

La tradautomaticidad: un concepto aplicado a la evaluación de sistemas de traducción automática

Joaquim Moré López

Universitat Oberta de Catalunya (UOC)
jmore@uoc.edu

Salvador Climent Roca

Universitat Oberta de Catalunya (UOC)
scliment@uoc.edu

Resumen: En este artículo, presentamos el concepto de *tradautomaticidad* y su aplicación a la evaluación de sistemas de traducción automática (TA). La tradautomaticidad se refiere al output generado por un traductor automático que el receptor no atribuiría a un traductor humano. La tradautomaticidad tiene una relación directa con la calidad de la traducción: cuantos más ejemplos de tradautomaticidad existan peor es la traducción. Para demostrar la utilidad de la detección de estos ejemplos presentamos un método de evaluación de coste bajo que consiste en identificar y cuantificar ejemplos de tradautomaticidad mediante búsquedas por Internet. Además, los ejemplos detectados pueden aprovecharse para otros usos, por ejemplo, la postedición automática de documentos traducidos automáticamente.

Palabras clave: evaluación, tradautomaticidad, traducción automática

Abstract: In this article we introduce the concept of *machine translationness* and its use for MT evaluations. Machine translationness is the output generated by an MT system which is unlikely to be attributed to a human translator. Machine translationness is closely related to the translation quality. The more instances of machine translationness the worse is the translation. In order to show the use of these instances we explain a cheap evaluation method that consists in identifying and quantifying instances of machine translationness by performing Internet searches. Besides, the instances detected can be reused, for example, in the automatic postedition of machine translated documents.

Keywords: evaluation, machine translationness, machine translation

1 Introducción

En este artículo presentamos un concepto que aquí denominaremos *tradautomaticidad* y su posible aplicación en un método de evaluación de traducciones automáticas. La tradautomaticidad es un término que hemos creado para referirnos al carácter que tienen algunos segmentos de texto generados por un sistema de TA que difícilmente podrían ser atribuibles a un traductor humano. Ejemplificaremos la relevancia de este concepto con un método de evaluación de TA basado en esta noción y mostraremos un prototipo que detecta de forma rápida y económica ejemplos de tradautomaticidad que afectan muy negativamente a la calidad de las traducciones.

Este método nació de la necesidad real de evaluar las traducciones automáticas de los documentos de la Universitat Oberta de Catalunya (UOC) y creemos que es una metodología prometedora porque permite tener una primera impresión de la calidad de las traducciones que puede ser suficiente según el propósito de la evaluación. Además, ahorra tiempo y dinero a organizaciones que requieren de la TA y necesitan evaluar continuamente su output para mejorar su producción.

2 La noción de tradautomaticidad

La tradautomaticidad es un término que hemos creado para referirnos a los rasgos característicos de una traducción automática que difícilmente se encuentran en las

traducciones humanas. Podríamos definir la tradautomaticidad como el ‘aroma de la traducción automática’, es decir, la aparición en un texto de rasgos lingüísticos que indican al receptor que es una traducción y que no ha sido realizada por un ser humano. En otras palabras, la tradautomaticidad correspondería a las características textuales que impedirían que la traducción pasara el test de Turing. En este artículo nos ocuparemos principalmente de las soluciones de traducción de un sistema que añade tradautomaticidad a su output. Por ejemplo, si un sistema catalán-español traduce el original *morin de set* (‘mueran de sed’) como *mueran de siete* añade un ejemplo de tradautomaticidad a su traducción.

3 Motivación de un método de evaluación basado en la tradautomaticidad

La UOC es una universidad virtual que traduce la mayoría de sus materiales educativos del catalán al español para los alumnos no catalanohablantes. Por otro lado, los documentos originalmente escritos en español se traducen al catalán para el Campus Virtual en catalán. El volumen de documentos es tal que ha sido necesario recurrir a la TA para reducir los costes en tiempo y dinero derivados de las necesidades de traducción de la institución. Puesto que los costes de la postedición dependen de la calidad del output, el Servicio Lingüístico de la UOC se ha ocupado de evaluar la calidad de su sistema de TA y también de detectar los errores sistemáticos que se pueden solucionar automáticamente. Además, el Servicio Lingüístico también dota al sistema de terminología nueva y recursos tales como memorias de traducción para mejorar la calidad del output. Por lo tanto, es necesario producir evaluaciones de forma continua para confirmar las mejoras que se han hecho.

