Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 215-222 recibido 1-05-2009; aceptado 5-06-2009

Procesamiento Lingüístico en Métricas de Evaluación Automática de Traducciones

Linguistic Processing in Automatic Translation Evaluation Metrics

Enrique Amigó
UNED
Madrid
enrique@lsi.uned.es

Jesús Gimenez
UPC
Barcelona
jgimenez@lsi.upc.edu

Felisa Verdejo
UNED
Madrid
felisa@lsi.uned.es

Resumen: A pesar de los esfuerzos por incluir procesamiento lingüístico en las métricas de evaluación automática de sistemas de traducción, las más usadas siguen siendo métricas basadas en solapamiento léxico. Esto se debe a que no se han clarificado aún las ventajas del uso de técnicas lingüísticas en este contexto. En este artículo se analiza en profundidad las ventajas de aplicar procesamiento lingüístico a nivel sintáctico y semántico en la evaluación automática de traducciones.
Palabras clave: Evaluación, traducción automática, procesamiento lingüístico

Abstract: Despite the efforts for incorporating linguistic knowledge into automatic MT evaluation methods, metrics based on mere lexical matching are still today the dominant approach. This is partly due to the fact that the advantages of using deeper linguistic information have not been yet clarified. In this work, we analyze the advantages of including syntactic and semantic information in automatic MT evaluation.
Keywords: Machine Translation, Evaluation Metrics, Linguistic Processing

1. Introducción

Por lo general, las métricas de evaluación automática de traducciones comparan la salida del sistema con traducciones modelo realizadas por asesores. Estas métricas permiten evaluar de forma iterativa sin necesidad de nuevos juicios humanos, acelerando el desarrollo de los sistemas de traducción. Además, estas métricas pueden aplicarse directamente a otras tareas que conlleven generación de texto, como resumen automático, reducción de frases, generación de lenguaje o cierto tipo de búsqueda de respuestas.

Las métricas más empleadas se basan en solapamiento léxico entre la traducción evaluada y las traducciones modelo. En adelante nos referiremos a estas métricas como métricas léxicas. Medir la similitud entre una traducción automática y las traducciones modelo no es una tarea trivial, y considerar únicamente el solapamiento léxico puede resultar insuficiente. El siguiente ejemplo muestra dos traducciones árabe-inglés evaluadas en la competición promovida por el NIST en 2005. La métrica léxica ROUGE$_L$ (Lin y Och, 2004), es una de las más fiables en términos de correlación con juicios humanos según trabajos anteriores. Esta métrica asigna una mayor puntuación a la traducción B que a la traducción A, aunque la traducción B contiene un gran número de errores gramaticales. Ésto se debe a que la traducción B, aun siendo incorrecta, tiene un mayor solapamiento léxico con los modelos.

Traducción A: *The Chinese President made unprecedented criticism of the leaders of Hong Kong after political failings in the former British colony on Monday .* Adecuación + fluidez=8.5.

Traducción B: *Chinese President Hu Jintao today unprecedented criticism to the leaders of Hong Kong wake political and financial failure in the former British colony.* Adecuación + fluidez=3.

Con el fin de atacar este problema se han propuesto métricas que emplean procesamiento lingüístico a niveles superiores de procesamiento: sintáctico o semántico (Owczarzak et al., 2006; Reeder et al., 2001; Liu y Gildea, 2005; Amigó et al., 2006; Mehay y Brew, 2007; Giménez y Màrquez, 2007; Owczarzak, van Genabith, y Way, 2007; Popovic y Ney, 2007; Giménez y Màrquez,

ISSN: 1135-5948 © 2009 Sociedad Española para el Procesamiento del Lenguaje Natural

2009). En adelante nos referiremos a estas métricas como métricas lingüísticas. Sin embargo, las métricas léxicas siguen siendo las más empleadas para evaluar sistemas de traducción. Por tanto, es interesante clarificar la aportación del procesamiento lingüístico en la evaluación de traducciones, y es éste el objetivo que nos planteamos en este artículo.

