Evaluación de un sistema de traducción automática basado en reglas o por qué BLEU sólo sirve para lo que sirve

Aingeru Mayor, Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, Kepa Sarasola
Euskal Herriko Unibertsitatea
Manuel Lardizabal 1, 20018 Donostia
aingeru@ehu.es

Resumen: *Matxin* es un sistema de traducción automática basado en reglas que traduce a euskera. Para su evaluación hemos usado la métrica HTER que calcula el coste de postedición, concluyendo que un editor necesitaría cambiar 4 de cada 10 palabras para corregir la salida del sistema. La calidad de las traducciones del sistema *Matxin* ha podido ser comparada con las de un sistema basado en corpus, obteniendo el segundo unos resultados significativamente peores. Debido al uso generalizado de BLEU, hemos querido estudiar los resultados BLEU conseguidos por ambos sistemas, constatando que esta métrica no es efectiva ni para medir la calidad absoluta de un sistema, ni para comparar sistemas que usan estrategias diferentes.
Palabras clave: Traducción automática basada en reglas, evaluación, HTER, BLEU

Abstract: *Matxin* is a rule-based machine translation system which translates to Basque. For its evaluation we have used the HTER metric which calculates the post-editing cost, concluding that 4 of each 10 words would have to be modified to correct the output generated by the system. We have compared the quality of *Matxin* translations with that of a corpus based system, and the results show that *Matxin* performs significantly better. Given the widespread use of BLEU, we have examined the BLEU scores for both systems, and we conclude that this metric is neither effective to measure the absolute quality of a system, nor suitable to compare systems based on different strategies.
Keywords: Rule-based machine translation, evaluation, HTER, BLEU

1. Introducción

En este artículo presentamos la evaluación del sistema de traducción automática (TA) basado en reglas *Matxin* (Mayor, 2007; Alegria et al., 2008) que traduce de español a euskera. Nuestro objetivo es proporcionar tanto una evaluación absoluta de la calidad de las traducciones del sistema, como una evaluación relativa, para lo que hemos podido comparar los resultados del sistema *Matxin* con los del sistema basado en corpus *Matrex* (Way et al., 2006; Labaka et al., 2007).

Para la evaluación hemos usado la métrica HTER (Snover et al., 2006; Przybocki, Sanders, y Le, 2006), que requiere que un editor humano postedite las traducciones del sistema de TA. El cálculo de esta métrica tiene un coste (que en el caso de evaluar uno o unos pocos sistemas de traducción es asumible) y, a cambio, nos proporciona una medida realista de la calidad de las traducciones del sistema, mostrándonos hasta qué punto son válidas.

Debido al uso (y a la exigencia de uso) generalizado de BLEU para la evaluación de sistemas de TA, hemos querido estudiar los resultados de esta métrica para los sistemas *Matxin* y *Matrex*, aún sabiendo que el uso de BLEU no es adecuado ni para evaluar la calidad absoluta de las traducciones, ni para comparar sistemas que usan estrategias de traducción diferentes. Esto nos dará la posibilidad de contrastar los resultados obtenidos con ambas métricas, pudiendo así sacar conclusiones al respecto.

En el apartado 2 presentamos el sistema *Matxin*. En el apartado 3 hacemos una introducción sobre los métodos de evaluación

de sistemas de TA, deteniéndonos en las dos métricas que vamos a usar: BLEU y HTER. El diseño de la evaluación y los resultados se muestran en el apartado 4. Finalizamos el artículo con las conclusiones derivadas de nuestros experimentos.

2. Matxin

Matxin es el primer sistema de TA públicamente disponible que traduce a euskera. Es un sistema basado en reglas, que sigue el modelo tradicional de transferencia.

Si bien en la última década la tendencia en el campo de la TA ha sido usar estrategias basadas en corpus, las estrategias tradicionales han de ser reconsideradas para poder hacer frente a las dificultades inherentes a los proyectos que trabajan con *lenguas no centrales* (Streiter, Scannell, y Stuflesser, 2006). De hecho, los sistemas estadísticos encuentran grandes dificultades para traducir a euskera. Por una parte, se necesitan corpus enormes para conseguir resultados aceptables, siendo muy limitados los corpus que hay para el euskera, y, por otra, como han demostrado Koehn y Monz (2006), los sistemas estadísticos obtienen peores resultados que los basados en reglas al traducir a lenguas con una morfología rica, como es el caso del euskera.

