TEISEL
Tecnologías para la investigación en segundas lenguas
ISSN 2696-676X
Universitat de Barcelona
Facultat de Filologia i Comunicació

VERTa: una métrica de evaluación de la traducción automática. Aplicaciones a la investigación sobre el español y el inglés como L2

Elisabet COMELLES PUJADAS (Autora para correspondencia)
Universitat de Barcelona (España)
elicomelles@ub.edu
https://orcid.org/0000-0002-4753-2712

Resumen: Este artículo presenta *VERTa* (https://github.com/jatserias/VERTa para la versión completa y http://grial.ub.edu:8080/VERTaDemo/ para la demo *online* en español), una métrica de evaluación de la traducción automática (TA) para el inglés y el español. *VERTa* es una métrica que utiliza información lingüística para evaluar frases traducidas automáticamente, a través de la comparación de dichas frases con frases traducidas por traductores humanos. A diferencia de otras métricas, *VERTa* proporciona no tan solo una puntuación por cada frase, sino también un análisis más cualitativo de los resultados obtenidos. El artículo analiza los pasos que se llevaron a cabo antes de diseñar e implementar la métrica: el estudio lingüístico del corpus de desarrollo para encontrar aquellas características más significativas que la métrica debía poder cubrir y las herramientas de procesamiento del texto que debían aplicarse a los segmentos comparados. Más adelante se describen los diferentes módulos que forman la métrica y la información que aportan, junto con ejemplos de la información que recibe el usuario. Aunque *VERTa* es una métrica de evaluación de la TA, se diferencia del resto en que durante su desarrollo se puso especial énfasis en analizar la información lingüística que debía aportar al usuario, para ir más allá de una mera puntuación del segmento traducido y poder servir como una primera guía cualitativa para detectar los errores de la traducción automática. En consecuencia, *VERTa* puede utilizarse en el área del aprendizaje, de la enseñanza y de la evaluación del inglés y del español como segundas lenguas y como lenguas extranjeras, así como en la investigación en este campo.

Palabras clave: traducción automática; lingüística informática; evaluación; inglés; español

Català:

VERTa: una mètrica de l'avaluació de la traducció automàtica. Aplicacions a la recerca sobre l'espanyol i l'anglès com a L2

***Resum**: Aquest article presenta VERTa (https://github.com/jatserias/VERTa per a la versió completa i http://grial.ub.edu:8080/VERTaDemo/ per a la demo online en castellà), una metrica d'avaluació de la traducció automàtica (TA) per a l'anglès i el castellà. VERTa utilitza informació lingüística per avaluar frases traduïdes automàticament, tot comparant-les amb frases traduïdes per traductors humans. A diferència d'altres mètriques, VERTa no només proporciona una puntuació per a cada frase traduïda, sinó que també fa una anàlisi més qualitativa dels resultats obtinguts. Aquest article descriu les passes que es van realitzar abans de dissenyar i implementar la mètrica: l'estudi lingüístic del corpus de desenvolupament per tal de trobar les característiques lingüístiques més rellevants que la mètrica havia de tractar i les eines de processament textual amb què s'haurien de tractar els segments comparats. A més, l'article analitza els mòduls que formen la mètrica i la informació que proporcionen, juntament amb exemples de la informació que rep l'usuari. Tot i que VERTa és una mètrica d'avaluació de la TA, es diferencia de la resta de mètriques pel fet que la informació que proporciona va més enllà d'una simple puntuació del segment avaluat i serveix com una primera guia qualitativa per detectar errors de traducció*

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

automàtica. En conseqüència, VERTa pot utilitzar-se per a l'aprenentatge, l'ensenyament i l'avaluació de l'anglès i el castellà com a segones llengües o com a llengües estrangeres, i també per a la recerca en aquesta àrea.

***Paraules clau**: traducció automàtica; lingüística computacional; avaluació; anglès; castellà*

English:

VERTa: a machine translation evaluation metric. Applications to L2 research on Spanish and English

***Abstract**: This article presents VERTa (https://github.com/jatserias/VERTa for the full version and http://grial.ub.edu:8080/VERTaDemo/ for the Spanish online demo), a machine translation (MT) evaluation metric for English and Spanish. VERTa uses linguistic information to evaluate machine-translated sentences by comparing them with sentences translated by human translators. Unlike other metrics, VERTa provides not only a score for each sentence compared, but also a more qualitative analysis of the results obtained. This article discusses the steps carried out before designing and implementing the metric: the linguistic study of the development corpus to find the most relevant linguistic features that the metric should be able to cover, and the text processing tools to be applied to the compared segments. In addition, it details the modules included in the metric and the information they provide, together with examples of the information the user receives. Although VERTa is an MT evaluation metric, it differs from the rest in that during its development special emphasis was placed on analyzing the linguistic information it should provide to the user, thus going beyond a mere scoring of the translated segment and serving as a first qualitative guide to detect machine translation errors. Consequently, VERTa can be used for the learning, teaching and evaluation of English and Spanish as second and/or foreign languages, as well as to carry out research studies in this area.*

***Keywords**: automatic translation; computational linguistics; evaluation; English; Spanish*

Introducción

La traducción automática (TA) es uno de los mayores retos dentro del área del Procesamiento del Lenguaje Natural (PLN), debido al conocimiento y la complejidad que esta tarea conlleva. Según Basnett (1980), cuando una persona traduce un texto "*a process of decoding and encoding takes place*" (p. 24), es decir, traducir es una operación cognitiva que implica descodificar el significado de un texto origen y codificarlo en la lengua meta. Así pues, el traductor debe entender y analizar el texto de partida, lo cual requiere un buen conocimiento de todas las dimensiones de la lengua origen (p. ej. el léxico, la gramática, la semántica) así como conocimiento de la cultura origen. Además, el proceso de codificación implica el mismo conocimiento de la lengua meta. Así pues, si tenemos en cuenta la complejidad de este proceso, entenderemos el reto que supone que esta misma operación la realice un ordenador. Sin embargo, en la última década el uso de redes neuronales ha ayudado a mejorar ostensiblemente la calidad de los textos traducidos automáticamente. Por este motivo la TA ya no solo se utiliza cuando se desea obtener una idea general de un texto escrito en una lengua que se desconoce, sino que muchos usuarios utilizan esta tecnología para escribir textos en una lengua que no dominan, e incluso la TA ha empezado a usarse en la enseñanza de las segundas lenguas (L2) y en la investigación en esta área. De hecho, Enríquez et al. (2019) acuñan el término *computer-assisted L2 learning and translation (CAL2T),* que pretende reflejar, entre otros aspectos, el incremento de las tecnologías basadas en datos (p. ej. el uso de corpus o de TA) en entornos de aprendizaje mediados por ordenador.