Para ello, tenemos en cuenta dos aspectos no abordados en trabajos anteriores. En primer lugar, la correlación de las métricas frente a juicios humanos (meta-evaluación) se computa tradicionalmente sobre la calidad promedio del sistema en todas sus traducciones (nivel de sistema) o sobre cada una de las traducciones evaluadas (nivel de segmento). Ésto afecta en gran medida a los resultados de la meta-evaluación de métricas. En este artículo consideramos el número de segmentos sobre los que se promedia la calidad del sistema como una variable de estudio.

En segundo lugar, consideramos las métricas léxicas y lingüísticas en su conjunto. Para las primeras consideramos una selección de 16 métricas léxicas representativas del estado de la cuestión. Para las segundas, consideramos 48 métricas lingüísticas y la métrica combinada ULC (Giménez y Màrquez, 2008), que integra métricas de ambos tipos.

Los resultados muestran que la métrica combinada ULC es más fiable en términos de correlación con juicios humanos que las métricas léxicas individuales, especialmente cuando la calidad del sistema se promedia sobre un conjunto lo suficientemente grande de frases. Además, en ciertos casos supera incluso la cota máxima de fiabilidad a la que podrían llegar las métricas léxicas en conjunto.

En la siguiente sección se revisa el estado de la cuestión en cuanto a meta-evaluación de métricas. La sección 3 describe las métricas y conjuntos de prueba empleados en este artículo. La sección 4 analiza las diferencias fundamentales entre métricas léxicas y lingüísticas en términos de correlación con juicios humanos. En la sección 5 se presenta la aproximación empleada para comparar métricas en los experimentos descritos en la sección 6. En la sección 7 se presentan las conclusiones obtenidas a partir de este trabajo.

2. *Trabajos previos en meta-evaluación de métricas de traducción*

Los criterios de meta-evaluación han ido evolucionando a medida que se han ido proponiendo nuevas métricas de evaluación. Por ejemplo, la métrica BLEU (Papineni et al., 2001) es meta-evaluada en términos de correlación con juicios humanos, de forma independiente para *adecuación* y *fluidez*. Para abordar las deficiencias de BLEU se presenta la métrica NIST (Doddington, 2002). Análogamente, NIST se ha meta-evaluado en términos de correlación con juicios humanos pero esta vez, sobre diferentes textos fuente, para un número variable de modelos y sobre textos de diferente tamaño. Tras ésto, se presenta la familia de métricas GTM (Melamed, Green, y Turian, 2003) argumentando que la correlación Pearson empleada en trabajos anteriores puede verse afectada por fenómenos de escala. Por ello, los autores proponen usar el coeficiente de correlación no paramétrico Spearman (Spearman, 1904) o la Tau de Kendall (Kendall, 1938). Sin embargo, en otro trabajo se compara un conjunto representativo de métricas léxicas sin obtener diferencias importantes entre los coeficientes Pearson y Spearman (Lin y Och, 2004).

En el trabajo en el que se propone la métrica METEOR (Banerjee y Lavie, 2005) se identifica un aspecto que afecta en gran medida a la fiabilidad de las métricas. Tradicionalmente se computa la correlación con juicios humanos tomando como muestra la calidad de cada sistema tras promediar la calidad de todas sus traducciones. Los autores proponen sin embargo computar la correlación frase a frase. Es lo que se denomina correlación a nivel de segmento frente a correlación a nivel de sistema. En la competición de sistemas promovida por el NIST en 2008 (Przybocki, Peterson, y Bronsart, 2008) se plantea también un nivel intermedio, evaluación de documentos.

En una línea paralela, varios autores sugieren la combinación de métricas de evaluación (Kulesza y Shieber, 2004) o rasgos de las traducciones (Albrecht y Hwa, 2007) tomando como criterio de optimización la correlación con juicios humanos o la capacidad de discriminar traducciones modelo frente a traducciones automáticas. Sin embargo, en ninguno de estos trabajos se

realiza un estudio sobre la necesidad de incluir procesamiento lingüístico en el proceso de traducción.