El prototipo *Matxin1.0* que traduce de español a euskera, puede usarse en Internet[1] y se distribuye como software de código abierto[2]. A día de hoy el sistema está siendo adaptado para traducir de inglés a euskera.

La traducción automática de español (o inglés) a euskera es una tarea muy compleja debido a que son lenguas tipológicamente muy lejanas, y con diferencias sintácticas muy grandes.

3. Métodos de evaluación

En el campo de la traducción automática la evaluación puede tener dos objetivos (Goutte, 2006): la evaluación absoluta, que da una medida total del comportamiento del sistema; y la evaluación relativa, que permite comparar diferentes sistemas de TA.

En cualquier caso, como apunta Koehn (2007), la pregunta correcta quizá no sea cuánto de buena es la traducción automática, sino cuán utilizable es.

A la hora de elegir un método de evaluación nos encontramos ante el dilema planteado por Eisele (2006): la evaluación manual es significativa, pero también cara y no reutilizable; la evaluación automática es rápida, repetible y objetiva[3], pero no se puede garantizar que sus resultados sean correctos.

Las medidas usadas tradicionalmente para evaluación manual han sido la fidelidad (*adequacy*), que mide si la traducción tiene el mismo significado que el texto original; y la fluidez (*fluency*), que mide si la traducción es gramaticalmente correcta o no. Estas medidas, aparte de ser caras, no son lo suficientemente concretas para medir el progreso de un sistema y no dan apenas información sobre lo que puede estar mal. Una alternativa puede ser la medida HTER (*Human-targeted Translation Edit Rate*) presentada en (Snover et al., 2006) y también llamada distancia de edición (Przybocki, Sanders, y Le, 2006), que calcula el coste de postedición de la traducción dada por el sistema.

Para juzgar la calidad de la TA de modo automático, se compara la traducción del sistema con traducciones humanas de referencia, asumiendo la hipótesis de que cuanto más parecidas sean, mejor será la calidad de la traducción. Una sola traducción de referencia no es suficiente ya que puede haber otra traducción correcta (u otras) que sea muy diferente a ésta. Por eso, la solución es usar un conjunto de traducciones de referencia (Popescu-Belis, King, y Benantar, 2002), si bien en la mayoría de los casos sólo se tiene disponible una única traducción (Giménez y Amigó, 2006). La proximidad entre la traducción del sistema y las de referencia se puede calcular tanto basándose en la correspondencia de cadenas (*string matching*), como en las métricas WER (Nießen et al., 2000), PER (Leusch, Ueffing, y Ney, 2003) o TER (Snover et al., 2006), o basándose en n-gramas, como es el caso de las métricas NIST (Doddington, 2002), WNM (Babych, 2004), F-measure (Melamed, Green, y J.P.Turian, 2003), Meteor (Lavie, Sagae, y Jayaraman, 2004) o BLEU (Papineni et al., 2002), que se ha convertido en la medida de evaluación de sistemas de TA más usada hoy en día.

[1] http://www.opentrad.org

[2] http://www.matxin.sourceforge.net

[3] En este contexto, que una evaluación sea *objetiva* significa que una misma traducción siempre tendrá la misma puntuación.

3.1. BLEU

BLEU calcula la media geométrica de la precisión de los n-gramas (n=1..4) multiplicada por una penalización de brevedad. La precisión de los n-gramas se calcula dividiendo el número de n-gramas de la traducción del sistema que aparecen en alguna de las traducciones de referencia entre el número de palabras de la traducción del sistema.

(Papineni et al., 2002; Doddington, 2002) afirman que BLEU es un método rápido, barato, independiente del lenguaje, y que tiene una gran correlación con las evaluaciones manuales.

Esta métrica ha guiado el progreso en el desarrollo de los sistemas estadísticos de traducción automática, puesto que la evaluación de los cambios incrementales del sistema y la optimización de los parámetros se hacen basándose en los resultados BLEU. Ha sido a su vez la medida elegida para comparar diferentes sistemas de TA en campañas de evaluación como las organizadas por la organización NIST (Lee y Przybocki, 2005).