Una de las tareas directamente ligada a la TA es la evaluación de los resultados obtenidos. En este sentido, hay diferentes tipos de evaluación, pero todos ellos tienen en común el objetivo de velar por la calidad lingüística del texto traducido automáticamente. Inicialmente el texto resultante de la TA era evaluado por humanos que se centraban en diferentes aspectos de la traducción. No cabe duda de que este tipo de evaluación es de gran utilidad, ya que estos

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

"jueces" tienen el conocimiento que la máquina intenta simular, aunque también debemos aceptar que se trata de una tarea subjetiva, lenta y que conlleva un alto coste económico. Como reacción a estas desventajas, en las últimas dos décadas se han desarrollado varias métricas automáticas de evaluación de la TA, que permiten a los desarrolladores de sistemas y estudiosos del área llevar a cabo evaluaciones que supuestamente son más rápidas, baratas y objetivas que las evaluaciones humanas. También dentro de la enseñanza de L2, en los últimos años, se han llevado a cabo diversos estudios sobre el uso de las métricas para la evaluación de las traducciones de los alumnos (Han y Lu, 2021; Munkova et al., 2021).

En este artículo se presenta VERTa[1], una de estas métricas disponible para el inglés y el español[2], se analizan los pasos que se han seguido para su diseño y desarrollo, se describe su funcionamiento y, por último, se discuten diferentes usos de la métrica dentro del área de la enseñanza, el aprendizaje y la evaluación de L2.

1. Breve descripción de las métricas de TA

Las métricas de TA funcionan como medidas de similitud y comparan el resultado de la TA con traducciones de referencia generadas por traductores expertos. Esta comparación se suele utilizar para diferentes tipos de evaluaciones:

a) *adequacy*: se evalúa que toda la información del texto origen esté en el texto meta, es decir, se centra en la semántica
b) *fluency*: se evalúa que el texto meta sea gramatical según la lengua de destino
c) *ranking*: se evalúan y ordenan diferentes salidas de TA de mejor a peor según su calidad global

Dentro de las métricas de evaluación de la TA, algunas no utilizan ningún tipo de información lingüística, como BLEU (Papineni et al., 2001), NIST (Doddington, 2002) o ROUGE (Lin y Och, 2004), otras funcionan a nivel de caracteres, como la familia de métricas ChrF (Popović, 2015; Popović, 2017). Otras métricas utilizan información lingüística más o menos sofisticada: algunas usan información léxica (p. ej. sinónimos) como METEOR (Denkowki y Lavie, 2014); otras emplean información morfológica (p. ej. información sobre sufijos, prefijos, raíces) como INFER (Popović 2012); otras se centran en la morfosintaxis y la sintaxis (p. ej. las categorías gramaticales, los constituyentes de la oración, las relaciones de dependencia) como SMT y HWCM (Liu y Gildea, 2005), UoWReVal (Gupta et al., 2015) y METEOR++2.0 (Guo y Hu, 2019); y otras utilizan información semántica (p. ej. roles semánticos) como MEANT 2.0 (Lo, 2017).

La mayoría de las métricas mencionadas anteriormente evalúan características parciales del texto, sin embargo, cuando un humano evalúa una traducción, lo hace de manera global. Por este motivo, algunos desarrolladores de métricas han explorado también diferentes maneras de combinar esta información, algunos mediante técnicas de

[1] El nombre de VERTa es un guiño a otras métricas de evaluación de la traducción automática, como BLEU o ROUGE. En nuestro caso, optamos por el color verde y lo combinamos con Ta (traducción automática), en un acrónimo que evoca un nombre de mujer ("Berta").

[2] Una demo de la métrica para el español está disponible *online* (http://grial.ub.edu:8080/VERTaDemo/), aunque se trata de una versión reducida que no utiliza todos los módulos de la métrica.

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

aprendizaje automático (Joty et al., 2014; Ma et al., 2017) y otros con enfoques más simples como la media de los resultados obtenidos para cada tipo de información lingüística (Giménez y Márquez, 2010; González et al., 2014).

La mayoría de las métricas de TA funcionan a nivel de segmento, es decir, comparan el segmento traducido automáticamente con un segmento traducido por una persona y dan una puntuación al segmento de TA. No obstante, estas métricas no suelen dar más información que esta evaluación cuantitativa y para el usuario o el desarrollador es difícil identificar los errores de traducción que está cometiendo el sistema. En otras palabras, estas métricas no suelen proporcionar una evaluación cualitativa que pueda servir como una primera guía para identificar los errores o aquellos aspectos de la traducción que no funcionan. VERTa, la métrica que se describe en este artículo, pretende ser un primer paso hacia una evaluación más cualitativa de la TA.

2. VERTa: Una métrica de evaluación de la traducción basada en información lingüística

VERTa se enmarca en el grupo de métricas que comparan el texto traducido automáticamente con la traducción de referencia creada por un traductor humano. VERTa utiliza información lingüística de diferentes niveles para comparar segmentos (frases) y como resultado produce una puntuación que indica cómo de similares son el segmento traducido automáticamente y la traducción de referencia, así como unos gráficos que pueden ayudar al usuario a interpretar esa puntuación.

2.1 Análisis de necesidades previo al diseño de la métrica

Para diseñar VERTa se tuvieron en cuenta dos cuestiones primordiales y que están entrelazadas: la primera, decidir cuál es la información lingüística relevante a la hora de evaluar una traducción automática, y la segunda, qué recursos y herramientas lingüísticas se deben utilizar para etiquetar el texto que se quiere evaluar y así poder tratar la información lingüística relevante.

Para abordar la primera, se llevó a cabo un detallado análisis cualitativo con la intención de identificar aquella información lingüística que la métrica debía poder considerar al comparar ambos segmentos y decidir la bondad del segmento traducido automáticamente. Para realizar este análisis, diseñar la métrica, implementarla y ajustar la combinación de los diferentes módulos que la forman se usaron dos corpus, uno para el inglés y otro para el español[3]. El corpus del inglés se extrajo de los datos proporcionados en dos campañas de evaluación[4], la NIST Open-MT05[5] y la MetricsMatr Evaluation Task 2010[6]. Del conjunto de datos proporcionados en la NIST Open-MT05 se usaron un total

[3] En una fase posterior se adaptó la métrica a la evaluación de la TA para el español.

[4] Las campañas de evaluación son competiciones que promueven el desarrollo de sistemas de TA.

[5] https://catalog.ldc.upenn.edu/LDC2010T14

[6] http://www.statmt.org/wmt10/evaluation-task.html

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

de 600 segmentos traducidos automáticamente del árabe al inglés mediante seis traductores automáticos diferentes, cuatro traducciones de referencia para cada segmento traducido y la evaluación de dos jueces humanos. Estos datos se utilizaron para realizar el análisis lingüístico, así como para desarrollar la métrica y encontrar la mejor combinación de módulos para evaluar la *fluency*. Por otro lado, de los datos ofrecidos por la MetricsMatr Evaluation Task 2010 se usaron un total de 800 segmentos traducidos automáticamente del árabe al inglés mediante ocho sistemas de traducción automáticos, cuatro traducciones de referencia por segmento traducido automáticamente y las correspondientes valoraciones de los jueces. Estos datos se utilizaron para el análisis lingüístico, para la implementación de la métrica y para la configuración de los módulos para evaluar la *adequacy*. En el caso del corpus del español, se utilizó una parte del corpus desarrollado en el proyecto KNOW2[7]. Este corpus consistía en 187 glosas de WordNet traducidas del inglés al español mediante dos sistemas de traducción automática[8], cuatro traducciones de referencia para cada segmento y la valoración de dos jueces. El análisis lingüístico del inglés está descrito en detalle en Comelles et al. (2017) y el del español en Comelles (2015). Siguiendo la tipología propuesta por Farrús et al. (2010), las características lingüísticas resultantes de este análisis se clasificaron en los siguientes niveles: información léxica, información morfológica, información sintáctica e información semántica. Dentro de la información léxica se tuvo en cuenta la semántica léxica, incluyendo relaciones de sinonimia, hiponimia e hiperonimia. En cuanto a la información morfológica, se incluyeron características morfosintácticas como la desinencia verbal, la información de género y número en los sustantivos, la categoría léxica, etc. En la información sintáctica se consideraron el orden de las palabras, las relaciones de dependencia y las diátesis. Por último, la información semántica se centraba en los roles semánticos, el reconocimiento de entidades nombradas (*named entities)*, la identificación de expresiones temporales equivalentes o la connotación positiva o negativa de la frase.

