Evaluación de estrategias para la traducción automática estadística de chino a castellano con el inglés como lengua pivote*

Evaluating Indirect Strategies for Chinese-Spanish statistical machine translation with English as Pivot language

Marta R. Costa-jussà*, Carlos Henríquez† and Rafael E. Banchs‡
*Barcelona Media Innovation Center
Av Diagonal, 177, 9th floor, 08018 Barcelona, Spain
marta.ruiz@barcelonamedia.org
† Universitat Politècnica de Catalunya-TALP
C/Jordi Girona, 08034, Barcelona
carlos.henriquez@upc.edu
‡ Institute for Infocomm Research
1 Fusionopolis Way 21-01, Singapore 138632
rembanchs@i2r.a-star.edu.sg

Resumen: El chino y el castellano son los idiomas más hablados en el mundo como lenguas maternas. Sin embargo, no existe mucha actividad de investigación en traducción automática entre este par de lenguas. Este artículo se enfoca en la investigación del estado actual de la cuestión de la traducción automática estadística entre chino-castellano, ya que hoy en día constituye una de las aproximaciones más usadas dentro del área de la traducción automática. Con este propósito en mente, describimos los corpus paralelos disponibles como el BTEC (Basic Traveller Expressions Corpora), la Biblia y las Naciones Unidas (UN). Concretamente, experimentamos con diferentes estrategias de traducción automática estadística directa e indirectas (denominadas pivotes). Entre las estrategias pivotes exploramos dos metodologías: la traducción de chino a pivote y de pivote a castellano; y el sistema entrenado con un pseudo-corpus chino-castellano, en el que el castellano se ha traducido previamente del pivote. Usamos el inglés como lengua pivote. Los resultados experimentales sugieren que el inglés podría constituir una lengua óptima para la intermediación de la traducción entre chino y castellano. Así pues, uno de los principales objetivos de este trabajo es motivar a la comunidad científica para investigar en este par de lenguas de alto impacto demográfico.
Palabras clave: Chino, castellano, traducción automática estadística, aproximaciones pivote

Abstract: Chinese and Spanish are the most spoken languages in the world. However, there is not much research done in machine translation for this language pair. This paper focuses on investigating the state-of-the-art of Chinese-Spanish Statistical Machine Translation, which nowadays is one of the more popular approaches in Machine Translation. For this purposes we report the details of the available parallel corpus which are the BTEC (Basic Traveller Expressions Corpora), *Holy Bible* and UN (United Nations). Additionally, we experiment with the biggest corpus (UN) to explore alternatives of SMT strategies which consist on using a pivot language. Two alternatives are shown for pivoting: translating from Chinese to Pivot and from Pivot to Spanish; and training on a Chinese-Spanish corpus, where the Spanish corpus has been previously translated from the Pivot language. We use English as Pivot language. Results show that English is quite a nice pivot language between Chinese and Spanish. One of the main objectives of this work is motivating and involving the research community to work in this important pair of languages given the demographic impact of these two languages.
Keywords: Chinese, Spanish, Statistical Machine Translation, Pivot strategies

1. Introducción

El chino y el castellano son dos lenguas muy distintas pero tienen en común que son dos de las lenguas más habladas en el mundo[1]. No hace falta decir la cantidad de intereses econónomicos que mueven estas lenguas. Sin embargo, es curioso que los recursos bilingües para este par son muy pocos.

Recientemente, nos hemos interesado en buscar y recolectar corpus bilingües para el par chino-castellano para investigación de técnicas de traducción automática. Y es sorprendente la baja cantidad de recursos que están disponibles para este par específico. De manera similar, hemos encontrado pocos trabajos de investigación relacionados con este par de idiomas y se pueden reducir a pocas referencias (Banchs et al., 2006a), (Bertoldi et al., 2008), (Banchs y Li, 2008), (Wang et al., 2008), (Banchs et al., 2006b).

