el primer caso, se propone un reordenamiento determinista en el cual el sistema TAE no interviene. En el segundo caso, el hecho de crear un grafo de hipótesis ponderadas de reordenamiento ofrece más robustez a la técnica RAE puesto que el reordenamiento se decide al mismo tiempo que se traduce. Los resultados muestran una mejora total de 2.4 puntos BLEU en la tarea EPPS inglés a español.

FUENTE:to remove the fascist or military dictatorships
TRAD SIST REF: a eliminar los fascistas o dictaduras militares
TRAD RAE_{1mejor}: para eliminar las **dictaduras fascistas o militares**
TRAD RAE_{grafo}: para eliminar las **dictaduras fascistas o militares**
REFERENCIA 1:con el fin de acabar con las dictaduras militares
REFERENCIA 2:para derrumbar las dictaduras fascistas o militares
FUENTE:and the totalitarian dictatorships which then ruled much of eastern and central Europe
TRAD SIST REF: y el totalitario dictaduras que luego dictaminó mucho de oriental y del centro de Europa
TRAD RAE_{1mejor}: y el totalitario dictaduras que entonces gobernaba mucho de la **Europa central y oriental**
TRAD RAE_{grafo}: y las **dictaduras totalitarias** que entonces gobernaba mucho de la **Europa central y oriental**
REFERENCIA 1: así como con las dictaduras totalitarias que controlaban en aquel momento gran parte de la Europa Central y del Este
REFERENCIA 2: y las dictaduras totali tarias que gobernaban en gran parte de Europa central y del este
FUENTE:exploit the fear factor in this matter
TRAD SIST REF: explotar el miedo factor en este asunto
TRAD RAE_{1mejor}: explotar el miedo factor en este asunto
TRAD RAE_{grafo}: explotar el **factor miedo** en este asunto
REFERENCIA 1: explotar el factor miedo en este asunto
REFERENCIA 2:abusar del factor miedo en este asunto

Figura 5: Ejemplos de traducciones.

Bibliografía

Casacuberta, F., E. Vidal, y J.M. Vilar. 2002. Architectures for speech-to-speech translation using finite-state models. *Proc. of the Workshop on Speech-to-Speech Translation: Algorithms and Systems*, páginas 39–44, July.

Costa-jussà, M.R. y J.A.R. Fonollosa. 2006. Statistical machine reordering. En *EMNLP*, páginas 71–77, Sydney, July. ACL.

Costa-jussà, M.R. y J. A. R. Fonollosa. 2007. Analysis of statistical and morphological classes to generate weighted reordering hypotheses on a statistical machine translation system. En *Proceedings of the Second Workshop on Statistical Machine Translation*, páginas 171–176, Prague, June. ACL.

Crego, J.M. y J.B. Mariño. 2007. Improving smt by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

Kanthak, S., D. Vilar, E. Matusov, R. Zens, y H. Ney. 2005. Novel reordering approaches in phrase-based statistical machine translation. En *Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, páginas 167–174, Ann Arbor, MI, June.

Knight, K. 1999. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4), December.

Koehn, P., F.J. Och, y D. Marcu. 2003. Statistical phrase-based translation. En *Proc. of the Human Language Technology Conference, HLT-NAACL'2003*, páginas 48–54, Edmonton, Canada, May.

Mariño, J.B., R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, y M.R. Costa-jussà. 2006. N-gram based machine translation. *Computational Linguistics*, 32(4):527–549, December.

Nelder, J.A. y R. Mead. 1965. A simplex method for function minimization. *The Computer Journal*, 7:308–313.

Och, F.J. y H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. En *Annual Meeting of the Association for Computational Linguistics*, páginas 295–302, Philadelphia, USA, July.

Och, F.J. y H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

Papineni, K., S. Roukos, T. Ward, y W-J. Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. En *40th Annual Meeting of the Association for Computational Linguistics*, páginas 311–318, Philadelphia, PA, July.

Stolcke, A. 2002. Srilm - an extensible language modeling toolkit. En *Proc. of the 9th Int. Conf. on Spoken Language Processing, ICSLP'02*, páginas 901–904, Denver, USA, September.

Zhang, Y., R. Zens, y H. Ney. 2007. Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation. En *Proc. of the Workshop on Syntax and Structure in Statistical Translation (SSST)*, páginas 1–8, Rochester, April.