El análisis lingüístico que se llevó a cabo fue de mucha ayuda a la hora de confirmar aquellas características lingüísticas que otras métricas del estado del arte utilizan, así como proponer nuevos rasgos como el uso de hiperónimos e hipónimos, o el uso de las diátesis. A la vez, este análisis afianza la combinación de información lingüística a diferentes niveles para asegurar un enfoque holístico más amplio de la evaluación de la TA. Por último, también enfatiza que la importancia de la información lingüística utilizada varía dependiendo de la lengua que se esté evaluando y del tipo de evaluación que se realice, ya esté más enfocada en la semántica del texto meta o en su gramaticalidad.

Para tratar la información lingüística relevante identificada en el análisis lingüístico, se seleccionaron y evaluaron diversos recursos lingüísticos y herramientas que permiten el procesamiento y anotación automáticos del texto traducido automáticamente y de los segmentos de referencia. En cuanto a los recursos utilizados, se optó por el

[7] http://ixa.si.ehu.es/know2/

[8] La traducción automática de las glosas de WordNet se realizó dentro del proyecto KNOW-2 (http://ixa.si.ehu.es/know2/) con el fin de actualizar los WordNets del catalán y español a la versión de WordNet 3.0 (Oliver y Climent, 2012). Las glosas se tradujeron utilizando un sistema de traducción basado en reglas y un sistema de traducción estadístico.

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

uso de WordNet (Fellbaum, 1998), una base de datos léxica que contiene nombres, adjetivos, verbos y adverbios organizados en *synsets* o grupos de sinónimos. Se decidió optar por este recurso porque proporciona una amplia cobertura tanto para el inglés como para el español, es de código libre y permite dar cuenta de las relaciones de sinonimia, hiperonimia e hiponimia, ya que los *synsets* están conectados por relaciones semántico-conceptuales y léxicas. Además, WordNet cuenta con una librería que permite lematizar, lo cual también fue muy útil a nivel léxico.

En cuanto a las herramientas para la anotación morfosintáctica y de dependencias, se optó por el analizador Standford CoreNLP suite (Manning et al., 2014) para analizar los datos del inglés. Esta herramienta se utiliza para etiquetar el corpus con categorías gramaticales, funciones y relaciones de dependencias, y para identificar y normalizar expresiones temporales que incluyen tanto momentos en el tiempo como la duración.

Para el tratamiento del texto en español, la herramienta seleccionada fue Freeling (Padró y Stanilovsky, 2012), porque integra diferentes tipos de análisis y es uno de los analizadores más populares para el tratamiento automático del español. Freeling se utiliza para lematizar, para etiquetar con categorías gramaticales y para obtener el análisis de dependencias del texto.

Además de estas herramientas, para el tratamiento del texto en inglés también se utilizan el etiquetador Supersense Tagger (Ciaramita y Altun, 2006), que identifica entidades nombradas (NEs), y la herramienta NEL, inspirada por Hachey et al., (2011), que vincula entidades nombradas que comparten el mismo referente a través de la Wikipedia. Por otro lado, para calcular la polaridad de un segmento, se usa una estrategia basada en diccionarios (Atserias et al., 2012) para determinar si la polaridad contextual de un segmento es negativa, positiva o neutra. Por último, se emplea el Modelo de Lenguaje (ML) News LM[9], un recurso desarrollado a partir de los corpus de noticias ofrecidos en la competición WMT11[10], que es de mucha utilidad para identificar segmentos que son gramaticalmente correctos a pesar de no estar representados en los segmentos de referencia con los que se comparan los segmentos traducidos automáticamente.

2.2 Diseño y arquitectura de VERTa

El análisis lingüístico fue el punto de partida para diseñar VERTa como un sistema modular, en el que cada uno de los módulos funciona primero individualmente y después se combinan según el tipo de evaluación que se quiere llevar a cabo y también según la lengua meta. Los módulos de VERTa son el módulo léxico, el módulo morfológico, el módulo de dependencias y el módulo semántico. Además de los módulos mencionados anteriormente, VERTa también incluye un módulo de n-gramas, que comprueba la similitud entre segmentos de texto, así como un módulo que utiliza un ML. En el campo de la evaluación de la TA, la validez de las métricas se suele medir a partir las correlaciones con los juicios

[9] http://www.quest.dcs.shef.ac.uk/quest_files/de-en/news.3gram.en.lm

[10] https://www.statmt.org/wmt11/

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

humanos: cuanto más fuerte es la correlación, mejor es la métrica. Es decir, para evaluar una métrica se compara la puntuación asignada por la métrica a cada uno de los segmentos traducidos automáticamente con la puntuación de los jueces humanos para cada uno de estos segmentos. Cuanto mejor se correlacionan ambos, mejor se supone que funciona la métrica. Este mismo proceso también se sigue cuando se desarrollan estas métricas automáticas. Sin embargo, para el diseño y el desarrollo de VERTa, nuestra intención era ir más allá de las correlaciones con los juicios humanos; es decir, ir más allá de la puntuación y centrarnos en los fenómenos lingüísticos que se escondían detrás de esa puntuación. Era importante conseguir buenas correlaciones, pero lo era más ver qué estaba pasando al utilizar y combinar cierta información lingüística, para así explorar qué podía aportar la métrica a la evaluación, además de una puntuación concreta. Con esta idea en mente, el resultado que la métrica ofrece no es tan solo un número que indica la calidad de un segmento traducido automáticamente en comparación con uno traducido por un humano, sino que la métrica proporciona una serie de ficheros .xml donde el usuario puede ver qué es lo que está pasando al aplicar cada módulo. Esto puede ser de gran ayuda para acometer un análisis de errores de traducción por niveles y, como se discutirá más adelante, para evaluar traducciones realizadas por estudiantes de L2.

Con los resultados de estos experimentos se desarrolló la versión definitiva de la métrica y se estableció la manera en la que los módulos funcionarían e interactuarían.