Además del corpus del BTEC (Basic Traveller Expressions) disponible a través de la evaluación internacional del IWSLT (International Workshop on Spoken Language Translation) y de la Biblia que se describen en (Bertoldi et al., 2008) and (Banchs y Li, 2008), respectivamente, en este trabajo experimentamos con el corpus de las Naciones Unidas (UN) que contiene chino y castellano (Rafalovith y Dale, 2009) además del inglés, francés, árabe y ruso. El BTEC, la Biblia y UN son corpus paralelos a nivel de oración, que es un requerimento de entrenamiento de los sistemas de traducción automática basados en segmentos (de aquí en adelante denominados sistemas de segmentos) (Koehn, Och, y Marcu, 2003).

Tomando el relativamente reciente corpus paralelo de naciones unidas como punto de partida, este trabajo se focaliza en el problema de desarrollar un sistema de segmentos para el chino-castellano a partir de recursos limitados. Pese a que el corpus de Naciones Unidas es mayor que el corpus del IWSLT y de la Biblia, es aún un corpus menor comparado con las grandes cantidades de datos con los que se entrenan hoy en día los sistemas de segmentos. Motivados por esta falta de datos, y conscientes de la enorme cantidad de datos que existe entre los pares chino-inglés e inglés-castellano, exploramos y evaluamos diferentes alternativas al problema mediante estrategias indirectas usando el inglés como lengua pivote. Más concretamente, usamos aproximaciones como cascada de sistemas y generación de pseudo-corpus y las compramos con un sistema de traducción automática estadístico directo chino-castellano. La primera aproximación consiste en traducir primero de chino a inglés y después de inglés a castellano. Mientras que la segunda aproximación consiste en construir un corpus sintético chino-castellano a partir de los corpus chino-inglés e inglés-castellano.

Este artículo se estructura de la siguiente manera. La sección 2 hace un breve resumen sobre los trabajos que podemos encontrar en la literatura relacionados con la traducción de chino a castellano que básicamente se concentran en los trabajos provenientes del IWSLT 2008. A continuación, la sección 3 describe las principales estrategias de traducción usadas en este trabajo: traducción automática estadística directa, traducción en cascada y traducción a partir de pseudo-corpus. La sección 4 presenta el entorno de evaluación que incluye los detalles sobre estadísticas del corpus, del sistema y detalles de evaluación. Después la sección 5 reporta los experimentos y los resultados obtenidos. Finalmente, la sección 6 concluye nuestro trabajo y propone futuras direcciones de investigación en esta línea.

2. Trabajo relacionado

Sorprendentemente, no hay mucho trabajo publicado en el área de la traducción automática entre el par chino-castellano pese a ser los dos idiomas más hablados en el mundo. Uno de los primeros trabajos lo encontramos en (Banchs et al., 2006a) en el cual usan dos corpus independientes chino-inglés e inglés-castellano para construir un sistema chino-castellano. Como sistema de referencia, utilizan un sistema basado en n-gramas que difiere del sistema de segmentos en el modelo de traducción y reordenamiento.

El evento de investigación más conocido se llevó a cabo en 2008 con la evaluación internacional del IWSLT. Este evento propuso dos tareas entre el par chino-castellano. La primera de ellas consistía en realizar una tra-

* Este trabajo ha sido parcialmente financiado por el Ministerio de Ciencia e Innovación mediante un programa Juan de la Cierva y el proyecto BUCEADOR (TEC2009-14094-C04-01). Asimismo los autores quieren agradecer al Institute for Infocomm Research y a Barcelona Media Innovation Center por su apoyo y permiso para publicar esta investigación.

[1] www.ethnologue.org

ducción automática estadística directa entre chino-castellano. La segunda, motivada por el hecho que en la práctica exiten pocos datos chino-castellano pero muchos datos en chino-inglés e inglés castellano, proponía hacer una traducción de chino-castellano pero pasando por el inglés.