El Servicio Lingüístico eligió el cálculo del BLEU (Papineni et al. 2001) como método de evaluación del sistema de TA por su coste en tiempo y dinero más reducido comparado con la evaluación humana. El Servicio Lingüístico también preparó los recursos necesarios para realizar futuras evaluaciones de sistemas con otros pares de lenguas como el catalán-inglés y el inglés-catalán.

Para cada lengua origen (inglés, español y catalán) se obtuvo un conjunto de 500 segmentos (correspondientes más o menos a

una frase) tomados de periódicos, páginas Web de turismo, documentos administrativos e informes económicos. El Servicio Lingüístico también se ocupó de las traducciones de referencia para cada segmento en los siguientes pares de lenguas: catalán-español, español-catalán, inglés-catalán y catalán-inglés. Estas traducciones fueron realizadas por cuatro traductores profesionales, nativos en la lengua destino que tenían una larga experiencia profesional, con licenciaturas y diplomas que acreditaban sus habilidades traductoras. A pesar de que los segmentos fueran descontextualizados y aparecieran de forma aleatoria, tenían sentido por ellos mismos y podían ser traducidos fielmente al original.

Se analizaron las referencias entregadas por los traductores humanos para garantizar que éstas no distorsionarían la evaluación pero pronto se hizo evidente que todos los 8,000 segmentos (2,000 para cada par de lenguas) tenían que revisarse porque apareció un número significativo de referencias que efectivamente podían distorsionar la evaluación. Casi todos los traductores, tanto los que trabajaron en la misma dirección de pares de lenguas, como los que trabajaron en direcciones distintas realizaron traducciones de referencia problemáticas. Veamos un ejemplo de traducción del español al catalán (1), que es problemático debido a la decisión por parte del traductor de no realizar una traducción palabra a palabra

- (1) Magnífico hotel ecológico rodeado de exuberante naturaleza. (Original)

Magnífic hotel ecològic envoltat de vegetació exuberant. (Referencia en catalán)

Magnífic hotel ecològic envoltat d'exuberant naturalesa. (Traducción automática)

Los revisores estuvieron mucho tiempo discutiendo si la traducción de *naturaleza* como *vegetació* (vegetación) era o no legítima; con lo cual la decisión sobre la legitimidad de la referencia supuso más esfuerzo que la evaluación manual de la traducción automática. La decisión era importante ya que si *naturalesa* no hubiera aparecido en ninguna referencia, un sistema que hubiera hecho una traducción

legítima palabra a palabra habría sido injustamente penalizado.

Concluimos, por tanto, que las revisiones de las referencias hacían que la evaluación fuera todavía más cara en tiempo y recursos. Así que intentamos diseñar un método alternativo que no necesitara de traducciones de referencia y que nos facilitara un diagnóstico rápido del comportamiento del sistema que ahorrara tiempo y dinero.

4 Diseño del método de evaluación

4.1 Motivación del diseño

De las propuestas de evaluación automática sin traducciones de referencia nos han interesado aquellas que parten de la asunción de que una traducción identificada como automática y no como humana es una mala traducción. Por consiguiente, la evaluación consiste en una clasificación de los textos evaluados en traducciones humanas y traducciones automáticas: cuanto más seguro está el evaluador de clasificar un texto como una traducción automática, su calidad es peor; y cuanto más seguro está de clasificarlo como una traducción humana, su calidad es mejor (Corston-Oliver et al (2001), Kulesza y Shieber (2004)). Nos interesa este enfoque principalmente porque, aunque una traducción automática que parece humana puede no ser fiel al original, consideramos que nos permite tener una primera impresión fiable de la calidad del output generado. Ya que queremos realizar evaluaciones continuas, esta visión puede ser suficiente y podemos dejar que evaluadores humanos realicen un análisis más profundo de la fluidez y la fidelidad de segmentos que aparecen con más frecuencia en los documentos de la institución y que tienen un contenido muy importante. Otras ventajas de este enfoque es que no es necesario evaluar un corpus muy grande para determinar si un sistema genera traducciones ‘muy automáticas’ (Reeder, 2001). Por otro lado, implica detectar errores de traducción sistemáticos que pueden aprovecharse para desarrollar un módulo de postedición automática de los textos generados por el sistema, con lo cual se reducen los costes de corrección (Gamon et al., 2005). Por último, estos evaluadores suelen clasificar las traducciones tras haber aprendido a identificar los rasgos característicos que le llevarán a clasificar un texto como una traducción automática y no humana. Una vez entrenado, el

