Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 143-146 recibido 20-03-2018 revisado 28-04-2018 aceptado 03-05-2018

QUALES: Estimación Automática de Calidad de Traducción Mediante Aprendizaje Automático Supervisado y No-Supervisado

QUALES: Machine Translation Quality Estimation via Supervised and Unsupervised Machine Learning

Thierry Etchegoyhen[1], **Eva Martínez Garcia**[1], **Andoni Azpeitia**[1], **Iñaki Alegria**[2], **Gorka Labaka**[2], **Arantza Otegi**[2], **Kepa Sarasola**[2], **Itziar Cortes**[3], **Amaia Jauregi**[3], **Igor Ellakuria**[4], **Eusebi Calonge**[5], **Maite Martin**[5]
[1]Vicomtech - {tetchegoyhen, emartinez, aazpeitia}@vicomtech.org
[2]IXA taldea, University of the Basque Country (UPV/EHU)
{i.alegria, gorka.labaka, arantza.otegi, kepa.sarasola}@ehu.eus
[3]Elhuyar - {i.cortes, a.jauregi}@elhuyar.eus
[4]ISEA - isantos@iseamcc.net
[5]Ametzagaiña - ecalonge@ametza.com, maite@adur.com

Resumen: La estimación automática de calidad (EAC) de la traducción automática consiste en medir la calidad de traducciones sin acceso a referencias humanas, habitualmente mediante métodos de aprendizaje automático. Un buen sistema EAC puede ayudar en tres aspectos del proceso de traducción asistida por medio de traducción automática y posedición: aumento de la productividad (descartando traducciones automáticas de mala calidad), estimación de costes (ayudando a prever el coste de posedición) y selección de proveedor (si se dispone de varios sistemas de traducción automática). El interés en este campo de investigación ha crecido significativamente en los últimos años, dando lugar a tareas compartidas a nivel mundial (WMT) y a una fuerte actividad científica. En este artículo, se hace un repaso del estado del arte en este área y se presenta el proyecto QUALES que se está realizando.
Palabras clave: Estimación de calidad, Traducción automática, Aprendizaje automático

Abstract: The automatic quality estimation (QE) of machine translation consists in measuring the quality of translations without access to human references, usually via machine learning approaches. A good QE system can help in three aspects of translation processes involving machine translation and post-editing: increasing productivity (by ruling out poor quality machine translation), estimating costs (by helping to forecast the cost of post-editing) and selecting a provider (if several machine translation systems are available). Interest in this research area has grown significantly in recent years, leading to regular shared tasks in the main machine translation conferences and intense scientific activity. In this article we review the state of the art in this research area and present project QUALES, which is under development.
Keywords: Quality Estimation, Machine Translation, Machine Learning

1 Participantes y entidades financiadoras

QUALES es un proyecto de investigación subvencionado por el Gobierno Vasco a través de la convocatoria de ayudas ELKARTEK 2017 de la Agencia Vasca de desarrollo empresarial Spri.[1]

El proyecto tiene una duración total de 21 meses, con comienzo el 1 de abril de 2017 y finalización el 31 de diciembre de 2018.

QUALES ha sido diseñado y se está lle-

[1]http://www.spri.eus

ISSN 1135-5948. DOI 10.26342/2018-61-18
© 2018 Sociedad Española para el Procesamiento del Lenguaje Natural

vando a cabo por el siguiente consorcio: Vicomtech[2], grupo IXA de la UPV/EHU[3], Elhuyar[4], ISEA[5] y Ametzagaina[6]. Son empresas adheridas Argia, Mondragon Lingua y Eleka. El proyecto tiene asignado el código KK-2017/00094 y el sitio web asociado es http://quales.eus/

2 Contexto y motivación

Con los avances obtenidos por los métodos basados en datos, estadísticos o enmarcados en redes neuronales profundas, la traducción automática (TA) ha logrado niveles de calidad suficientes para su uso en la industria. En particular, los proveedores de servicios lingüísticos de traducción requieren poder integrar componentes de traducción automática para dar una respuesta eficiente y de alta calidad a las demandas de traducción en entornos y dominios variados. Pese a estos progresos notables, la calidad de las traducciones automáticas puede variar significativamente según el dominio, los idiomas considerados o la complejidad de los segmentos individuales por traducir. Esta variabilidad genera problemas bien identificados, entre los cuales los principales son:

- *Productividad*: las traducciones automáticas de pésima calidad requieren un esfuerzo cognitivo importante por parte de los traductores profesionales, en particular para determinar si ciertas partes de la traducción automática son recuperables y qué correcciones aplicar. El enfrentarse a traducciones de baja calidad genera así pérdidas de productividad y frustración para los profesionales del sector. De forma similar, traducciones automáticas correctas a nivel gramatical pero con errores a nivel semántico implican un esfuerzo importante de identificación y corrección de los errores de traducción.
- *Estimación de costes*: los proveedores de servicios lingüísticos ofertan varios servicios según las demandas y la complejidad de las traducciones. Estos servicios pueden ser traducción automática completa, con posedición humana, o realizada por completo por traductores humanos, con o sin el apoyo de memorias de traducción existentes. A cada uno de estos servicios le corresponde una tarificación precisa según el esfuerzo necesario, desde el uso de TA únicamente, con coste mínimo, hasta la traducción humana por completo, con coste máximo. Para poder establecer los costes correctos, es necesario poder determinar la calidad de cada traducción automática y establecer así automáticamente el tipo de servicio óptimo correspondiente.
- *Selección de traducciones*: cada industria con necesidades multilingües puede acceder a una gama variada de servicios de traducción automática, desde sistemas propietarios entrenados para diferentes dominios hasta las ofertas de traducción genéricas online. Cada sistema de TA suele producir traducciones diferentes debidas a los diferentes datos y métodos empleados para el modelado del sistema, con niveles de calidad variables. Resulta imprescindible, entonces, la estimación automática de la calidad de cada una de las traducciones generada por los diferentes sistemas, para poder seleccionar así el mejor conjunto de traducciones.

Tradicionalmente, la calidad de las traducciones automáticas se ha medido de forma estática, comparando un subconjunto de las traducciones con referencias creadas por profesionales humanos. Las comparaciones se establecen usando métricas automáticas que calculan, con cierta aproximación, la distancia entre las traducciones automáticas y las referencias. Este enfoque sigue siendo fundamental para obtener una medida de la calidad general de los sistemas de TA en los dominios considerados y permitir avances incrementales en el desarrollo de los sistemas en función de los resultados objetivos obtenidos con estas métricas.

Pese a estos aspectos positivos, este tipo de medida de calidad es problemático en dos aspectos principales. En primer lugar, la correlación entre los resultados de las métricas automáticas y las valoraciones humanas es baja a nivel de frases o segmentos, lo que no permite una estimación adecuada de la calidad de los segmentos individuales. En segundo lugar, este enfoque implica disponer

[2]http://www.vicomtech.org
[3]http://ixa.eus
[4]http://www.elhuyar.eus/
[5]http://www.iseamcc.net
[6]http://www.ametza.com

de traducciones de referencia, lo cual no es realista a la hora de evaluar la calidad de los amplios volúmenes de traducciones automáticas generadas.

Esta limitación de las métricas estáticas a la hora de medir la calidad de las traducciones automáticas impone el desarrollo de métodos alternativos. La estimación automática de calidad (EAC) (Blatz y otros, 2004; Specia, Raj, y Turchi, 2010) se centra en responder a este reto, a través de sistemas que permitan medir la calidad de traducciones individuales, sin acceso a referencias humanas, empleando habitualmente métodos de aprendizaje automático. Este campo de investigación y desarrollo ha crecido significativamente en los últimos años, dando lugar a tareas compartidas a nivel mundial y una fuerte actividad científica.

Una EAC exitosa permitiría aportar una solución a los tres problemas principales indicados anteriormente, al ofrecer mecanismos adecuados de medida de calidad para cada segmento de traducción automática generado por cualquier sistema de TA. A estos retos responde el proyecto Quales, mediante el desarrollo de métodos supervisados y no-supervisados para la estimación automática de calidad.

3 Estado del arte

Los enfoques supervisados han sido el paradigma dominante en EAC, donde traducciones automáticas anotadas o poseditadas se usan para entrenar clasificadores (Blatz y otros, 2004; Quirk, 2004) o regresores (Specia y otros, 2009). Los sistemas participantes en las tareas compartidas de las conferencias WMT han sido así tipicamente basados en enfoques supervisados, con diferencias centradas en diferentes conjuntos de características (*features*) o en los métodos de aprendizaje automático utilizados, p. ej. Support Vector Machines o Gaussian Processes (Callison-Burch y otros, 2012).

Los sistemas baseline estándares para la tarea se generan habitualmente con las herramientas QUEST (Specia et al., 2013), o QUEST++ (Specia, Paetzold, y Scarton, 2015), en base a 17 características que incluyen puntuación de perplejidad en base a modelos de lenguaje, probabilidades de traducción léxica, o ratios de ocurrencias de palabras, entre otros.

En trabajos recientes, los enfoques basados en redes neuronales han sido empleados también de forma exitosa para la tarea de estimación automática de calidad, bien sea mediante características adicionales (Shah y otros, 2016) o como sistemas EAC completos (Kim, Lee, y Na, 2017; Martins, Kepler, y Monteiro, 2017). En la última edición de la tarea compartida en WMT 2017, los mejores sistemas basados en redes neuronales incrementaron notablemente las prestaciones de la baseline (Bojar y otros, 2017). Así, en la traducción de alemán a inglés los valores del índice de Pearson de los dos mejores sistemas fueron de 0,728 y 0,715, muy por encima del valor baseline (0,441). Para el sentido de traducción inverso, los resultados fueron similares con 0,695 y 0,673 para los mejores sistemas, y 0,307 para la baseline.