3. Métricas y conjuntos de prueba

3.1. Métricas

Para la realización de los experimentos descritos en este artículo hemos considerado un conjunto amplio de métricas a tres niveles de procesamiento: léxico, sintáctico y semántico. A nivel léxico, hemos tomado métricas estandar basadas en distancia de edición (WER, PER and TER), precision léxica (BLEU y NIST), cobertura léxica (ROUGE), y medida F (GTM y METEOR). A nivel sintáctico hemos considerado diferentes familias de métricas basadas en procesamiento sintáctico superficial y analisis completo: relaciones de dependencia (DP) y árboles de constituyentes (CP). A nivel semántico hemos incluido métricas pertenecientes a tres familias que emplean respectivamente entidades nombradas (NE), roles semánticos (SR) y estructuras de discurso (DR). En Gimenez et al. (2007) se describe en detalle todas estas métricas.

Para analizar en conjunto la fiabilidad de las métricas lingüísticas hemos empleado la métrica combinada ULC (Giménez y Màrquez, 2008). Ésta es una aproximación simple que consiste en promediar aritméticamente y sin entrenamiento un subconjunto de métricas a diferentes niveles de procesamiento lingüístico. Además de haber dado buenos resultados en campañas de meta-evaluación recientes (Callison-Burch et al., 2008; Callison-Burch et al., 2009), su simplicidad asegura que sus resultados no están sesgados por el efecto de sobre-ajuste para un conjunto de prueba específico. El cómputo de las métricas se ha realizado mediante la herramienta IQ*MT* (Giménez, 2007)[1].

3.2. Conjuntos de prueba

Hemos empleado los conjuntos de prueba generados en las campañas de evaluación del NIST en 2004 y 2005 (Le y Przybocki, 2005)[2]. En estas campañas se llevaron a cabo dos tareas de traducción: árabe a inglés (AE) y chino a inglés (CE). Los juicios humanos dados por dos anotadores para cada traducción incluyen adecuación de las traducciones y fluidez escalados de uno a cinco.

	2004		2005	
	AE	**CE**	**AE**	**CE**
Modelos	5	5	5	4
Sistemas	5	10	5+1	5
Segmentos	347	447	266	272

Cuadro 1: Descripción de los conjuntos de prueba en las campañas de evaluación NIST 2004/2005

El cuadro 1 describe numéricamente estos conjuntos de prueba, incluyendo el número de traducciones modelo por segmento evaluado, el número de sistemas y el número de segmentos evaluados por sistema. El corpus AE05 incluye además un sistema de traducción asistido, lo que afectará de manera relevante a los resultados en nuestros experimentos.

4. Correlación a nivel de segmento versus sistema

Un primer paso para contrastar las métricas léxicas frente a métricas lingüísticas es calcular la correlación con juicios humanos de cada una de ellas. La figura 1 muestra la correlación obtenida para cada métrica a nivel de sistema (eje horizontal) y a nivel de segmento (eje vertical). La métricas linguïsticas están representadas por rombos grises. Las métricas léxicas están representadas por cuadrados negros. El círculo representa las correlaciones obtenidas para la métrica combinada ULC.

Cada conjunto de prueba (AE05, CE04, etc.) representa un escenario distinto, con sistemas de traducción de diferentes características y para diferentes lenguas. Para dar igual peso a cada uno de los escenarios, se ha promediado los valores de correlación de cada métrica sobre cada uno de los diferentes conjuntos de prueba.

El aspecto más importante de esta gráfica es que la fiabilidad relativa de las métricas depende en gran medida de si se mide a nivel de sistema o de segmento. De hecho, parece haber un relación inversa entre ambos tipos de correlación. Esta gráfica presenta una correlación Pearson negativa de 0,44. Es decir, una métrica puede ser más fiable que otra en términos de correlación a nivel de sistema e invertirse esta relación según la correlación a nivel de segmento.