Repasando ambas tareas, el mejor sistema de cada una de ellas se desarrolló por el mismo grupo de autores (Wang et al., 2008). Respecto al sistema directo, usaron un sistema estándar de segmentos. Lo que marcó la diferencia con los otros sistemas que participaban es básicamente que proporcionaron su propia segmentación del chino y que usaron el diccionario bilingüe del LDC (Linguistic Data Consortium). Respecto a la tarea pivote, compararon dos aproximaciones distintas: la primera, entrenar dos modelos de traducción en chino-inglés e inglés-castellano, y después construir un modelo de traducción pivote para chino-castellano, es decir, se inspiraron en el sistema pivote que se presentó en (Wu y Wang, 2007); la segunda se basaba en el sistema en cascada y obtuvo mejores resultados. Recordemos que el sistema de cascada consiste en obtener la traducción de chino-castellano a través de la concatenación de traducciones de chino-inglés e inglés-castellano.

Entre ambas tareas, los mejores resultados entre chino-castellano resultaron ser los que se construyeron con sistemas pivotes, es decir, los resultantes de la segunda tarea.

Otros participantes también propusieron usar la metodología de cascada. La idea aquí es traducir de chino a inglés y después de inglés a castellano, lo que significa concatenar dos traducciones. Esta aproximación se puede hacer con una o con listas de las n mejores traducciones (Khalilov et al., 2008).

Finalmente, otra propuesta es la de generar pseudo-corpus que significa traducir o bien el inglés a chino o a castellano creando de este modo un corpus paralelo sintético chino-castellano. Este pseudo-corpus se usa para entrenar el sistema chino-castellano con el que se traducirá (Utiyama et al., 2008).

Además de la investigación explicada que directamente aborda el par chino-castellano, también encontramos en la literatura otros trabajos relacionados con estrategias pivote. Está el trabajo (previamente nombrado) de Wu (2007) y el de Cohn y Lapata (2007). Ambos usan diferentes idiomas pivote para crear segmentos fuente-destino que después usarán en el sistema directo construido a partir de corpus paralelo creado con un corpus paralelo fuente-destino. Entre estos dos trabajos básicamente hay las siguientes diferencias: Wu y Wang (2007) manejan las frecuencias relativas y los pesos léxicos mientras que Cohn y Lapata (2007) solamente tratan con las frecuencias relativas. Además, usan diferentes idiomas pivote. El primero, construye una tabla de segmentos fuente-destino para cada idioma pivote y después las interpola todas con el sistema TAE directo. El segundo considera todos los idiomas intermedios al mismo tiempo para constuir una única tabla de segmentos e interpolarla con el sistema TAE directo. De todos modos, hemos visto que la comparación de Wang et al. (2008) muestra que el sistema en cascada es mejor que la combinación de tablas de segmentos para la tarea pivote de chino-castellano.

Aunque existen páginas de traducción que proporcionan traducciones entre chino y castellano, la calidad de las traducciones resultantes no suele ser muy satisfactoria. Y, por lo que hemos visto en esta sección, realmente no hay mucha investigación de traducción automática estadística en este par de idiomas. El principal motivo puede ser la falta de recursos, especialmente de corpus paralelo. Este estudio pretende mostrar progreso e involucrar investigadores en el área de traducción automática estadística mediante la comparación de dos tecnologías pivote diferentes: la cascada (Wang et al., 2008; Khalilov et al., 2008) y la generación de pseudo-corpus (Utiyama et al., 2008; Bertoldi et al., 2008). Si conseguimos demostrar que la diferencia de calidad de traducción entre el sistema TAE directo y las estrategias pivote no es muy alta, entonces, se podría enfocar la traducción de chino-castellano directamente con una estrategia pivote. Lo que permite una estrategia pivote en el caso de usar inglés es que los recursos son mucho mayores tanto en el caso chino-inglés como inglés-chino.