A la vez que su uso se ha generalizado, han ido surgiendo serias dudas en torno a BLEU y el resto de métricas basadas en n-gramas. Por un lado, es difícil interpretar lo que expresa un resultado BLEU: *What does a Bleu score of 0,016 mean?* (Turian, Shen, y Melamed, 2003). Por otro lado, Callison-Burch, Osborne, y Koehn (2006) han demostrado que en determinadas condiciones una mejora de BLEU *no es suficiente* para reflejar una mejora en la calidad de la traducción, y que *no es necesario* mejorar BLEU para conseguir una mejora notable en la calidad de la traducción.

Y es que, siendo cierto que podemos afirmar que la traducción del sistema es mejor cuanto más se parezca a las traducciones de referencia, en principio no podemos afirmar que la calidad sea peor si se parece menos a las referencias, a no ser que dispongamos de todas las traducciones correctas posibles (Gispert Ramis, 2006).

(Callison-Burch, Osborne, y Koehn, 2006; Koehn y Monz, 2006) han demostrado que BLEU no tiene una correlación tan alta como se cree. Por ejemplo, de los datos de la campaña de evaluación NIST 2005 (Lee y Przybocki, 2005), se puede subrayar que el sistema que quedaba 1º en la evaluación manual, se clasificó como 6º con BLEU, y que, en general, los sistemas estadísticos reciben una puntuación BLEU más alta, castigándose a los sistemas basados en reglas con puntuaciones muy bajas.

Además, Homola, Kubon, y Pecina (2009) señalan que los resultados de BLEU son aún más inadecuados para lenguas con una morfología más rica y para aquellas que tienen un mayor grado de libertad en el orden de las palabras.

(Boitet et al., 2006) afirman que las métricas basadas en n-gramas, como BLEU, son inadecuadas, porque no miden la calidad de la traducción sino su parecido con las traducciones de referencia; y caras, porque el coste de preparar las traducciones de referencia (que han de ser varias) es muy alto. Algunos autores (Hamon y Rajman, 2006) consideran que al necesitar traducciones de referencia creadas manualmente, estas métricas no se pueden considerar automáticas, sino semi-automáticas.

Por todo ello, como subrayan Callison-Burch, Osborne, y Koehn (2006), resulta imprescindible distinguir qué usos de BLEU, y del resto de métodos de evaluación basados en n-gramas, son adecuados y cuáles no:

- Usos adecuados:
  - Seguimiento de los cambios incrementales de un sistema
  - Comparación de sistemas que usen estrategias similares
  - Optimización de los valores de los parámetros de sistemas estadísticos
- Usos inadecuados:
  - Comparación de sistemas que usen estrategias diferentes
  - Comparación de sistemas cuando el par de lenguas, el número de referencias o el tamaño de n-gramas es diferente
  - Identificación de mejoras de aspectos de la traducción que la métrica no modela bien
  - Monitorización de mejoras que aparecen poco en el corpus de test

A pesar de todas estas investigaciones críticas con el uso inadecuado de BLEU, no ha cambiado mucho la situación de excesiva confianza hacia BLEU que se describía en (Callison-Burch, Osborne, y Koehn, 2006):

los artículos de los congresos presentan mejoras en la calidad de sistemas de TA, dando únicamente mejores resultados de BLEU, o de métricas similares, sin mostrar ni un solo ejemplo de traducción; o se comparan sistemas usando BLEU sin contrastar con evaluaciones manuales.

De hecho, una y otra vez nos encontramos con que en la revisión de artículos para su aceptación en congresos se está pidiendo que se incluyan resultados BLEU, en situaciones en las que su uso es totalmente inadecuado.

3.2. HTER

Para calcular HTER, un editor humano modifica la traducción del sistema de TA de manera que la versión editada tenga todo el significado del texto de origen y esté escrita de manera entendible, realizando el mínimo número de modificaciones posible (Przybocki, Sanders, y Le, 2006). Las modificaciones posibles son:

- inserción, borrado y sustitución de palabras
- movimiento de grupos de palabras

El resultado HTER se obtiene dividiendo el número de modificaciones entre el número de *tokens* de la traducción editada[4].

Esta métrica mide la calidad de las traducciones del sistema de manera realista, mostrándonos, y esto es de gran importancia, hasta qué punto son válidas. Es decir, da la medida de cuánto trabajo de postedición se requiere para que las traducciones del sistema puedan ser consideradas correctas.