[1] `http://www.lsi.upc.edu/~nlp/IQMT`

[2] `http://www.nist.gov/speech/tests/mt`

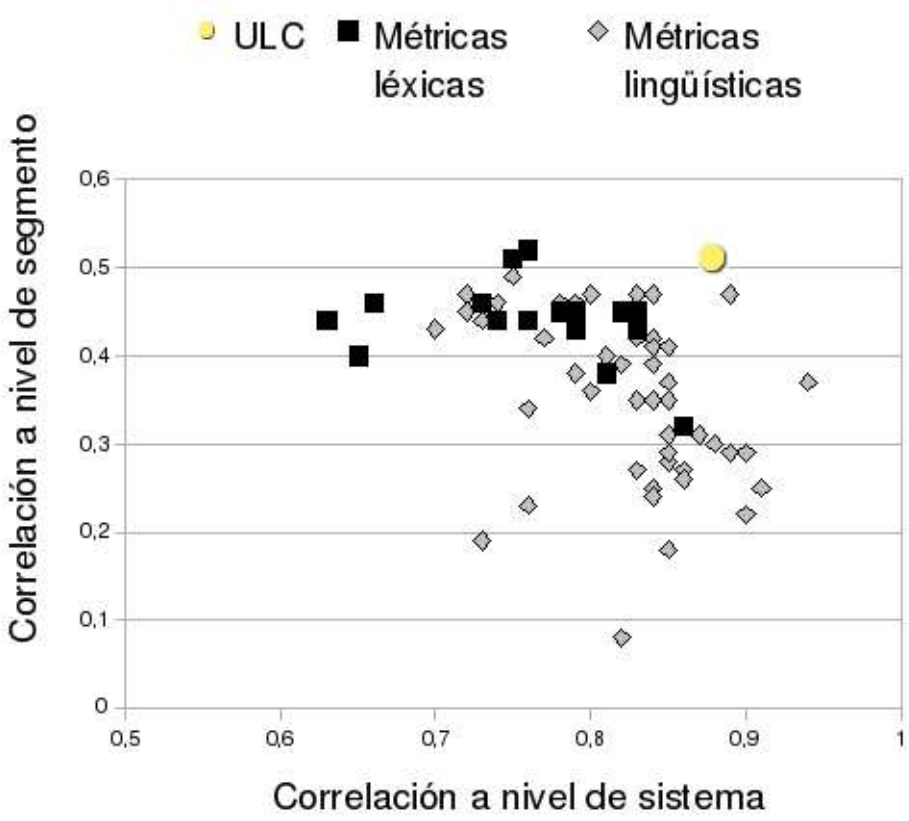

Figura 1: Correlación a nivel de sistema versus a nivel de segmento

En general puede verse que las métricas lingüísticas tienden a dar mejores resultados a nivel de sistema, aunque a la vista de la gráfica no podemos asegurar esta tendencia dada la dispersión de los datos. Sin embargo, lo que sí es cierto es que la métrica combinada ULC obtiene valores de correlación relativamente altos tanto a nivel de sistema como a nivel de frase. Ésto representa una primera evidencia a favor del uso de métricas lingüísticas en combinación con métricas léxicas.

Dado que el nivel al que se establece la correlación resulta tan determinante, en este trabajo planteamos el número de traducciones sobre las que se promedia la calidad como una variable de estudio. Por otro lado, es posible que combinando únicamente métricas léxicas pudiesen obtenerse valores de correlación similares a ULC. Es necesario comprobar por tanto hasta qué punto la combinación de métricas léxicas puede ser efectiva sin necesidad de aplicar métricas lingüísticas.

5. *PWA y MaxPWA*

Como muestran los resultados de la sección anterior y en concordancia con experimentos realizados en otros trabajos, la fiabilidad de las métricas varía según calculemos correlaciones a nivel de segmento o sistema. Para cuantificar este aspecto, consideraremos el número de traducciones sobre las que se promedia la calidad de los sistemas como una de las variables de estudio.