2.3 Funcionamiento de VERTa

VERTa es una herramienta de código abierto que funciona en UNIX/Linux. El código para instalar VERTa puede descargarse de https://github.com/jatserias/VERTa

Para utilizar VERTa, el usuario debe tener un texto hipótesis (el texto que quiere evaluar) y un texto de referencia (aquel con el que el texto hipótesis será comparado). Ambos deben ser texto plano en UTF8. En la línea de comandos el usuario deberá especificar las lenguas de los textos (inglés o español) y el tipo de evaluación que quiere realizar (*adequacy, fluency o ranking*).

Una vez introducidos los textos hipótesis y referencia en VERTa, estos textos serán procesados y etiquetados con las herramientas y los recursos de PLN explicados en el apartado 2.1. El resultado de este etiquetado se muestra en la Figura 1, donde se puede ver la frase *The French minister arrived in Marrakech on Friday night* etiquetada. Así, de izquierda a derecha encontramos la etiqueta de palabra *arrived*, el lema correspondiente (*arrive*), la etiqueta morfosintáctica correspondiente (VBD, verbo en pasado), la etiqueta y relación de dependencia que nos indica que *minister* es el sujeto de la palabra número 4 (*arrived*), la entidad nombrada de tipo lugar *Marrakech* y la expresión temporal de fecha y tiempo *Friday night*.

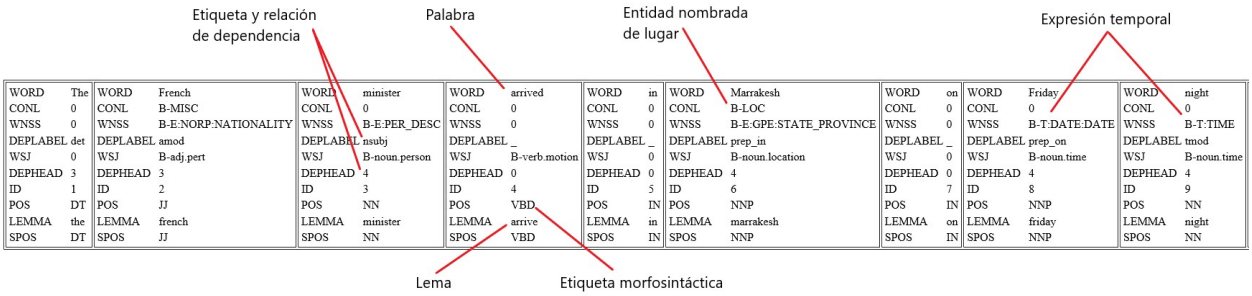
Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

Figura 1

Texto etiquetado



El primer módulo que se utiliza en VERTa es el módulo léxico, que se encarga de hacer el alineamiento de palabras según las características predefinidas. Este módulo compara los ítems léxicos de los segmentos traducidos automáticamente con aquellos de las referencias, aplicando la siguiente información lingüística de manera ordenada:

1. Formas léxicas
2. Lemas
3. Sinónimos
4. Hiperónimos
5. Hipónimos
6. Lema parcial (los primeros caracteres de un lema)

De esta manera, si las formas léxicas coinciden, las dos palabras se alinean. Si no coinciden, se comparan los lemas. Si estos tampoco coinciden, se comprueba en WordNet si las palabras son sinónimas. Si son sinónimas, se alinean. Si no lo son, se comprueba si una es un hiperónimo de la otra y viceversa y, finalmente, se mira si los lemas comparten los primeros caracteres. En caso de que no se pueda producir ningún alineamiento, se continua con la siguiente palabra.

El resultado del módulo léxico se ve en la Figura 2, donde la primera línea corresponde al segmento hipótesis, la tercera es la del segmento referencia y en la franja del medio aparecen las palabras alineadas a partir de la información lingüística incluida en este módulo: *Burned churches 10 Ten Days*. En el caso de *burning* y *Burned*, aunque las formas léxicas no son idénticas, se han podido alinear a través del lema parcial. Las palabras que no han podido ser alineadas aparecen en color rojo. Esto puede indicar que alguna palabra no está suficientemente bien traducida o incluso que alguna palabra no se ha llegado a traducir.

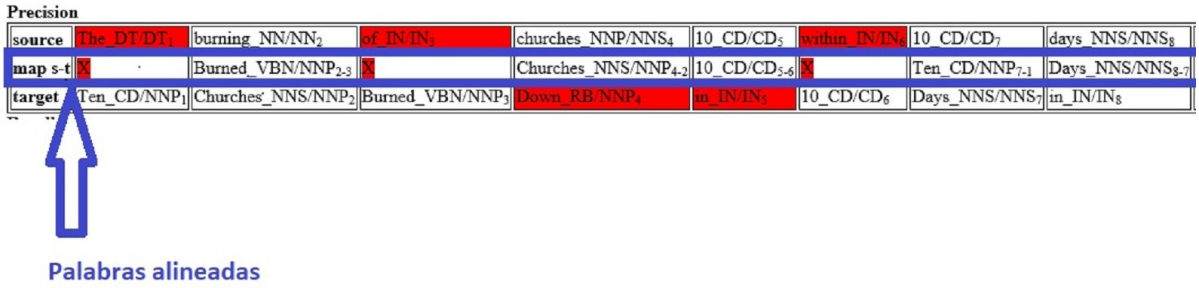
Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

Figura 2

Alineación de palabras a partir del módulo léxico



El segundo módulo que se aplica es el morfológico, que combina las alineaciones realizadas por el módulo léxico con la etiqueta de categoría morfosintáctica de cada palabra. La función de este módulo es restringir la holgada alineación del módulo léxico, que, aunque puede ser muy conveniente a la hora de evaluar el contenido semántico, también puede resultar un problema al evaluar la gramaticalidad de un segmento.

El siguiente módulo que entra en juego es el de dependencias, que pretende hacer de puente entre la sintaxis y la semántica. Este módulo es particularmente útil para encontrar similitudes entre segmentos que presentan un orden de constituyentes diferente o entre segmentos que expresan contenido semántico similar utilizando diferentes estructuras sintácticas. El resultado de este módulo (ver Figura 3) es una tabla donde se pueden ver las tripletas para cada una de las relaciones de dependencia identificadas tanto en el texto origen como en el texto meta, y una última columna con el tipo de emparejamiento de tripletas y el peso asignado. Así por ejemplo, vemos que las tripletas _(TOP:0,*burning*:2) y _(TOP:0,*Burned*:3) cuentan como un emparejamiento completo y se les asigna el peso máximo. Aunque la forma léxica es diferente, el módulo de dependencias también utiliza información de sinonimia, hiperonimia, hiponimia y lema, para proporcionar una mayor cobertura. En el caso de la pareja prep_of(*burning*:2,*churches*:4) y nsubj(*Burned*:3,*Churches*:2), el tipo de emparejamiento no es completo ya que la etiqueta difiere; sin embargo, el tipo de preposición es equivalente y, dado que tanto el núcleo como el modificador coinciden, el peso asignado es el peso máximo. Así pues, mediante la información aportada por este módulo podemos concluir que las estructuras *Burning of churches* y *Churches burned* son semánticamente equivalentes, aunque estén realizadas mediante diferentes construcciones sintácticas. El módulo de dependencias también puede ayudar a identificar estructuras no gramaticales mediante la detección de las etiquetas *dep* y *nomatch*.