En la evaluación de un sistema de TA esta información es fundamental si queremos saber cuánto de utilizables son sus traducciones para ser publicadas, es decir, para saber si es más barato encargar la traducción a un traductor humano o posteditar la salida del sistema de TA.

Los experimentos presentados en (Snover et al., 2006; Przybocki, Sanders, y Le, 2006) demuestran que la métrica HTER tiene mejor correlación que BLEU con los juicios humanos de fidelidad y fluidez. Es más, HTER ha demostrado ser más consistente y de más detalle que las anotaciones humanas de fidelidad y fluidez.

[4]El software desarrollado en (Snover et al., 2006) para el cálculo automatizado de los resultados HTER se encuentra públicamente disponible en: http://www.cs.umd.edu/∼snover/tercom

El principal problema que surge con HTER es su coste, que se calcula entre 550 y 800 palabras/hora. Eso sí, para evaluar el progreso de un sistema no sería necesario posteditar todas las traducciones en cada evaluación, puesto que muchas de las traducciones no cambiarían. En el caso de evaluar un sistema de diseminación, cuya salida siempre se postedita para su publicación, calcular el coste de postedición es automático. En las campañas de evaluación GALE de la organización NIST (Przybocki, Sanders, y Le, 2006) se estudió el posible uso de HTER para hacer frente a las limitaciones de los métodos basados en n-gramas, pero en ese contexto, al tener que evaluar muchos sistemas, el coste sí que es muy grande, siendo seguramente esa la razón por la que ha sido descartado HTER en esas campañas.

El trabajo de postedición para calcular HTER, por razones prácticas, ha venido siendo realizado por editores monolingües, usando una traducción de referencia en vez del texto original, pero, como señala Font Llitjós (2006), eso no garantiza que el significado completo del texto original se mantenga. Por ello sería más adecuado contar con editores bilingües para realizar el trabajo de postedición.

4. Evaluación del sistema Matxin

Para realizar la evaluación se han usado dos corpus diferentes: *Eitb*, corpus periodístico general, que recoge las noticias de la radio y televisión vasca EITB; y *Consumer*, corpus sobre consumo que recoge los artículos publicados en la revista *Consumer* de la empresa Eroski (Alcázar, 2006).

Como afirma Koehn (2007) la evaluación de las traducciones de oraciones largas puede resultar muy complicada, puesto que los sistema de TA generan traducciones muy confusas y con errores en diferentes partes. Por ello hemos optado por evaluar oraciones de entre 5 y 25 palabras.

Para la evaluación con BLEU de cada uno de los corpus se han usado 1.500 oraciones de entre 5 y 25 palabras elegidas al azar, de las cuales se tiene una sola traducción de referencia.

El cálculo de los resultados HTER se ha realizado posteditando de cada uno de los corpus 50 oraciones de entre 5 y 25 palabras elegidas al azar. El trabajo de postedición ha sido realizado por un único editor bilingüe

			Ins	Bor	Sus	Mov	TM	Ed	Tok	HTER	
Matxin	Normal	*Eitb*	20	13	147	35	47	215	532	40,41 %	%42,00
		Cons	27	30	152	50	56	259	594	43,60 %	
	Segmentado	*Eitb*	37	21	136	60	93	254	727	34,94 %	%37,06
		Cons	50	50	149	66	87	315	804	39,18 %	
Matrex	Normal	*Eitb*	35	101	205	27	28	368	512	71,87 %	%64,92
		Cons	47	55	187	60	71	349	602	57,97 %	
	Segmentado	*Eitb*	32	173	289	60	68	554	726	76,31 %	%62,96
		Cons	46	112	145	84	113	387	780	49,61 %	

Cuadro 1: Resultados HTER

y las instrucciones de postedición se basan en las del programa GALE[5], que han sido adaptadas, sobre todo, por el hecho de que en nuestro caso el trabajo de postedición es bilingüe. Las normas básicas para la postedición señalan que la traducción editada ha de mantener el *mismo significado* que el texto de origen, ser *comprensible* y ser *gramaticalmente correcta*, todo ello realizando el *mínimo número de modificaciones* posible.