3. *Aproximaciones de traducción automática estadística directa y pivote*

Existen diferentes aproximaciones para traducir un par de idiomas en traducción automática estadística. Las siguientes secciones presentan los detalles de las aproximaciones automática estadística directa y pivotes que usamos en este trabajo.

3.1. Sistema de traducción automática estadística directa

Nuestro sistema TAE directo es un sistema basado en segmentos (o sistema de segmentos) (Koehn, Och, y Marcu, 2003). Este conocido sistema implementa un modelo log-lineal en el cual una oración en el lenguaje fuente $f^J = f_1, f_2, \ldots, f_J$ se traduce en una oración en el lenguaje destino $e^I = e_1, e_2, \ldots, e_I$ buscando una hipótesis de traducción $\hat{e}^I$ que maximice la combinación log-lineal de varias funciones características (Och, 2003):

$$\hat{e}_I = \arg\max_{e^I} \{ \sum_{m=1}^{M} \lambda_m h_m(e^I, f^J) \} \qquad (1)$$

donde la función característica h_m hace referencia a la función característica m y la λ_m hace referencia al correspondiente peso optimizado de la misma función característica.

Las principales funciones características son el modelo de traducción y el modelo de lenguaje. El primero trata el problema de qué segmento destino f_j traduce la segmento fuente e_i y el último estima la probabilidad de la hipótesis de traducción. A parte de estos dos modelos, destaca el modelo de reordenamiento, que en particular nosotros usamos el reordenamiento lexicalizado que se define en (Tillman, 2004).

3.2. Sistema cascada

Esta aproximación utiliza dos traducciones independientemente: fuente-pivote y pivote-destino. Ambos se constuyen y optimizan para mejorar su calidad de traducción, independientemente de la tarea final, fuente-destino. Posteriormente se concatenan para traducir del lenguaje fuente al lenguaje destino en dos pasos: primero, se calcula la mejor traducción del sistema fuente-pivote y, después, esta mejor traducción se traduce nuevamente con el sistema pivote-destino para obtener el resultado final.

3.3. Sistema Pseudo-Corpus

Esta aproximación traduce el entrenamiento pivote a idioma destino del corpus paralelo fuente-pivote usando un sistema de traducción construido previamente con los datos pivote-destino. Con lo cual, estamos creando un corpus paralelo fuente-destino pero sintético. Después, se construye un sistema fuente-destino con este corpus paralelo sintético. El sistema basado en pseudo-corpus es de mayor calidad si se puede optimizar usando un conjunto de desarrollo basado en corpus fuente-destino original.

4. Entorno de experimentación

En esta sección introducimos los detalles del entorno de experimentación y evaluación. Describimos las estadísticas del corpus de las naciones unidas, así como realizamos la construcción de los sistemas y la evaluación.

4.1. Estadísticas del corpus

Por lo que sabemos, sólo están disponibles tres corpus para investigación en el par chino-castellano: el BTEC, la Biblia y las UN. El primero se usó en la evaluación del IWSLT 2008 y se reportan experimentos con estrategias pivote en papers como (Bertoldi et al., 2008). La Biblia se uso para própositos similares en el trabajo (Henríquez Q., Banchs, y Mariño, 2010). En este estudio, decidimos usar las UN porque es el mayor corpus de los tres. Por lo que conocemos, podría ser el mayor corpus paralelo que existe en el par chino-castellano (para fines de investigación) y que todavía no se ha explotado para hacer investigaciones entre este par ni para estrategias pivote.