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

Figura 3

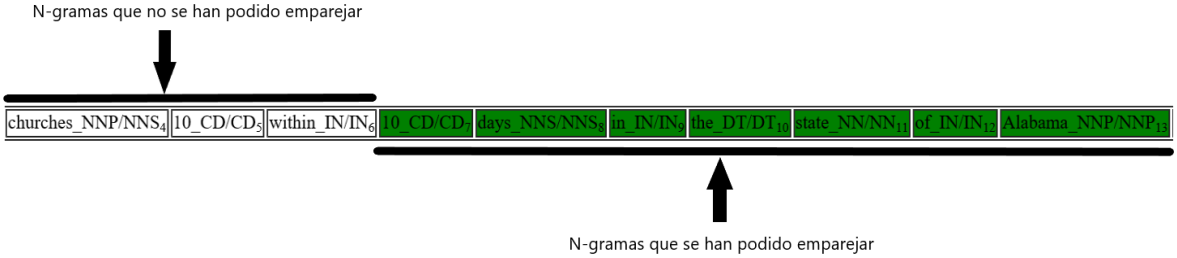
Alineación de tripletas del módulo de dependencias

source	target	Pattern
det(burning:2,The:1)	NO MATCH	nomatch
_(TOP:0,burning:2)	_(TOP:0,Burned:3)	(X,X,X) : 1.0
_(TOP:0,of:3)	NO MATCH	nomatch
prep_of(burning:2,churches:4)	nsubj(Burned:3,Churches:2)	(O,X,X) : 1.0
num(churches:4,10:5)	num(Churches:2,Ten:1)	(X,X,X) : 1.0
_(TOP:0,within:6)	_(TOP:0,in:5)	(X,X,O) : 0.9
num(days:8,10:7)	num(Days:7,10:6)	(X,X,X) : 1.0
prep_within(burning:2,days:8)	prep_in(Burned:3,Days:7)	(O,X,X) : 1.0
_(TOP:0,in:9)	_(TOP:0,in:8)	(X,X,X) : 1.0
det(state:11,the:10)	det(State:11,the:9)	(X,X,X) : 1.0
prep_in(days:8,state:11)	prep_in(Days:7,State:11)	(X,X,X) : 1.0
_(TOP:0,of:12)	_(TOP:0,of:12)	(X,X,X) : 1.0
nn(U.S:14,Alabama:13)	NO MATCH	nomatch
prep_of(state:11,U.S:14)	prep_of(State:11,Alabama:13)	(X,X,O) : 0.9
_(TOP:0,.:15)	NO MATCH	nomatch

El módulo que se aplica a continuación es el de n-gramas, que actúa a partir de la alineación de palabras establecida por el módulo léxico. Este módulo puede configurarse para decidir la longitud de los n-gramas y hacerlo más restrictivo (en el caso de evaluar la *fluency*) o más laxo (en el caso de evaluar la *adequacy*). Los n-gramas que se han podido emparejar aparecen en color verde, mientras que los no emparejados no salen coloreados, tal y como se puede ver en la Figura 4, donde se muestra parte del resultado del módulo de n-gramas al comparar las frases *The burning of churches 10 within 10 days in the State of Alabama U.S.* y *Ten Churches Burned Down in 10 Days in the American State of Alabama*. Al comparar estas dos frases, el módulo de n-gramas empareja la secuencia *10 days in the state of Alabama* (en color verde), pero no puede emparejar la secuencia *churches 10 within* (sin colorear).

Figura 4

Alineación por el módulo de n-gramas



El siguiente módulo que se utiliza es el semántico, que consiste en el análisis de sentimientos (para detectar si las emociones y las opiniones que se expresan son positivas, negativas o neutras), en el reconocimiento y vinculación

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

de entidades nombradas, y en el reconocimiento y vinculación de expresiones temporales. En la Figura 5 se puede ver el resultado del módulo semántico para el par de frases siguiente:

- HYP: *In the 1950's and 1960's of the 20th century, targeting several church frequented by the state of Alabama in the soda.*
- REF: *In the 1950's and 1960's, many churches frequented by blacks were targeted in the state of Alabama.*

Se puede comprobar que hay una entidad nombrada de tipo lugar *Alabama* que es igual en ambas frases y que apunta a la misma página de la Wikipedia, así como dos expresiones temporales del tipo fecha que coinciden en ambas frases. Así, el módulo semántico (en particular la vinculación de entidades nombradas) nos permite emparejar representaciones de una misma entidad, por ejemplo, "George Washington" y "el primer presidente de Estados Unidos", ya que ambas apuntan a la misma página de la Wikipedia.

El último módulo que se aplica es el módulo de ML. Este módulo se diferencia de los anteriores en que no compara el segmento traducido con la referencia, sino que se aplica únicamente al segmento de TA y mediante una serie de cálculos indica la probabilidad de que el segmento traducido se encuentre en el corpus utilizado para generar el ML. De esta manera, se puede comprobar la gramaticalidad del segmento traducido, aunque no esté incluido en el segmento de referencia, y asegurar una cobertura más amplia.

Figura 5

Información del módulo semántico

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

Todos los módulos presentados en esta sección se combinan de manera diferente dependiendo del tipo de evaluación (Comelles y Atserias, 2019): i) evaluación de la *adequacy*, para comprobar si la información semántica del texto origen se ha trasladado al texto meta; ii) evaluación de la *fluency*, para comprobar la correcta gramaticalidad de la frase según la estructura de la lengua meta; ii) evaluación de *ranking*, para evaluar la traducción de manera holística (semántica y gramaticalidad) y, dadas diversas traducciones del mismo segmento, poderlas organizar según su calidad.

Las combinaciones de módulos necesarias para cada tipo de evaluación están implementadas en diferentes ficheros de configuración, que el usuario puede elegir en función de si pretende evaluar la *adequacy* o la *fluency*, o si desea tener en cuenta tanto la semántica como la gramaticalidad del segmento meta.

3. VERTa y la investigación en segundas lenguas (L2)

VERTa es una métrica de similitud inicialmente desarrollada para la evaluación de la traducción automática, aunque también puede ser de gran utilidad para los profesores y estudiantes de segundas lenguas y lenguas extranjeras, así como para los investigadores en esta área.

La última década parece haber visto el resurgimiento de la traducción, tanto como una herramienta más dentro de la enseñanza de L2 (Laviosa y González-Davies, 2019) como una habilidad importante en la comunicación intercultural y mediación, tal como se refleja en el *Marco común europeo de referencia para las lenguas* (MCER, Consejo de Europa, 2001). Traducir de una lengua extranjera a la propia lengua permite que el aprendiz verifique y consolide la comprensión de textos en la lengua extranjera y, por otra parte, traducir de la lengua propia a una lengua extranjera puede ayudar a mejorar la producción escrita y oral del estudiante.