El cálculo de los resultados HTER contabiliza como *token* cada palabra y signo de puntuación. Para nuestros experimentos hemos calculado también los valores HTER segmentando las palabras tanto de la traducción del sistema como de la posteditada, de modo que se contabilizan como *token* los lemas, las posposiciones o casos gramaticales, y los signos de puntuación. De esta manera conseguimos un resultado más informativo, puesto que una palabra traducida en euskera puede contener errores tanto en la elección del lema, como en la asignación de la posposición o caso gramatical (que corresponde a una preposición en español o a una función sintáctica, además de contener información de determinación y número). Los resultados así calculados serían, de este modo, más adecuados para compararlos con los de sistemas que traducen a idiomas no aglutinativos, como el español o el inglés.

Además de la evaluación absoluta de la calidad del sistema *Matxin*, hemos podido comparar sus resultados con los obtenidos en la evaluación del sistema *Matrex*. El sistema basado en corpus *Matrex*, desarrollado en Dublin, ha sido adaptado para traducir de español a euskera usando herramientas de procesamiento del euskera desarrolladas en el grupo IXA, consiguiendo mejores resultados que un sistema de TA estadístico estándar (Way et al., 2006; Labaka et al., 2007). El sistema fue entrenado con 50.000 oraciones (975.000 palabras en español y 785.000 en euskera) del corpus de la revista *Consumer* (por supuesto, diferentes de las usadas para la evaluación).

La evaluación, tanto usando BLEU como HTER, del sistema *Matrex* ha sido realizada en las mismas condiciones que las del sistema *Matxin*, y usando el mismo corpus de test.

4.1. Resultados HTER

El cuadro 1 muestra los resultados HTER conseguidos por los dos sistemas, tanto el cálculo normal como el realizado sobre los textos segmentados, para las 50 oraciones de cada corpus. El resultado HTER se calcula dividiendo el número de ediciones (*Ed*) entre el número de *tokens* (*Tok*). El número de ediciones es la suma de todos los tipos de modificaciones: inserción (*Ins*), borrado (*Bor*), sustitución (*Sus*) y movimiento de grupo de palabras (*Mov*). En la tabla también aparece el número de *tokens* movidos (*TM*).

El sistema *Matxin* obtiene para el corpus *Eitb* un resultado HTER normal de 40,41 %, es decir, de cada 100 palabras 40 han de ser editadas. Para el corpus *Consumer*, el valor es un poco mayor, 43,60 %. El cálculo HTER segmentado da unos resultados mejores (34,94 % y 39,18 %) puesto que, a veces, el lema de la palabra es erróneo y la posposición o caso gramatical es correcta, o viceversa. Podemos observar que las ediciones más frecuentes, con mucha diferencia, son las sustituciones, seguidas muy de lejos por los movimientos.

El sistema *Matrex*, en cambio, consigue una puntuación HTER normal de 57,97 % para el corpus *Consumer*, y de 71,87 % para el corpus *Eitb*, resultando lógico que la puntuación para el corpus *Consumer* sea mejor, puesto que es el corpus con el que se ha entrenado el sistema. Si bien el cálculo HTER segmentado da unos resultados mejores (49,61 %) para el corpus *Consumer*, no

[5]http://projects.ldc.upenn.edu/gale/Translation/Editors/GALEpostedit_guidelines-3.0.2.pdf

sucede así para el corpus *Eitb* (76,31 %), donde los resultados son peores porque, a veces, tanto el lema de la palabra como la posposición o caso gramatical son incorrectos, con lo que aumenta en proporción el número de errores cometidos.

Para el sistema *Matrex* las ediciones más realizadas son la sustitución y el borrado. Este dato coincide con los presentados en (Snover et al., 2006), donde se evaluaban sistemas de TA estadísticos. El hecho de que el borrado de palabras sea una operación frecuente en la postedición de sistemas de TA estadísticos y no así en la postedición de las traducciones de *Matxin* nos hace suponer que, en general, en las traducciones de sistemas de TA estadística aparecen muchas más palabras sobrantes que en las de los sistemas basados en reglas.

Si comparamos los resultados HTER normales de ambos sistemas podemos observar que para el corpus *Consumer* la calidad de las traducciones de *Matxin* es mejor que la de *Matrex* (43,60 % vs 57,97 %), y que para el corpus *Eitb*, en el cual *Matrex* no ha sido entrenado, la diferencia es todavía mayor (40,41 % vs 71,87 %). Si comparamos los resultados HTER sobre las traducciones segmentadas, la diferencia para el corpus *Consumer* no es tan grande (39,18 % vs 49,61 %), pero para el corpus *Eitb* la diferencia llega a más del doble (34,94 % vs 76,31 %)

La primera conclusión, por lo tanto, es que el sistema basado en reglas *Matxin* es significativamente mejor que el sistema basado en corpus *Matrex*.