evaluador puede evaluar traducciones nuevas con rapidez y con un coste mínimo. Sin embargo, el aprendizaje automático de las características propias de las traducciones humanas y automáticas requiere de corpus de entrenamiento con traducciones de los dos tipos cuya creación puede ser costosa a lo que hay que añadir la anotación de este corpus con la información lingüística, semántica, etc. necesaria para el aprendizaje (Gamon et al., 2005). El diseño de evaluación que presentamos pretende detectar características propias de las traducciones automáticas sin necesidad de realizar un corpus de entrenamiento.

4.2 Descripción del método

El método de evaluación que presentamos consiste en detectar y cuantificar los ejemplos de tradautomatización de un corpus de evaluación. Según la explicación del apartado 2, un ejemplo de tradautomatización es una solución de traducción elegida por el sistema entre un conjunto de posibles soluciones y cuya generación es muy poco probable si se compara con la solución que un traductor humano habría elegido y que el receptor habría aprobado.

Relacionamos la probabilidad de que un traductor humano genere una posible traducción con el número de apariciones de dicha traducción en un corpus representativo del uso general de la lengua destino. Por otro lado, relacionamos la reacción de los receptores ante una solución de traducción con la expectativa de encontrarla en un texto fluido. La expectativa se infiere comparando el número de apariciones de cada posible solución de traducción en el corpus representativo. Por tanto, un ejemplo de tradautomatización cumple con esta condición: dado un segmento de texto original SC y un segmento TC_i que es la traducción de SC generada por el sistema entre TC_1, TC_2, \dots, TC_n posibles traducciones, TC_i es un ejemplo de tradautomatización si el número de apariciones de TC_i en un corpus representativo de la lengua destino es ampliamente superado por el número de resultados de cualquiera de las otras soluciones posibles.

Por razones prácticas, tomamos el total de páginas Web publicadas como corpus representativo, siempre que la lengua destino esté ampliamente presente en la Red. Si una solución de traducción no aparece en ninguna página Web se hipotetiza que la probabilidad de ser generada por un traductor humano es escasa

y si una solución tiene un número de apariciones mucho mayor que otra solución, se hipotetiza que la primera solución es más probable de ser generada por el traductor humano que la segunda y, por tanto, la reacción del receptor será probablemente más positiva que si se enfrentara con la solución menos probable. Por ejemplo, el ejemplo de tradautomatización *vuelan esconder* del original catalán *volen amagar*, no se encuentra en ninguna página Web utilizando los motores de búsqueda Yahoo y Google (última consulta 10-02-06), mientras que el motor de Google encuentra 772 resultados de *quieren esconder*, que es una solución de traducción posible.

El método tiene las siguientes fases: etiquetaje de las traducciones automáticas, creación de los segmentos de traducciones automáticas, creación de segmentos alternativos, detección de ejemplos de tradautomatización y, si se comparan sistemas o versiones del mismo sistema, comparación de resultados.

Etiquetaje de traducciones automáticas.

El output se etiqueta con un etiquetador automático. Para la confección del prototipo (ver sección 5) utilizamos el etiquetador del analizador de código abierto Freeling para el español (Atserias et al. (2006)). Cada palabra se etiqueta con su forma, lema y categoría gramatical.

Creación de segmentos. Las palabras etiquetadas de cada frase se agrupan en segmentos. Los segmentos que hemos establecido son los siguientes: sintagmas nominales, verbos (simples y complejos), sintagmas adjetivales que realizan la función de complemento verbal, sintagmas adverbiales, y sintagmas preposicionales que son adjuntos y sintagmas relacionados. Entendemos por sintagmas relacionados aquellos que se concatenan sin ningún signo de puntuación en medio y que expresan una relación entre dos conceptos. De momento consideramos la concatenación de un sintagma nominal con un verbo, un sintagma nominal relacionado con un sintagma adjetival mediante un verbo, dos sintagmas nominales juntos y un verbo con un sintagma preposicional que realiza la función de argumento.

Creación de segmentos alternativos. Para cada segmento se crea una traducción alternativa. Una alternativa del segmento *C* es un segmento *C'* creado automáticamente

mediante una de las siguientes acciones, que referiremos como A1 y A2:

A1. Sustituir la traducción de una palabra en mayúscula por su correspondiente palabra original (e.g: Catalán: *memòria RAM* ; C: *memoria RAMO*; C': *memoria RAM*).