Las UN proporcionan un corpus paralelo a nivel de oración. De este corpus paralelo, hemos de extraer un conjunto de entrenamiento, de desarrollo y de test. Para hacer los experimentos comparables, todos los pares de idiomas (chino-castellano, chino-inglés e inglés-castellano) usan el mismo conjunto de entrenamiento, de desarrollo y de test que se construyó siguiendo las pautas que describimos a continuación:

1. Todos los corpus se tokenizaron, usando el tokenizador disponible en MOSES (Koehn et al., 2007) para castellano e inglés; se usó el ictclass (Zhang et al., 2003) para el chino.
2. El castellano y el inglés se pasaron a minúsculas.
3. Si una oración tiene más de 100 palabras en cualquier idioma, se elimina de todos los corpus.
4. Si una oración tiene una ratio de palabras mayor que tres para cualquier par de lenguas, se elimina de todos los corpus.

5. Para todos los idiomas, identificamos todas las oraciones que ocurren una única vez en el corpus y que difieren de cualquier otra oración presente en el corpus, después, extraemos una muestra de este subcorpus como desarrollo y test. El resto de oraciones se dejan como corpus de entrenamiento.

La tabla 1 muestra las principales características de todos los corpus. Se ha usado chino mandarín simplificado.

4.2. Detalles de los sistemas

Para construir los sistemas de segmentos, usamos tecnología Moses (Koehn et al., 2007). Para todos los sistemas, usamos los parámetros por defecto de Moses que incluye alineamiento *grow-diag-final-and*, reordenamiento lexicalizado, modelo de lenguaje 5-gramas usando suavizado Kneser-Ney y segmentos hasta longitud 10. La optimización se hizo usando MERT (Minimum Error Rate Training) (Och, 2003).

4.3. Detalles de la evaluación

Para evaluar la calidad de la traducción usamos la métrica estándar BLEU (Papineni et al., 2001). Además, para evaluar la significancia de los resultados utilizamos la técnica denominada "Pair Bootstrap Resampling" presentada en (Koehn, 2004).

5. Experimentación con las aproximaciones automática estadística directa y pivotes

Básicamente llevamos a cabo tres aproximaciones de traducción automática estadística para traducir el par chino-castellano: aproximación automática estadística directa, aproximación pivote por cascada y aproximación basada en pseudo-corpus.

Utilizamos inglés, castellano y chino para construir y comparar diferentes sistemas de traducción. El objetivo es estudiar el impacto de la estrategia pivote para traducir de chino a castellano y ver si estas pueden mejorar el sistema TAE directo. Específicamente, construimos tres sistemas: el sistema TAE directo chino-castellano; el sistema pivote basado en cascada; y el sistema pivote basado en pseudo-corpus.

Para los sistemas pivotes necesitamos construir los sistemas chino-inglés e inglés-chino. La tabla 2 muestra el BLEU conseguido con los sistemas intermedios entrenados con el corpus de las UN.

	BLEU
Chinese-English	58.80
English-Spanish	71.93

Tabla 2: Sistemas pivote con el corpus de las UN

La tabla 3 muestra los resultados chino-castellano usando el corpus de las naciones unidas. Como podemos ver en la tabla 2 el BLEU en las tareas chino-inglés e inglés-castellano es más alto que el BLEU para chino-castellano. Esto parece indicar que la tarea chino-castellano es más compleja.

	Sistema	BLEU
chino-castellano	directo	**53.66**
chino-inglés-castellano	cascada	**53.39**
chino-inglés-castellano	pseudo	**53.07**

Tabla 3: Resultados para el sistema directo y los sistemas pivote

La aproximación basada en cascada mejora mínimamente la aproximación basada en pseudo-corpus, pero la diferencia no es significativa como hemos comprobado con el método (Koehn, 2004). Estos resultados no coinciden con los mostrados en trabajos previos para el mismo par de lenguas (Bertoldi et al., 2008) (Henríquez Q., Banchs, y Mariño, 2010). En estos trabajos previos el corpus con el que se trabajaba era menor y con un vocabulario muy restringido. Asimismo, no se mostraron resultados de significancia, lo cual pone en cuestión las diferencias entre los trabajos.