Si bien los resultados de *Matrex* no son nada buenos (para el corpus *Eitb* de cada 10 palabras 7 han de ser corregidas), eventualmente podrían mejorarse entrenando el sistema con un corpus más grande y haciendo ajustes. Hay que añadir que también se podría hacer un trabajo de ajuste de *Matxin* para un corpus concreto, mejorando sus resultados.

4.2. Resultados BLEU

En el cuadro 2 podemos ver un resumen de los resultados de la evaluación, tanto del cálculo HTER sin segmentar, como del de BLEU. Así como para HTER valores más pequeños indican una mejor calidad, para BLEU valores más pequeños deberían señalar una peor calidad de la salida del sistema.

Mirando los resultado BLEU difícilmente podemos sacar ninguna conclusión sobre la calidad absoluta de las traducciones obtenidas por cada sistema.

	HTER		BLEU	
	Matxin	*Matrex*	*Matxin*	*Matrex*
Eitb	40,41 %	71,87 %	9,30	9,02
Consumer	43,60 %	57,97 %	6,31	8,03

Cuadro 2: HTER vs BLEU

Matxin obtiene 9,30 puntos para el corpus *Eitb*, y un resultado peor de 6,31 para el corpus *Consumer*.

Los resultados para *Matrex* son, sorprendentemente, mejores para el corpus *Eitb* (9,02) que para el corpus *Consumer* con el que se ha entrenado el sistema (8,03).

Si comparamos los resultado BLEU de ambos sistemas, vemos que mientras que para el corpus *Eitb* los resultados de *Matxin* son un poco mejores que los de *Matrex* (9,30 vs 9,02), para el corpus *Consumer* los resultados de *Matxin* son peores (6,31 vs 8,03). Si quisiésemos sacar de estos resultados alguna conclusión sobre la calidad de las traducciones realizadas con estos sistemas, diríamos que *Matrex* consigue una calidad de traducción mejor que *Matxin* para textos del corpus *Consumer* con el que se ha entrenado. Pero los resultados HTER basados en la postedición manual invalidan totalmente esta conclusión.

Se puede criticar el uso de una sola referencia para el cálculo de BLEU y, de hecho, estamos de acuerdo con ello, pero si quisiésemos evaluar con más referencias, habría que crearlas manualmente, y esto sería muchísimo más caro que el trabajo de postedición necesario para calcular HTER. Además, Callison-Burch, Osborne, y Koehn (2006) han demostrado que puede suceder lo mismo aún usando múltiples referencias.

5. Conclusiones y trabajo futuro

Hemos evaluado el sistema de TA *Matxin 1.0* que traduce de español a euskera y, simplificando los resultados, podemos decir que un editor necesitaría cambiar 4 de cada 10 palabras para corregir la salida del sistema. Estos resultados demuestran que el sistema tiene mucho que mejorar y que aún está muy lejos de poder usarse como sistema para diseminación de información, es decir, que sea rentable usar el sistema posteditando sus traducciones para ser publicadas.

Para valorar estos resultados hemos de te-

ner en cuenta que traducir de español a euskera es una tarea compleja. Esto explica también los pobres resultados obtenidos por el sistema de TA basado en corpus *Matrex* traduciendo a euskera.

El uso de HTER nos ha proporcionado una evaluación, a nuestro entender, fiable sobre la calidad de los sistemas de TA presentados. Tanto una evaluación absoluta, para saber hasta qué punto pueden servir las traducciones realizadas, como una evaluación relativa que nos permite comparar entre diferentes sistemas. El coste del trabajo de postedición necesario no ha sido excesivo. Para postreditar 200 oraciones (50 de cada corpus por cada sistema) se han necesitado menos de 7 horas, con una velocidad media de edición de 2 minutos/oración (350 palabras/hora).

BLEU es una herramienta de gran valor, ya que es esencial en la construcción de sistemas de TA estadística. Pero, como hemos podido volver a constatar en nuestros experimentos, BLEU no ofrece ninguna información sobre la calidad absoluta de un sistema de TA, no muestra en qué medida son utilizables las traducciones realizadas, y además tampoco sirve para comparar sistemas que usan estrategias diferentes.