A2. Cuando haya una palabra traducida TW cuyo original correspondiente pueda tener una traducción distinta TW', sustituir TW por TW'.

Catalán	C	SW	TW	TW'	C'
Sortida vol	Salida quiere	vol	quiere	vuelo	Salida vuelo

Figura 1: Ejemplo A2

Hasta ahora hemos establecido estas dos acciones pero se podrían realizar otras acciones que fueran más allá de la selección léxica.

Para crear automáticamente segmentos alternativos hace falta tener los formularios de palabras de las lenguas origen y destino, con información sobre la forma, el lema y la categoría gramatical de cada palabra, y una lista de pares <palabra original, palabra traducida>, donde 'palabra traducida' es el equivalente de traducción de la palabra original. Por ejemplo, la alternativa *morir de sed* para *morir de siete* se crea cuando se encuentran los siguientes pares <set, siete> y <set, sed>.

Detección de ejemplos de tradautomatización. De modo similar a cómo se seleccionan candidatos de traducción en (Grefenstette, 1999), para cada segmento de la traducción automática el detector obtiene el número de páginas Web en el que aparece gracias a un motor de búsqueda por Internet. Cuando no hay resultados, el segmento se pone en la lista de candidatos a ejemplos de tradautomatización. Cuando el segmento tiene alternativas éstas también se buscan por Internet y sus resultados se comparan con los del segmento a evaluar. Si el número de resultados de una alternativa supera ampliamente el número de resultados del segmento, éste se considera un ejemplo de tradautomatización. Los ejemplos de tradautomatización se guardan en una lista. De forma provisional establecemos que un segmento de traducción automática con menos de 50 resultados es un ejemplo de

tradautomatización. Si supera este límite y existe una solución alternativa con un número como mínimo cinco veces mayor el segmento también es considerado como un ejemplo de tradautomatización.

Comparación de resultados. El número de ejemplos de tradautomatización del sistema A o su versión más actualizada se compara con el número de ejemplos del sistema B o de su versión previa. Cuanto menor sea el número mejor es el sistema o la versión. Las listas de candidatos de A y B también se comparan. Si una de las listas tiene un candidato que no está en la otra lista, este candidato se cuenta como un ejemplo de tradautomatización.

5 Prototipo del método de evaluación

Para probar la viabilidad del método, intentamos encontrar ejemplos de tradautomatización en las traducciones automáticas al español de 500 segmentos en catalán preparados por el Servicio Lingüístico de la UOC (ver Sección 2). Las traducciones fueron realizadas con el sistema de código abierto Internostrum¹ porque los recursos de este sistema pueden obtenerse libremente por lo que los formularios en catalán y español y la lista de pares <palabra original, palabra traducida> se han generado automáticamente. Elegimos la dirección catalán-español porque es la dirección con mayor producción de la institución. De los 396 errores detectados manualmente destacamos los siguientes:

- **Confusión a nivel de lema (34,4 %)**

Entre los sentidos variados de una palabra del original, el sistema traduce según un sentido que no es fiel al original. Cuando la frase catalana *morin de set* ('mueran de sed') se traduce como *mueran de siete* el sistema ha interpretado un sentido distinto de la palabra *set*.

- **Confusión a nivel de forma (13%)**

El sistema se confunde por la coincidencia de forma de la palabra original con otra palabra original cuyo significado no corresponde al contexto. Por ejemplo, el sustantivo catalán *vol* ('vuelo') coincide con la tercera persona del

singular del presente de indicativo del verbo *voler* ('querer'). Esto explica que *sortida vol* se traduzca como *sortida quiere* en la Figura 1.

- **Traducción ilegítima palabra a palabra (11,4%)**

Traducción no adecuada de acrónimos (e.g: *la memoria RAMO*), traducción de modismos (e.g: *hacer el melocotón*, como traducción de *fer el préssec* que significa 'hacer el canelo'), artículos delante de nombres propios (*el Irán*), etc.

- **No apocopación (1,7%)**

Ejemplos como *un grande momento* o *el primero ministro*.

- **Uso impropio de 'ser' y 'estar' (0,7%)**

Ejemplos como *el disco es lleno*.