Aunque permitan obtener las estimaciones las más precisas a día de hoy, los enfoques supervisados dependen de anotaciones o posediciones humanas. Este aspecto es problemático dada la gran cantidad de diferentes dominios y pares de idiomas en los que se requiere aplicar la tecnología. Considerando los recursos y esfuerzos necesarios para anotar o poseditar conjuntos de muestras adecuados para entrenar sistemas de calidad, el coste de los métodos supervisados puede resultar prohibitivo.

Enfoques no-supervisados que no necesiten datos anotados, basados estrictamente en las características de las frases de origen y/o de las frases traducidas, tienen la notable ventaja de poder adaptarse más fácilmente a distintos dominios y pares de idiomas. Pese a ofrecer este tipo de ventajas, pocos estudios se han centrado en enfoques no-supervisados para EAC. Uno de ellos es (Moreau y Vogel, 2012), donde la estimación de calidad se ejecuta en base a amplios conjuntos de n-gramas y medidas de similitud. (Popovic, 2012) es otra alternativa, basada en la combinación de puntuaciones obtenidas por modelos de lenguaje y probabilidades de traducción léxica sobre morfemas y categorías gramaticales. Ninguno de estos enfoques ha logrado superar hasta hoy las baselines supervisadas.

4 QUALES

En el marco de QUALES, se investigan tanto métodos supervisados avanzados basados en redes neuronales como métodos no-supervisados que permitan desarrollar estimadores de calidad de forma eficiente en dis-

tintos casos de uso.

Tras su puesta en marcha en 2017, el proyecto ha logrado los primeros resultados siguientes:

- Creación de datos de entrenamiento y validación manuales y sintéticos para los pares de idiomas euskera-castellano e inglés-castellano en el dominio de las noticias.
- Despliegue de baselines supervisadas basadas en QUEST++.
- Desarrollo de sistemas de EAC basados en redes neuronales y aprendizaje profundo, explotando espacios vectoriales bilingües.
- Desarrollo de un sistema de EAC basado en características mínimas, en versión supervisada y no-supervisada. Ambas versiones superan significativamente las baselines sobre los datos de las tareas compartidas WMT 2015, 2016 y 2017.

Los primeros resultados del proyecto son satisfactorios, en particular los obtenidos mediante métodos minimalistas no-supervisados que superan significativamente a sistemas supervisados robustos y permiten un despliegue eficiente de estimadores fiables para nuevos dominios.

QUALES aportará además los primeros resultados para el par de idiomas euskera-castellano en el campo de la estimación automática de calidad, lo cual constituye un objetivo importante del proyecto.

Durante 2018, el esfuerzo se centrará en extender y mejorar los primeros sistemas desarrollados, y en validar los resultados obtenidos. La convocatoria en la que se enmarca el proyecto apoya a proyectos de investigación con alto potencial industrial y se validará el potencial de los métodos explorados para un uso en entornos profesionales.

Bibliografía

Blatz, J. et al. 2004. Confidence estimation for machine translation. En *Proceedings of COLING*, páginas 315–321.

Bojar, O. et al. 2017. Findings of the 2017 conference on machine translation. En *Proceedings of the Second Conference on Machine Translation*, páginas 169–214, Copenhagen, Denmark.

Callison-Burch, C. et al. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. En *Proceedings of the Seventh Workshop on Statistical Machine Translation*.

Kim, H., J.-H. Lee, y S.-H. Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. En *Proceedings of the Second Conference on Machine Translation*, páginas 562–568, Copenhagen, Denmark.

Martins, A. F. T., F. Kepler, y J. Monteiro. 2017. Unbabel's participation in the wmt17 translation quality estimation shared task. En *Proceedings of the Second Conference on Machine Translation*, páginas 569–574, Copenhagen, Denmark.

Moreau, E. y C. Vogel. 2012. Quality estimation: an experimental study using unsupervised similarity measures. En *Proceedings of the Seventh Workshop on Statistical Machine Translation*, páginas 120–126.

Popovic, M. 2012. Morpheme- and pos-based IBM1 and language model scores for translation quality estimation. En *Proceedings of the Seventh Workshop on Statistical Machine Translation*, páginas 133–137.

Quirk, C. 2004. Training a sentence-level machine translation confidence measure. En *Proceedings of LREC*, páginas 825–828.

Shah, K. et al. 2016. SHEF-LIUM-NN: Sentence level Quality Estimation with Neural Network Features. En *Proceedings of the First Conference on Machine Translation*, volumen 2, páginas 838–842.

Specia, L. et al. 2009. Estimating the sentence-level quality of machine translation systems. En *13th Conference of the European Association for Machine Translation*, páginas 28–37.

Specia, L., G. Paetzold, y C. Scarton. 2015. Multi-level translation quality prediction with QUEST++. *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, páginas 115–120.

Specia, L., D. Raj, y M. Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.