Lo que nos llama poderosamente la atención es cómo, a pesar de que las limitaciones de BLEU han sido puestas al descubierto y se han señalado claramente cuáles son sus usos adecuados y cuáles no, hoy en día seguimos viendo que se usa BLEU (y se exige su uso) en situaciones para las que no es apropiado.

Estos usos parecen basarse en la siguiente premisa: *Usamos BLEU, suponiendo que nos sirve, porque es barato.* Y en ello se están cometiendo dos grandes errores, porque:

1. BLEU no sirve para lo que no sirve. Por ejemplo, para calibrar la calidad absoluta de un sistema, o para comparar sistemas de estrategias diferentes.

2. BLEU sí tiene coste. Para usar BLEU de manera *fiable* es necesario disponer de tres traducciones de referencia, y la mayoría de los corpus bilingües solo tienen una. Por ello, para poder usar BLEU hay que crear manualmente las referencias, y el trabajo que esto supone es muy caro.

Por ello, es urgente dejar a un lado los usos inapropiados de BLEU, usando esta métrica sólo cuando realmente sea adecuada, y analizando su coste.

En las evaluaciones futuras del sistema *Matxin* seguiremos usando el método de evaluación manual HTER. Como afirma Koehn (2007), es importante investigar para reducir el coste de las evaluaciones manuales, facilitando su realización y haciéndolas reutilizables. Con ese objetivo, estamos diseñando un entorno gráfico que haga más cómodo el trabajo de postedición y gestione las postediciones de evaluaciones anteriores de manera que sólo se tengan que posteditar las traducciones que cambian, reduciendo así el coste de manera considerable.

Queremos estudiar el uso de la mejora presentada en (Snover et al., 2009) para el cálculo de HTER, que incorpora el uso de morfología, sinónimos y paráfrasis, y ajusta los costes para diferentes tipos de errores, ya que la versión que hemos usado en nuestros experimentos puntúa igual todas las ediciones realizadas, no haciendo distinción entre los errores que son graves y los que no.

Asimismo, en futuras evaluaciones, nuestro objetivo es obtener datos más significativos aumentando el corpus de evaluación, y estudiar la consistencia del trabajo de postedición, usando más de una persona para la postedición del mismo corpus.

Otra línea de investigación atractiva será comparar cualitativamente las traducciones de *Matxin* y *Matrex*, para estudiar en qué acierta y en qué falla cada sistema, investigando así las estrategias de hibridación más eficientes.

Agradecimientos

Esta investigación ha recibido ayuda del Ministerio de Educación y Ciencia a través de los proyectos OpenMT: Open Source Machine Translation using hybrid methods (TIN2006-15307-C03-01) y Ricoterm-3 (HUM2007-65966-CO2-02). El corpus *Consumer* ha sido cedido por Asier Alcázar y la Fundación Eroski.

Bibliografía

Alcázar, A. 2006. Towards linguistically searchable text. En *Proceedings of BIDE 2005*, Deusto. Bilbao.

Alegria, I., X. Arregi, X. Artola, A. Díaz de Ilarraza, G. Labaka, M. Lersundi, A. Mayor, y K. Sarasola. 2008. Strate-

gies for sustainable MT for basque: incremental design, reusability, standardization and open-source. En *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, páginas 59–64, Hyderabad, India.

Babych, B. 2004. Weighted N-gram model for evaluating Machine Translation output. En *Proceedings of the CLUK '04. Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*, páginas 15–22, Birmingham, UK.

Boitet, C., Y. Bey, M. Tomokiyo, C. Cao, y H. Blanchon. 2006. IWSLT-06: Experiments with commercial MT systems and lessons from subjective evaluations. En *Proceedings of the International Workshop on Spoken Language Translation, IWSLT-06*, páginas 23–30, Kyoto, Japan.

Callison-Burch, C., M. Osborne, y P. Koehn. 2006. Re-evaluating the role of BLEU in Machine Translation research. En *Proceedings of EACL-2006*.

Doddington, G. 2002. Automatic evaluation of Machine Translation quality using n-gram co-occurrence statistics. En *Proceedings of the HLT 2002*, páginas 138–145, San Diego, California.