El uso de la TA en la enseñanza y aprendizaje de L2 ha generado mucho debate y controversia entre los profesores. En muchos casos, se ha relacionado el uso de TA con una práctica deshonesta por parte de los estudiantes, como si el uso de TA se considerara una forma de engañar al profesor sobre el conocimiento que el alumno tiene de la L2. En otros casos, se ha considerado que el uso de la TA no ayudaba a los aprendices de L2, ya que los alejaba del proceso de aprendizaje y no les permitía distanciarse de su L1 (Briggs, 2018). Incluso en algunos momentos, se ha desestimado el uso de esta tecnología debido a las malas traducciones que generaba, ya que podían confundir a alumnos de niveles bajos al no tener suficientes conocimientos para discriminar entre buenas o malas traducciones. Sin embargo, a raíz de los avances producidos en la última década en el área de la traducción automática, especialmente desde que se empezaron a utilizar redes neuronales, la calidad de los textos traducidos automáticamente ha mejorado ostensiblemente. Esta es una de las razones por las que la TA es cada vez más popular y algunos profesores han empezado a incluirla como un recurso pedagógico más para el aprendizaje de segundas lenguas y lenguas extranjeras, en concreto como una herramienta de ayuda a la escritura (Yoo-Jean, 2021; Chon et al., 2021; Chung y Ahn, 2021; Lee, 2022).

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

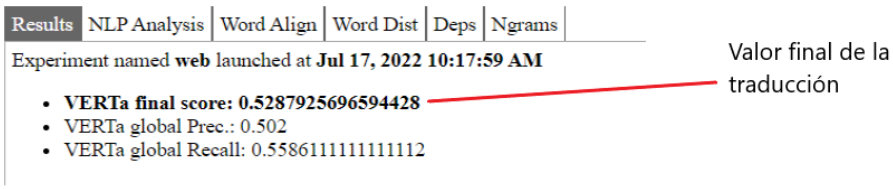
 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

De hecho, algunos investigadores del área (Jiménez-Crespo, 2017; Ducar y Sohocket, 2018) han empezado a abogar por la incorporación de este tipo de tecnología en la enseñanza y aprendizaje de L2.

El uso de la traducción automática está siendo investigado también dentro de la enseñanza del inglés para usos específicos, como herramienta de ayuda tanto para la lectura (Borsatti y Riess, 2021) como para la escritura (Groves and Mundt, 2015; Kol et al., 2018). En este sentido, una métrica como VERTa puede ayudar a los profesores cuando tengan que recomendar a sus estudiantes un traductor automático entre todos los que están disponibles actualmente, calibrando cuál proporciona la mejor calidad o da mejores resultados dentro de su área de especialidad. Para ello, los profesores deberán contar de entrada con dos textos: un texto en la lengua de origen, que será traducido automáticamente por los sistemas de traducción que se quieran evaluar, y el mismo texto traducido por un humano, que será utilizado como referencia por VERTa. Los textos traducidos automáticamente y la traducción de referencia serán procesados y analizados mediante VERTa, que proporcionará como salida un valor del 0 al 1 para cada una de las traducciones (ver Figura 6), siendo la mejor traducción la más cercana a 1. De esta manera, el profesor podrá elegir aquel traductor que haya obtenido la puntuación más alta y recomendar su uso a los alumnos.

Figura 6
Puntuación final de la traducción



La TA y VERTa también pueden utilizarse para la creación de actividades para el aprendizaje de L2 a través de la traducción, en concreto para la implementación de actividades de comparación y crítica de traducciones. Dada una frase en la lengua origen, el profesor utiliza diferentes sistemas de TA para generar varias traducciones y a continuación las evalúa con VERTa para desestimar aquellas que no muestran una buena calidad y así quedarse solo con las traducciones aceptables. A partir del conjunto filtrado de oraciones, el profesor podrá diseñar un ejercicio para que los alumnos las compararen y discutan las diferentes propuestas de traducción.

Otra aplicación de VERTa tiene que ver con la adaptación que estamos haciendo de la métrica dentro del proyecto de investigación TAGFACT[11], donde la estamos utilizando para medir la similitud semántica textual y la similitud de eventos. El objetivo es que a partir de dos textos de noticias podamos saber si son semánticamente similares. Esta adaptación puede llevarse al campo de la enseñanza de L2 para la creación de ejercicios. Así, tomando como punto

[11] "Del texto al conocimiento: factualidad y grados de certeza en español": http://grial.edu.es/web/es/grial/del-texto-al-conocimiento-factualidad-y-grados-de-certeza-en-espanol-tagfact/

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

de partida un corpus textual paralelo (por ejemplo, artículos periodísticos sobre el mismo tema), VERTa puede utilizarse para comparar estos textos y obtener aquellas frases que son semánticamente similares, es decir aquellas que VERTa ha puntuado con un 1 o un valor cercano a 1. Este primer filtro facilitaría al profesor la selección de estructuras distintas utilizadas para expresar el mismo significado. De esta manera los profesores podrían crear ejercicios para trabajar el uso de la sinonimia y la paráfrasis, dando así respuesta a una habilidad importante en nuestra sociedad, tal y como se indica en el MCER: "Las actividades de mediación de tipo lingüístico, que (re)procesan un texto existente, ocupan un lugar importante en el funcionamiento lingüístico normal de nuestras sociedades" (Consejo de Europa, 2001, p. 15).

Asimismo, VERTa se puede utilizar como una herramienta de ayuda a la evaluación de las traducciones de los alumnos de grados y posgrados de lenguas modernas o traducción e interpretación. La evaluación de la traducción es una tarea difícil, que exige mucho tiempo y muchas veces es cognitivamente agotadora, por eso han ido apareciendo estudios en los que se analiza la viabilidad del uso de las métricas de evaluación de la TA para evaluar las traducciones producidas por los alumnos (Michaud y McCoy, 2013; Chung 2020; Han y Lu, 2021; Munkova et al. 2021). En estos estudios se han usado métricas como BLEU y METEOR, que proporcionan una evaluación cuantitativa basada en una puntuación de similitud. VERTa puede hacer una aportación más allá de la evaluación cuantitativa y proporcionar información sobre aquellos niveles lingüísticos que presentan más errores y, por tanto, se puede usar como una guía para identificar aquellas áreas en las que los alumnos tienen más problemas.

Un ejemplo de esta aplicación se ofrece a continuación, donde se comparan dos frases traducidas por dos alumnos de un curso de traducción con una frase de referencia. La frase original estaba en inglés y se les pidió que la tradujeran al español. A la vez, el profesor ya contaba con la traducción al español del texto inglés, que se utilizó como referencia en VERTa.

- FRASE ORIGEN: *During this short time almost a million people visit the park, located around half an hour from Amsterdam*.
- TRAD. 1: Durante este poco tiempo casi un millón de personas visitan el parque, el cual está alrededor de una hora y media de Ámsterdam.
- TRAD. 2: Durante este breve periodo, casi un millón de personas visitan el parque, situado aproximadamente a media hora de Ámsterdam.
- REF.: Durante este breve periodo, cerca de un millón de personas visitan el parque, situado a media hora de Ámsterdam aproximadamente.

Una vez evaluadas ambas frases mediante VERTa, obtenemos una puntuación de 0,59 para la primera traducción y de 0,87 para la segunda. De entrada, esto ya indica que la segunda traducción es mejor que la primera, pero si vamos al resultado del módulo de n-gramas, vemos que hay partes importantes de las frases que no están alineadas: por una parte, "poco tiempo" (TRAD. 1) y "breve periodo" (REF.) no se han vinculado, puesto que "poco" y "breve" no funcionan como sinónimos exactos en este contexto; por otro lado, "una hora y media" (TRAD. 1) y "media hora" (REF.) tampoco están alineados, ya que el significado es completamente diferente.