Además, la correlación con juicios humanos de métricas individuales no da información relativa al potencial de las métricas en su conjunto. Dicho de otro modo, ¿hasta donde podemos llegar combinando métricas léxicas? ¿Es estrictamente necesario introducir métricas lingüísticas? Una forma de abordar esta cuestión consiste en aplicar algún método concreto de aprendizaje automático para entrenar combinaciones de métricas. Sin embargo, los resultados de este análisis pueden estar sesgados por el algoritmo de aprendizaje empleado.

Para evitar este sesgo, extendemos para conjuntos de métricas la medida Pairwise Accuracy (PWA). Para una métrica de evaluación, PWA se define en términos de probabilidad como, dados dos sistemas o conjuntos de traducciones tales que uno supera a

otro según juicios humanos, la probabilidad de que la métrica automática confirme este resultado. Formalmente:

$$\text{PWA}(x) = \quad (1)$$
$$\text{Prob}(x(T_1) > x(T_2) | Q(T_1) > Q(T_2))$$

Siendo $x(T)$ la calidad de la traducción o conjunto de traducciones T según la métrica x y $Q(T)$ la calidad de T según juicios humanos.

En la competición de métricas de traducción promovida por el NIST en 2008 (Przybocki, Peterson, y Bronsart, 2008) se empleó una meta-métrica análoga (Pairwise System Comparison). En este caso se preguntaba explícitamente a los jueces humanos qué traducción preferían de entre dos traducciones. En nuestro caso, consideramos que una traducción es mejor que otra si obtiene mejor puntuación según jueces humanos.

Pero además, podemos extender esta medida para estimar la cota máxima de fiabilidad de un conjunto de métricas, con independencia del criterio de combinación empleado. Definimos la medida de máximo PWA bajo el supuesto de que:

- Si ninguna métrica asigna mayor calidad a T_1 que a T_2, entonces ningún criterio de combinación podría afirmar que T_1 es mejor que T_2.
- Si todas las métricas asignan mayor calidad a T_1 que a T_2, entonces cualquier criterio de combinación concluirá que T_1 es mejor que T_2.
- Si existen divergencias entre métricas, es decir para algunas métricas T_1 es mejor que T_2 y a la inversa para otras, entonces en el mejor caso (con el mejor criterio de pesado de métricas) la combinación de métricas llegará a la misma conclusión que los jueces humanos, dando mayor peso a la métrica más apropiada en cada caso.

Bajo estos supuestos podemos definir la cota superior de fiabilidad de un conjunto de métricas como:

$$\text{MaxPWA}(X) = \quad (2)$$
$$\text{Prob}(\exists x \in X.x(T_1) > x(T_2) | Q(T_1) > Q(T_2))$$

En general, esta medida no puede ser tomada como referencia para estimar la calidad de cualquier conjunto de métricas. Por ejemplo, un conjunto lo suficientemente grande de métricas aleatorias podría obtener un MaxPWA=1. De hecho, MaxPWA no es un método estandar para la selección de rasgos en el campo del aprendizaje automático. Sin embargo, dado que se trata de una cota superior, si se diera el caso de que una métrica individual superara el MaxPWA de otro conjunto de métricas que no incluyera a la primera, podemos asegurar que la primera métrica es más fiable que el conjunto aún aplicando el mejor criterio de combinación posible. Este es el caso de los resultados obtenidos en nuestros experimentos.

Finalmente, para PWA y MaxPWA podemos fijar el número de segmentos sobre los que se promedia la calidad del par de sistemas comparado en cada muestra. Es decir, siendo X un conjunto de métricas, T y T' dos conjuntos de traducciones generados por dos sistemas y n el número de segmentos considerado por muestra:

$$\text{MaxPWA}_n(X) = \quad (3)$$
$$\text{Prob}(\exists x \in X.x(T) > x(T') |$$
$$Q(T) > Q(T') \wedge |T| = |T'| = n)$$

6. Resultados

Las gráficas de la figura 2 muestran los resultados obtenidos tras aplicar PWA sobre cada una de las métricas léxicas (lineas grises) y sobre la métrica ULC (linea negra), y MaxPWA sobre el conjunto de métricas léxicas (línea negra discontinua). Cada gráfica corresponde a un conjunto de prueba distinto. El eje horizontal representa el número de segmentos sobre los que se promedia la calidad de los sistemas. Las conclusiones obtenidas se describen en los siguiente subapartados.