Asimismo, no hay diferencia significativa entre el sistema TAE directo y el en cascada. Lo cual nos lleva a concluir que para realizar un sistema chino-castellano podemos usar un sistema TAE directo o un sistema basado en traducción chino-inglés e inglés-castellano. Esto es muy ventajoso dada la gran cantidad de recursos que existe entre estos últimos pares comparado con los recursos entre chino-castellano.

Con el objetivo de realizar un estudio más detallado, realizamos un análisis manual para observar los casos en que la aproximación cascada mejoraba el sistema TAE directo. Evaluamos el BLEU usando todas las oraciones de test individualmente.

	entrenamiento			desarrollo			test		
	or	pal	vocab	or	pal	vocab	or	pal	vocab
chino	58,679	1,689,324	18,040	1,004	30,927	3,512	1,005	32,560	3,616
castellano	58,679	2,296,123	20,530	1,004	42,176	4,374	1,005	44,006	4,509
inglés	58,679	2,005,591	14,728	1,004	36,681	3,710	1,005	38,285	3,836

Tabla 1: Estadísticas del corpus UN. Or significa oraciones; pal, palabras; y vocab, vocabulario

DIRECTO	cuestiones como a que consideren seriamente la posibilidad de ratificar la tortura y otros tratos o penas crueles , inhumanos o degradantes ;
CASCADA	como cuestiones a que consideren seriamente la posibilidad de ratificar **la convención contra** la tortura y otros tratos o penas crueles , inhumanos o degradantes ;
REF	considere seriamente la posibilidad de ratificar , con carácter prioritario , la convención contra la tortura y otros tratos o penas crueles , inhumanos o degradantes
DIRECTO	habiendo examinado el segundo informe de la comisión y la recomendación que figura en él
CASCADA	habiendo examinado el segundo informe de la comisión **de verificación de poderes** y las recomendaciones que figuran en él
REF	habiendo examinado el segundo informe de la comisión de verificación de poderes y la recomendación que figura en él
DIRECTO	pide al secretario general que prepare un informe sobre la aplicación de esta resolución a la asamblea general , quincuagésimo sexto período de sesiones
CASCADA	pide al secretario general que prepare un informe sobre la aplicación de la presente resolución **para su examen por la asamblea general** en su quincuagésimo sexto período de sesiones
REF	pide al secretario general que prepare un informe sobre la aplicación de la presente resolución , que será examinado por la asamblea general en su quincuagésimo sexto período de sesiones

Tabla 4: Ejemplos donde el sistema cascada es mejor que el sistema directo.

Obtuvimos que 460 oraciones eran mejores en el sistema TAE directo y 433 eran mejores en el sistema cascada (el resto eran similares). Revisamos algunas de las oraciones que eran mejores en la aproximación cascada y reportamos algunos ejemplos en la tabla 4.

6. Conclusiones

Este trabajo reporta un breve estudio del estado de la cuestión en la traducción estadística chino-castellano. Este par de lenguas es muy interesante en términos económicos y culturales si tenemos en cuenta el alto número de parlantes chinos y castellanos. Además, la traducción automática estadística es una de las aproximaciones más usadas en evaluaciones internacionales.

La comparación entre sistemas pivote y un sistema de TAE directo es la principal aportación de este trabajo. Hemos visto que el sistema directo mejora los sistemas pivotes pero la calidad de traducción no es muy diferente entre el sistema directo y el mejor de los sistemas pivotes que ha resultado ser el sistema basado en cascada en lugar del sistema basado en pseudo-corpus.

Estas conclusiones son muy ventajosas porque sabemos que podemos construir un sistema de chino-castellano a partir de corpus chino-inglés e inglés-castellano, que presentan muchos más recursos que el par chino-castellano.

Como trabajo futuro, planificamos investigar en extracción automática de corpus paralelo a partir de corpus comparable para mejorar la calidad chino-castellano así como investigar si el inglés es el mejor idioma pivote para el par chino-castellano o si la combinación de diferentes idiomas pivotes junto con diferentes aproximaciones pivote puede ayudar en este par.