Eisele, A. 2006. Improving Machine Translation quality via hybrid sistems and refined evaluation methods. Presentation in November 2006 at IST Event 2006, Workshops and Networking Sessions, Multilinguism and Language Technology a Challenge for Europe, Helsinki.

Font Llitjós, A. 2006. Giving the power to bilingual speakers. En *Automated Post-Editing Workshop at AMTA*.

Giménez, J. y E. Amigó. 2006. IQMT: a framework for automatic Machine Translation evaluation. En *Proceedings of the Fifth LREC*, páginas 685–690, Genoa, Italy.

Gispert Ramis, A. 2006. *Introducing Linguistic Knowledge into Statistical Machine Translation*. Ph.D. tesis, Universitat Politécnica de Catalunya.

Goutte, C. 2006. Automatic evaluation of Machine Translation quality. Presentation at the European Community.

Hamon, O. y M. Rajman. 2006. X-score: automatic evaluation of Machine Translation grammaticality. En *Proceedings of the Fifth LREC*, páginas 155–160, Genoa, Italy.

Homola, P., V. Kubon, y P. Pecina. 2009. A simple automatic mt evaluation metric. En *Proceedings of the Fourth Workshop on SMT. EACL.*

Koehn, P. 2007. Evaluating evaluation - Lessons from the WMT 2007 shared task. En *Proceedings of the MT Summit Workshop on MT Evaluation.*

Koehn, P. y C. Monz. 2006. Manual and automatic evaluation of Machine Translation between European languages. En *Proceedings on the Workshop on SMT*, páginas 102–121, New York City, June. ACL.

Labaka, G., N. Stroppa, A. Way, y K. Sarasola. 2007. Comparing rule-based and data-driven approaches to Spanish-to-Basque machine translation. En *Proceedings of the MT-Summit XI*, Copenhagen.

Lavie, A., K. Sagae, y S. Jayaraman. 2004. The significance of recall in automatic metrics for MT evaluation. En *Proceedings of the 6th conference of the AMTA*, páginas 134–143, Washington.

Lee, A. y M. Przybocki. 2005. NIST 2005 Machine Translation evaluation official results. Informe técnico, NIST.

Leusch, G., N. Ueffing, y H. Ney. 2003. A novel string-to-string distance measure with applications to Machine Translation evaluation. En *Proceedings of the MT Summit IX*, páginas 240–247, New Orleans, USA.

Mayor, A. 2007. *Matxin: Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz*. Ph.D. tesis, University of the Basque Country, Donostia, Euskal Herria.

Melamed, I.D., R. Green, y J.P.Turian. 2003. Precision and recall of Machine Translation. En *Proceedings of the HLT-NAACL 2003: Conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, Edmonton, Canada.

Nießen, S., F.J. Och, G. Leusch, y H. Ney. 2000. An evaluation tool for Machine Translation: fast evaluation for MT research. En *Proceedings of the Second LREC*, páginas 39–45, Athens, Greece.

Papineni, K., S. Roukos, T. Ward, y W. Zhu. 2002. BLEU: a method for automatic evaluation of Machine Translation. En *Proceedings of the 40th Annual Meeting of the ACL*, páginas 311–318.

Popescu-Belis, A., M. King, y H. Benantar. 2002. Towards a corpus of corrected human translations. En *Proceedings of the Workshop: MT evaluation, human evaluators meet automated metrics. Third LREC*, páginas 17–21.

Przybocki, M., G. Sanders, y A. Le. 2006. Edit distance: a metric for Machine Translation evaluation. En *Proceedings of the Fifth LREC*, páginas 2038–2043, Genoa, Italy.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, y J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. En *Proceedings of the AMTA*.

Snover, M., N. Madnani, B. Dorr, y R. Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. En *Proceedings of the Fourth Workshop on SMT. EACL.*

Streiter, O., K.P. Scannell, y M. Stuflesser. 2006. Implementing NLP projects for noncentral languages: instructions for funding bodies, strategies for developers. *Machine Translation*, 20(4):267–289.

Turian, J. P., L. Shen, y I.D. Melamed. 2003. Evaluation of Machine Translation and its evaluation. En *Proceedings of the MT Summit IX*, New Orleans, USA.

Way, A., N. Stroppa, K.Sarasola, y D. Groves. 2006. Example-Based Machine Translation of the Basque language. En *Proceedings of the 4th LREC*, Boston, USA.