6.1. Competiciones en 2004 versus 2005

En general, el PWA de las métricas en las competiciones de 2005 es inferior al de las competiciones en 2004. Ésto puede deberse a varias razones. Por un lado, aunque el número de sistemas evaluados en 2005 no es mayor, sí puede ser mayor su heterogeneidad. Es decir, resulta más complejo comparar traducciones de sistemas que difieren en las técnicas de traducción empleadas. Por ejemplo, el conjunto de test AE2005 incluye un sistema asistido por el usuario.

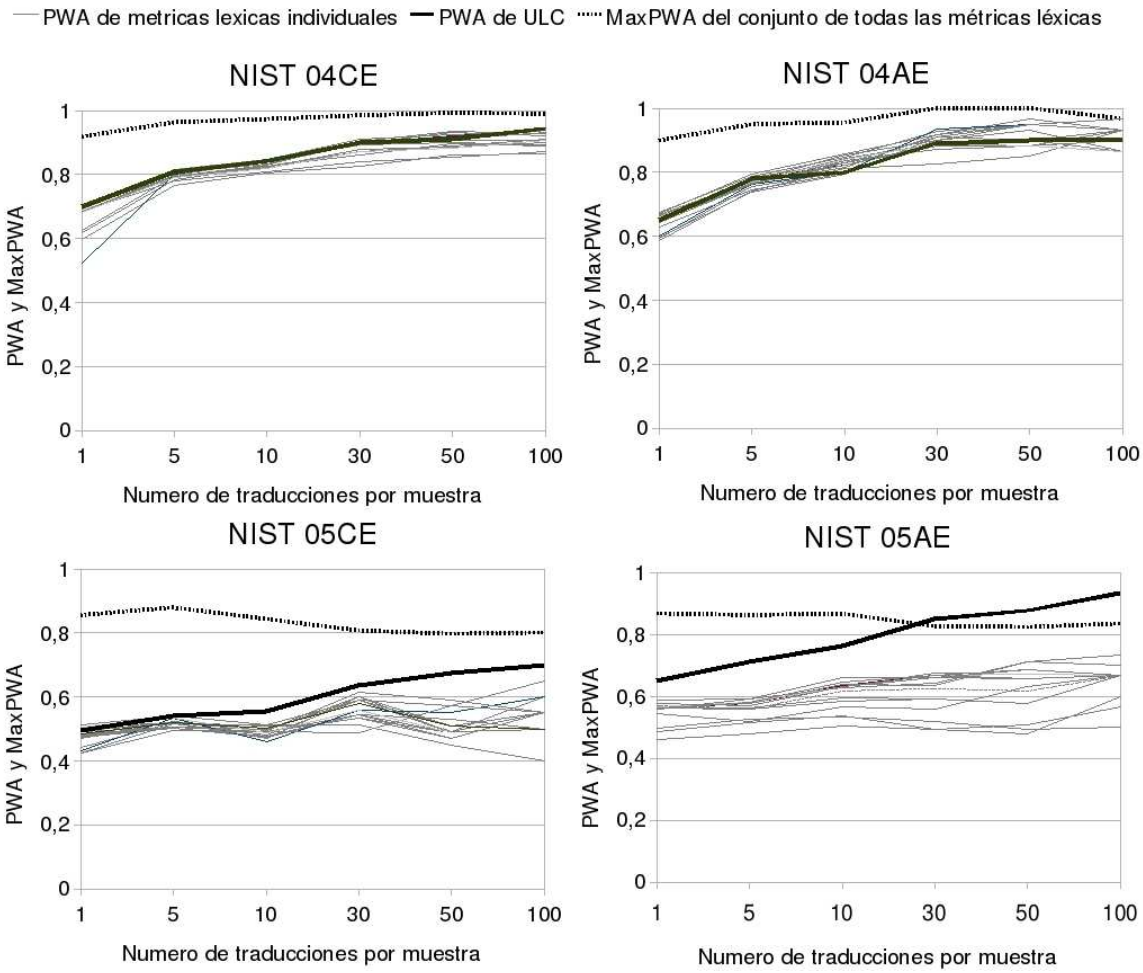

Figura 2: Resultados de PWA y MaxPWA

Otra razón importante, aunque difícil de demostrar, es que a partir de 2004 los desarrolladores emplearon métricas léxicas, fundamentalmente BLEU, para optimizar sus sistemas con vistas a la competición de 2005. Ésto hace que los sistemas mejoren en solapamiento de n-gramas, sin considerar otros aspectos de la calidad de los sistemas. Esta hipótesis concuerda con el hecho de que sin embargo, la métrica ULC que incluye métricas basadas en niveles lingüísticos superiores se vea menos afectada por el descenso de fiabilidad en los conjuntos de prueba de 2005.

6.2. Efecto del número de traducciones por muestra

El eje horizonal representa el número de traducciones sobre las que se promedia la calidad del sistema. Es decir, el número de traducciones por muestra de evaluación. En concordancia con la correlación a nivel de sistema frente a nivel de segmento, a medida que aumenta el número de traducciones la fiabilidad de las métricas aumenta, especialmente para el caso de la métrica combinada ULC. A la vista de los datos, no existe un número fijo de traducciones a partir del cual la fiabilidad de las métricas se estabilice, dado que depende de otros factores. En los conjuntos de prueba de 2004 la fiabilidad de las métricas parece estabilizarse a partir de 30 traducciones. En los conjuntos de 2005 no aparece un punto de inflexión.