Bibliografía

Banchs, R. y H. Li. 2008. Exploring spanish morphology effects on chinese-spanish smt. En *MATMT 2008: Mixing Approaches to Machine Translation*, páginas 49–53, Donostia-San Sebastian, Spain, February.

Banchs, R. E., J. M. Crego, P. Lambert, y J. B. Mariño. 2006a. A Feasibility Study For Chinese-Spanish Statistical Machine Translation. En *Proc. of the 5th Int. Symposium on Chinese Spoken Language Processing (ISCSLP)CONLL*, páginas 681–692, Kent Ridge, Singapore, December 13–16.

Banchs, R. E., J. M. Crego, P. Lambert, y J. B. Mariño. 2006b. Chinese-Spanish statistical machine translation experiments. En *Memorias de las 4tas Jornadas de Tecnologías del Habla*, Zaragoza, Spain.

Bertoldi, N., R. Cattoni, M. Federico, y M. Barbaiani. 2008. FBK @ IWSLT-2008. En *Proc. of the International Workshop on Spoken Language Translation*, páginas 34–38, Hawaii, USA.

Cohn, T. y M. Lapata. 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. En *Proc. of the ACL*.

Henríquez Q., C. A., R. E. Banchs, y J. B. Mariño. 2010. Learning reordering models for statistical machine translation with a pivot language. Internal Report TALP-UPC.

Khalilov, M., M. R. Costa-Jussà, C. A. Henríquez, J. A. R. Fonollosa, A. Hernández, J. B. Mariño, R. E. Banchs, B. Chen, M. Zhang, A. Aw, y H. Li. 2008. The TALP & I2R SMT Systems for IWSLT 2008. En *Proc. of the International Workshop on Spoken Language Translation*, páginas 116–123, Hawaii, USA.

Koehn, P. 2004. Statistical significance tests for machine translation evaluation. En *Proceedings of EMNLP*, volumen 4, páginas 388–395.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, y E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. En *Proc. of the ACL*, páginas 177–180, Prague, Czech Republic.

Koehn, P., F. Och, y D. Marcu. 2003. Statistical phrase-based translation. En *HLT-NAACL*, páginas 48–54.

Och, F.J. 2003. Minimum error rate training in statistical machine translation. En *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*, páginas 160–167.

Papineni, K., S. Roukos, T. Ward, y W.J. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. IBM Research Report, RC22176, September.

Rafalovith, A. y R. Dale. 2009. United nations general assembly resolutions: A six-language parallel corpus. En *Proc. of the MT Summit XII*, páginas 292–299, Ottawa.

Tillman, C. 2004. A block orientation model for statistical machine translation. En *HLT-NAACL*.

Utiyama, M., A. Finch, H. Okuma, M. Paul, H. Cao, H. Yamamoto, K. Yasuda, y E. Sumita. 2008. The NICT/ATR Speech Translation System for IWSLT 2008. En *Proc. of the International Workshop on Spoken Language Translation*, páginas 77–84, Hawaii, USA.

Wang, H., H. Wu, X. Hu, Z. Liu, J. Li, D. Ren, y Z.Ñiu. 2008. The TCH Machine Translation System for IWSLT 2008. En *Proc. of the International Workshop on Spoken Language Translation*, páginas 124–131, Hawaii, USA.

Wu, H. y H. Wang. 2007. Pivot Language Approach for Phrase-Based Statistical Machine Translation. En *Proc. of the ACL*, páginas 856–863, Prague.

Zhang, H., H. Yu, D. Xiong, y Q. Liu. 2003. HHMM-based chinese lexical analyzer ICTCLAS. En *Proc. of the 2nd SIGHAN Workshop on Chinese language processing*, páginas 184–187, Sapporo, Japan, July.