Estos fenómenos son la causa del 61,2% de los errores detectados. El resto se reparte en errores que pueden ser fácilmente detectados por cualquier corrector ortográfico y gramatical como palabras no traducidas por errores tipográficos del original (19,2%) o por no tener un equivalente en el diccionario bilingüe (10%), errores de concordancia de género o persona verbal (4,3%), o errores de contracción ortográfica y fonología sintáctica como *de el* o *hizo* (0,7%). Finalmente, un 4,3% de los errores tienen causas variadas pero no sistemáticas.

En la Tabla 2 mostramos algunos ejemplos de tradautomatización que se pueden detectar con el método. En la columna **SO** aparecen los segmentos originales. En la columna **STA** los segmentos traducidos automáticamente y en **STalt** las traducciones alternativas de STA. En las columnas con **Res.** se indican los resultados de la búsqueda por Internet de cada tipo de segmento.

¹ www.internostrum.com

Tipo de error	SO	STA	Res. STA	STalt	Res. STAlt
<i>Confusión a nivel de lema</i>	Morin de set	Mueran de siete	0	Mueran de sed	164
	Jornada sagnant	Jornada sangrante	0	Jornada sangrienta	32100
<i>Confusión a nivel de forma</i>	Sortida vol	Salida quiere	61	Salida vuelo	310
	Sortir a sopar	Salir a cena	7	Salir a cenar	19200
	Endeumentament net	Endeudamiento limpio	0	Endeudamiento neto	1450
<i>No apocopcación</i>	Una gran festa	Una grande fiesta	167	Una gran fiesta	188000
	Primer contacte	Primero contacto	416	Primer contacto	492000
<i>Traducción ilegítima palabra a palabra</i>	Fer el préssec	Hacer el melocotón	0		
	Memòria RAM	Memoria RAMO	6	Memoria RAM	1320000
<i>Uso impropio de ser-estar</i>	El disc és ple	El disco es lleno	0	El disco está lleno	398
	És previst d'arribar	Es previsto llegar	0	Está previsto llegar	200

Tabla 2: Ejemplos de tradautomatidad detectados con búsquedas a Internet

6 Discusión

Más de un 90% de traducciones erróneas del test de evaluación se pueden detectar con nuestro método, si se le añade además un corrector ortográfico y gramatical. Por consiguiente se pueden detectar la mayoría de ejemplos de tradautomatidad. Con recursos

gratuitos (páginas Web de la Red, formularios y un etiquetador de código libre) y recursos asequibles como los correctores ortográficos y gramaticales de los programas de edición más extendidos, se puede tener una 'primera impresión' de la calidad de un sistema de TA. En ocasiones, esta primera impresión es suficiente para el propósito de la evaluación y es el primer paso para un análisis más profundo y exhaustivo. Por otra parte el método permitiría elaborar una estrategia de mejora del sistema y obtener información para el desarrollo de un módulo de postedición automática. Por lo tanto, los costes serían superados ampliamente por los beneficios de los resultados obtenidos y por la posibilidad de reutilizar estos resultados.

Hay, sin embargo, dos aspectos que merecen ser comentados. En primer lugar, el hecho de que un segmento no aparezca en ninguna página Web no siempre es una indicación directa de que el segmento sea un error de traducción. Por ejemplo, un segmento gramatical en español como *mataron a Rigoberto Mallofré* no aparece en ninguna página Web porque *Rigoberto Mallofré* es un individuo que no tiene ninguna referencia en la Red.

En segundo lugar, la aparición de un segmento en muchas páginas Web no siempre es un dato significativo para determinar que no es un ejemplo de tradautomatidad. Por ejemplo, *plano de estudio*, que es una traducción incorrecta del catalán *pla d'estudis*, coincide con el término portugués. Por otro lado, hay que tener en cuenta la presencia de blogs, páginas Web con un uso lingüístico descuidado e incluso páginas Web que han sido traducidas automáticamente y que no han sido posteditadas. Por ejemplo, *disco llevar* como traducción del catalán *disc dur* aparece en una página Web traducida automáticamente. Sin embargo, el número de resultados de la mayoría de estos segmentos está ampliamente superado por la alternativa de traducción correcta; con lo cual se demuestra que son ejemplos de tradautomatidad. Por ejemplo, *disco llevar* tiene 63 resultados mientras que *disco duro* tiene 8.540.000. El segmento también puede desaparecer cuando éste coexiste con otro segmento en una petición de búsqueda más larga. Por ejemplo, *los Bocados* como traducción del catalán *els Mossos*, refiriéndose al cuerpo de policía catalán, tiene 369 resultados pero *los Bocados detectan* no tiene ninguno.