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

Con todo, hay que tener en cuenta las limitaciones de VERTa. Al tratarse de una métrica automática que utiliza herramientas y recursos de PLN para analizar los textos, se debe contar con la posibilidad de que haya algunos errores en la evaluación, ya sean producidos por la misma métrica o debidos al análisis automático previo de los textos. Por este motivo, se recomienda el uso de VERTa como una herramienta de ayuda al profesor a la hora de evaluar y no tanto para el aprendizaje autónomo.

Finalmente, en los cursos de traducción e interpretación, VERTa puede utilizarse como una herramienta más dentro del proceso de la traducción, de la postedición, y de la evaluación de la postedición. Según Doherty et al., (2018), aunque la mayoría de los cursos de grado y posgrado incluyen tecnologías de la traducción como la TA o las memorias de traducción, la formación en herramientas para evaluar la calidad de la traducción continúa siendo una asignatura pendiente. En este sentido, una herramienta como VERTa podría utilizarse para formar a los alumnos en métricas de evaluación automática. De esta manera, aprenderían a evaluar un texto traducido automáticamente, a decidir si hay que posteditarlo o no, y a usar la información proporcionada por VERTa como guía para detectar aquellos cambios que sean necesarios.

Conclusiones

En este artículo se ha presentado una detallada descripción de VERTa, una métrica de evaluación de la TA, haciendo especial énfasis en el proceso de análisis y reflexión que se llevó a cabo antes del diseño y la implementación de la métrica, en la información que cada módulo de la métrica proporciona al usuario, y en la utilidad de VERTa para la enseñanza e investigación en L2.

Aunque VERTa es una métrica de evaluación de la TA, se diferencia del resto en que durante su desarrollo se puso especial énfasis en analizar la información lingüística que cada módulo proporcionaba al usuario, con la finalidad de ir más allá de una mera puntuación del segmento traducido y así poder servir como una primera guía cualitativa para detectar los errores de la traducción automática. En consecuencia, VERTa se puede utilizar en el área del aprendizaje, de la enseñanza y de la evaluación del inglés y del español como L2, ya sea para seleccionar herramientas de traducción automática como ayuda a la escritura, o como herramienta para evaluar las traducciones de los estudiantes, o incluso para la elaboración de ejercicios sobre semántica léxica o de reflexión y descripción lingüística.

Agradecimientos

Agradezco al Dr. Jordi Atserias Batalla su trabajo en la implementación de la métrica y a la Dra. Natalia Judith Laso sus recomendaciones, sugerencias y comentarios a la hora de escribir este artículo. Esta investigación se ha llevado a cabo gracias al proyecto TAGFACT, financiado por el Ministerio de Economía, Industria y Competitividad (FFI2017-84008-P).

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

Referencias

Atserias, Jordi; Blanco, Roi; Chenlo, Jose M; Rodriguez, Carlos (2012). FBM-Yahoo at RepLab 2012. En Pamela Forner, Jussi Karlgren, Christia Womser-Hacker y Nicola Ferro (Eds.), *CLEF 2012 Working Notes*. CEUR-WS.org. https://tec.citius.usc.es/ir/pdf/RepLab2012.pdf

Bassnett, Susan (1980). *Translation Studies*. Routledge.

Borsatti, Débora y Riess, Adriana (2021). Using machine translator as a pedagogical resource in English for specific purposes courses in the academic context. *Revista de Estudos da Linguagem*, *29*(2), 829-858. http://dx.doi.org/10.17851/2237-2083.29.2.829-858

Briggs, Neil (2018). Neural machine translation tools in the language learning classroom: Students' use, perceptions and analyses. *Jalt Call Journal*, *14*(1), 3-24. https://files.eric.ed.gov/fulltext/EJ1177331.pdf

Chon, Yuah V.; Shin, Dongkwang; Kim, Go Eun (2021). Comparing L2 learners' writing against parallel machine-translated texts: Raters' assessment, linguistic complexity and errors. *System*, *96*. https://doi.org/10.1016/j.system.2020.102408

Chung, Eun Seon; Ahn, Soojin (2021). The effect of using machine translation on linguistic features in L2 writing across proficiency levels and text genres. *Computer Assisted Language Learning*, https://doi.org/10.1080/09588221.2020.1871029

Chung, Hye-Yeon (2020). Automatic Evaluation of Human Translation: BLEU vs. METEOR. *Lebende Sprachen*, *65*(1), 181-205. https://doi.org/10.1515/les-2020-0009

Ciaramita, Massimiliano; Altun, Yasemin (2006). Broad-Coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger. En Dan Jurafsky y Eric Gaussier (Eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 594–602). Association for Computational Linguistics. https://aclanthology.org/W06-1670.pdf

Comelles, Elisabet (2015). *Automatic Machine Translation Evaluation: A Qualitative Approach*. [Tesis Doctoral, Universitat de Barcelona]. http://diposit.ub.edu/dspace/bitstream/2445/65906/1/ECP_PhD_THESIS.pdf

Comelles, Elisabet; Arranz, Victoria; Castellón, Irene (2017). Guiding Automatic MT Evaluation by Means of Linguistic Features. *Digital Scholarship in the Hunanities*, *32*(4), 761-778. https://doi.org/10.1093/llc/fqw042

Comelles, Elisabet; Atserias, Jordi (2019). VERTa: a linguistic approach to automatic machine translation evaluation. *Language Resources and Evaluation, 53*, 57-86. https://doi.org/10.1007/s10579-018-9430-2

Consejo de Europa (2001). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación (MCER)*. MECD-Anaya. https://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/cvc_mer.pdf

Denkowski, Michael; Lavie, Alon (2014). Meteor Universal. Language Specific Translation Evaluation for Any Target Language. En Ondřej Bojar, Christian Buck, Christian Federman, Barry Haddow, Philipp Koehn, Christoff Monz, Matt Post y Lucia Specia (Eds.), *Proceedings of the Ninth Workshop on Statistical Machine Translation* (pp. 376-380). Association for Computational Linguistics. https://aclanthology.org/W14-3348/

Doddington, George (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. En Mitchell Marcus (Ed.), *Proceedings of the 2nd International Conference on Human Language Technology* (pp. 138-145). Morgan Kaufmann Publishers Inc. https://dl.acm.org/doi/10.5555/1289189.1289273

Doherty, Stephen; Moorkes, Joss; Gaspari, Federico; Castilho, Sheila (2018). On Education and Training in Translation Quality Assessment. En Joss Moorkens, Sheila Castilho, Federico Gaspari y Stephen Doherty (Eds.), *Translation Quality Assessment. From Principles to Practice* (pp. 95-106). Springer. https://doi.org/10.1007/978-3-319-91241-7_5

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

Ducar, C; Shocket, D. H (2018). Machine translation and the L2 classroom: Pedagogical solutions for making peace with Google translate. *Foreign Language Annals*, *51*(4), 779-795. https://doi.org/10.1111/flan.12366

Enríquez, Vanessa; Austermühl, Frank; Sánchez, Marina (2019). Computer-assisted L2 learning and translation (CAL2T). En Sara Laviosa y Maria González-Davis (Eds.), *The Routledge Handbook of Translation and Education* (1st ed.) (pp. 278-299). Routledge. https://doi.org/10.4324/9780367854850

Farrús, Mireia; Costa-Jussà, Marta R.; Mariño, José B.; Fonollosa, José A. R. (2010). Linguistic-based Evaluation Criteria to identify Statistical Machine Translation Errors. En Francois Yvon y Viggo Hansen (Eds.), *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation. https://aclanthology.org/2010.eamt-1.12/

Fellbaum, Charles. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Giménez, Jesús; Márquez, Lluís (2010). Asiya: An Open Toolkit for Automatic Machine Translation (Meta-) Evaluation. *The Prague Bulletin of Mathematical Linguistics*, *94*, 77-86.