Sin embargo, las gráficas muestran que las métricas lingüísticas incluidas en ULC explotan en mayor medida las ventajas de promediar la calidad de los sistemas sobre conjuntos de traducciones. Es decir, sin perder fiabilidad a nivel de segmento respecto a la métricas léxicas, incluir métricas lingüísticas consigue mejorar la fiabilidad sobre conjuntos de traducciones.

6.3. Combinación de métricas léxicas

La última cuestión es hasta qué punto la fiabilidad de las métricas léxicas puede aumentar combinándolas entre sí. Este aspecto puede analizarse mediante el MaxPWA de las métricas léxicas representado en las gráficas. Como se indicó anteriormente no podemos asegurar que un método de aprendizaje automático sea capaz de alcanzar las cotas de fiabilidad dadas por el MaxPWA. Sin embargo, el dato más relevante es que para el caso de NISTAE05, la métrica combinada ULC supera la mejor combinación posible de métricas léxicas cuando se promedia la calidad sobre más de 30 segmentos. Nótese que en NISTAE05 aparece el sistema asistido, lo que supone un reto para las métricas. Este resultado representa una evidencia clara de que las técnicas lingüísticas permiten superar barreras en el problema de la evaluación que no pueden ser superadas por las métricas léxicas.

7. Conclusiones

Los experimentos realizados en este trabajo muestran que, en términos de correlación con juicios humanos, combinar métricas léxicas con lingüísticas supera a las métricas léxicas individuales. Especialmente, ésto es así cuando la calidad de los sistemas es promediada sobre un número suficientemente grande de traducciones y cuando los sistemas evaluados son heterogéneos, como es el caso del sistema asistido evaluado en uno de los conjuntos de prueba estudiados. Además, una combinación no entrenada de métricas léxicas y lingüísticas supera en algunos la cota máxima teórica de las métricas léxicas en su conjunto. Estas conclusiones motivan, por un lado, el desarrollo de métricas a niveles superiores de procesamiento lingüístico y, por otro lado, la combinación de las mismas con métricas léxicas, ya sea mediante técnicas de aprendizaje automático (Kulesza y Shieber, 2004) o mediante modelos no entrenados (Amigó et al., 2005).