7 Conclusiones

Nuestro método, combinado con un corrector gramatical y ortográfico, puede detectar más del 90% de los errores de traducción de nuestro test de evaluación y, por consiguiente, la mayoría de ejemplos de tradautomaticidad. El método todavía está en una fase preliminar pero los primeros resultados obtenidos nos animan a continuar desarrollándolo. Es un método barato cuyos recursos son gratuitos o al alcance de cualquier usuario, y sin necesidad de realizar corpus de entrenamiento. Además, los resultados son significativos porque son consecuentes con la idea de que los evaluadores humanos penalizan las traducciones que identifican como automáticas y no las que creen que pueden pasar por humanas. Por otro lado, es posible tener un rápido diagnóstico de la calidad de las traducciones que puede ser suficiente para el propósito de la evaluación, elaborando un análisis más profundo cuando el propósito lo requiera o la impresión primera de la calidad no sea del todo concluyente.

Los ejemplos de tradautomaticidad detectados pueden aprovecharse para un módulo de postedición automático. Incluso el método de detección podría utilizarse para comprobar si una página Web ha sido traducida automáticamente y su publicación no ha pasado por el proceso de postedición.

8 Trabajo futuro

Hasta el momento, hemos analizado los resultados de los segmentos que son nominales y ahora estamos analizando los resultados de los segmentos nominales combinados con verbos. Por lo tanto, estamos en plena evaluación del método. Nuestro principal objetivo es, una vez hayamos obtenido todos los segmentos del corpus de evaluación y sus combinaciones, hayamos cuantificado su presencia en la red, y hayamos comparado sus resultados con los de sus alternativas de traducción, poder tener datos sobre el porcentaje de ejemplos de tradautomaticidad detectados correctamente y ejemplos detectados erróneamente. Además, deberemos comprobar si la eficacia del método depende de la similitud sintáctica del catalán y español. Por esta razón, aplicaremos el método para evaluar las traducciones de sistemas cuyos pares de lenguas tienen una sintaxis más alejada (español-inglés,

por ejemplo). Trataremos de encontrar nuevas acciones que generen automáticamente traducciones alternativas para detectar ejemplos de tradautomaticidad sintáctica, además de las de tipo léxico que hemos visto mayormente aquí. No sólo será interesante ver cómo funciona el método para otros pares de lenguas sino comprobar si hay casos típicos de tradautomaticidad según el método del motor de traducción. Saber, por ejemplo, si los sistemas basados en reglas y los sistemas estadísticos tienen cada uno de ellos ejemplos de tradautomaticidad que les son característicos.

Por otro lado, intentaremos elaborar una estrategia que no penalice los segmentos de traducción que contienen nombres propios inexistentes o muy poco presentes en la Red. Finalmente, aprovecharemos nuestro método para tareas de postedición.

Bibliografía

- Atserias, J., B. Casas, E. Comelles, M. González, L. Padró y M. Padró (2006). *FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*. Génova, Italia
- Corston-Oliver, S., M. Gamon y C. Brockett (2001). A machine learning approach to the automatic evaluation of machine translation. *Proceedings of the Association for Computational Linguistics*. Toulouse, France. pp. 140-14.
- Gamon, M., A. Aue, y M. Smets (2005). Sentence-level MT Evaluation Without Reference Translations: Beyond Language Modeling. *Proceedings of the 10th Annual EAMT Conference*. Budapest.
- Grefenstette, G. (1999). The www as a resource for example-based mt tasks. Machine Translation Task, Proc. Of Aslib Conference on Translating and the Computer. London
- Kulesza, A. y S. M. Shieber (2004). A learning Approach to Improving Sentence-Level MT Evaluation. *Proceedings of the 10th*

International Conference on Theoretical and Methodological Issues in Machine Translation. Baltimore.

Moré, J. y S. Climent (2006). A Cheap MT-Evaluation based on Internet Searches. *Proceedings of the 11th Annual Conference of the EAMT Conference*. Oslo.

Papineni, K., S. Roukos, T. Ward i W-J. Zhu (2001). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of ACL, Philadelphia, PA*.

Reeder, F. (2001). In One Hundred Words or Less. *MT Evaluation Workshop MT Summit VIII*. Santiago de Compostela.