Agradecimientos

Este trabajo ha sido financiado parcialmente por la Comunidad de Madrid a través del proyecto MAVIR S0505-TIC0267.

Bibliografía

Albrecht, Joshua y Rebecca Hwa. 2007. A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation. En *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, páginas 880–887.

Amigó, Enrique, Jesús Giménez, Julio Gonzalo, y Lluís Màrquez. 2006. MT Evaluation: Human-Like vs. Human Acceptable. En *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, páginas 17–24.

Amigó, Enrique, Julio Gonzalo, Anselmo Pe nas, y Felisa Verdejo. 2005. QARLA: a Framework for the Evaluation of Automatic Summarization. En *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, páginas 280–289.

Banerjee, Satanjeev y Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. En *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, y Josh Schroeder. 2008. Further meta-evaluation of machine translation. En *Proceedings of the Third Workshop on Statistical Machine Translation*, páginas 70–106.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, y Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. En *Proceedings of the Fourth Workshop on Statistical Machine Translation*, páginas 1–28.

Doddington, George. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. En *Proceedings of the 2nd International Conference on Human Language Technology*, páginas 138–145.

Giménez, Jesús. 2007. IQMT v 2.0. Technical Manual (LSI-07-29-R). Informe técnico, TALP Research Center. LSI

Department. http://www.lsi.upc.edu/~nlp/IQMT/IQMT.v2.1.pdf.

Giménez, Jesús y Lluís Màrquez. 2007. Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. En *Proceedings of the ACL Workshop on Statistical Machine Translation*, páginas 256–264.

Giménez, Jesús y Lluís Màrquez. 2008. Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations. En *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP)*, páginas 319–326.

Giménez, Jesús y Lluís Màrquez. 2009. On the Robustness of Syntactic and Semantic Features for Automatic MT Evaluation. En *Proceedings of the Fourth Workshop on Statistical Machine Translation*, páginas 250–258.

Kendall, Maurice. 1938. A New Measure of Rank Correlation. *Biometrika*, 30:81–89.

Kulesza, Alex y Stuart M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. En *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, páginas 75–84.

Le, Audrey y Mark Przybocki. 2005. NIST 2005 machine translation evaluation official results. En *Official release of automatic evaluation scores for all submissions, August.*

Lin, Chin-Yew y Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. En *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL).*

Liu, Ding y Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. En *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, páginas 25–32.

Mehay, Dennis y Chris Brew. 2007. BLEUATRE: Flattening Syntactic Dependencies for MT Evaluation. En *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI).*

Melamed, I. Dan, Ryan Green, y Joseph P. Turian. 2003. Precision and Recall of Machine Translation. En *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL).*

Owczarzak, Karolina, Declan Groves, Josef Van Genabith, y Andy Way. 2006. Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation. En *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, páginas 148–155.

Owczarzak, Karolina, Josef van Genabith, y Andy Way. 2007. Labelled Dependencies in Machine Translation Evaluation. En *Proceedings of the ACL Workshop on Statistical Machine Translation*, páginas 104–111.

Papineni, Kishore, Salim Roukos, Todd Ward, y Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation, RC22176. Informe técnico, IBM T.J. Watson Research Center.

Popovic, Maja y Hermann Ney. 2007. Word Error Rates: Decomposition over POS classes and Applications for Error Analysis. En *Proceedings of the Second Workshop on Statistical Machine Translation*, páginas 48–55, June.

Przybocki, M., K. Peterson, y S. Bronsart. 2008. Official results of the NIST 2008 "Metrics for MAchine TRanslationÇhallenge (MetricsMATR08). Informe técnico, National Institute of Standards and Technology.

Reeder, Florence, Keith Miller, Jennifer Doyon, y John White. 2001. The Naming of Things and the Confusion of Tongues: an MT Metric. En *Proceedings of the Workshop on MT Evaluation "Who did what to whom?.ªt Machine Translation Summit VIII*, páginas 55–59.

Spearman, Charles. 1904. The Proof and Measurement of Association Between Two Rings. *American Journal of Psychology*, 15:72–101.