González, Meritxell; Barrón-Cedeño, Alberto; Márquez, Lluís (2014). IPA and STOUT: Leveraging Linguistic and Source-based Features for Machine Translation Evaluation. En Ondřej Bojar, Christian Buck, Christian Federman, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post y Lucia Specia (Eds.), *Proceedings of the Ninth Workshop on Statistical Machine Translation* (pp.394-401). Association for Computational Linguistics. https://aclanthology.org/W14-3351/

Groves, Michael; Mundt, Klaus. (2015). Friend or foe? Google Translate in language for academic purposes. *English for Specific Purposes, 37*, 112-121. https://doi.org/10.1016/j.esp.2014.09.001

Gupta, Rohit; Orăsan, Constantin; van Genabith, Josef (2015). ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks. En Lluís Márquez, Chris Callison-Burch y Jian Su (Eds.) *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1066-1072). Association for Computational Linguistics. https://aclanthology.org/D15-1124/

Guo, Yinuo; Hu, Junfeng (2019). Meteor++ 2.0: Adopt Syntactic Level Paraphrase Knowledge into Machine Translation Evaluation. En Ondřej Bojar, Rajen Chatterjee, Christian Federman, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martin, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi y Karin Verspoor (Eds.), *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)* (pp. 501-506). Association for Computational Linguistics. https://aclanthology.org/W19-5357

Hachey, Ben; Radford, Will; Curran, James R. (2011). Graph-based named entity linking with Wikipedia. En *Proceedings of the 12th International conference on Web information system engineering*, 213-226, Springer-Verlag. Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/978-3-642-24434-6_16

Han, Chao; Lu, Xiaolei (2021). Can automated machine translation evaluation metrics be used to assess students' interpretation in the language learning classroom? *Computer Assisted Language Learning*, https://doi.org/10.1080/09588221.2021.1968915

Jimenez-Crespo, Miguel A. (2017). The role of translation technologies in Spanish language learning. *Journal of Spanish Language Teaching*, 4(2), 181-193. https://doi.org/10.1080/23247797.2017.1408949

Joty, Shafiq; Guzmán, Francisco; Márquez, Lluís; Nakov, Preslav (2014). DiscoTK: Using Discourse Structure for Machine Translation Evaluation. En Ondřej Bojar, Christian Buck, Christian Federman, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post y Lucia Specia, (Eds.), *Proceedings of the Ninth Workshop on Statistical Machine Translation* (pp. 402-408). Association for Computational Linguistics. https://aclanthology.org/W14-3352/

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

Kol, Sara; Schcolnik, Miriam; Spector-Cohen, Elana (2018). Google Translate in Academic Writing Courses? *The EuroCALL Review, 26*(2), 50-57, https://doi.org/10.4995/eurocall.2018.10140

Laviosa, Sara; González-Davies, Maria (Eds.) (2019). *The Routledge Handbook of Translation and Education* (1st ed.). Routledge. https://doi.org/10.4324/9780367854850

Lee, Sangmin-Michelle (2022). An investigation of machine translation output quality and the influencing factors of source texts. *ReCALL, 34*(1), 81-94. https://doi.org/10.1017/S0958344021000124

Lin, Chin-Yew; Och, Franz Josef (21-26 de julio de 2004). Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. *Proceedings of the 42nd Annual Meeting of the Association for Computationl Linguistics* (ACL-04). https://aclanthology.org/P04-1077

Liu, Ding; Gildea, Daniel (2005). Syntactic Features for Evaluation of Machine Translation. En Goldstein, Jade; Lavie, Alon; Lin, Chin-Yew; Voss, Clare (Eds.), *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization* (pp. 25-32). Association for Computational Linguistics. https://www.cs.rochester.edu/~gildea/pubs/liu-gildea-eval05.pdf

Lo, Chi-kiu (2017). Meant 2.0: Accurate semantic MT evaluation for any output language. En Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federman, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn y Julia Kreutzer (Eds.), *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers* (pp. 589-597). Association for Computational Linguistics. https://aclanthology.org/W17-4767.pdf

Ma, Q Qingsong; Graham, Yvette; Wang, Shugen; Liu, Qun (2017). Blend: a novel combined mt metric based on direct assessment casict-dcu submission to wmt17 metrics task. En Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federman, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn y Julia Kreutzer (Eds.), *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers* (pp. 598-603). Association for Computational Linguistics. https://aclanthology.org/W17-4768

Manning, Christopher; Surdeanu, Mihai; Bauer, John; Finkel,Jenny; Bethard, Steven; McClosky, David (2014). The Stanford CoreNLP Natural Language Processing Toolkit. En Kalina Bontcheva y Jinggo Zhu, (Eds.), *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55-60). Association for Computational Linguistics. https://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf

Michaud, Lisa N.; McCoy, Patricia Ann (2013). Applying Machine Translation Metrics to Student-Written Translations. En Joel Tetreault, Jill Burstein y Claudia Leackock (Eds.), *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 306-311). Association for Computational Linguistics. https://aclanthology.org/W13-1740.pdf

Munkova, Dasa; Munk,, Michal; Benko, L'ubomír; Hajek, Petr (2021). The role of automated evaluation techniques in online professional translator training. *PeerJ Computer Science*, 7:e706, https://doi.org/10.7717/peerj-cs.706

Oliver, Antoni; Climent, Salvador (2012). Building WordNets by machine translations of sense tagged corpora. En Christian Fellbaum y Piek Vossen (Eds.), *Proceedings of the Global WordNet Conference 2012* (pp. 232-239). Tribun EU.

Padró, Lluís; Stanilovsky, Evgeny (2012). FreeLing 3.0: Towards Wider Multilinguality. En Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk y Stelios Piperidis (Eds.), *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022

ISSN 2696-676X

Papineni, Kishore; Roukos, Salim; Ward, Todd; Zhu, Wei-Jing (2002). BLEU: a method for automatic evaluation of machine translation. En Pierre Isabele, Eugene Charniak y Dekang Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311-318). Association for Computational Linguistics. https://aclanthology.org/P02-1040.pdf

Popović, Maja (2012). Class error rates for evaluation of machine translation output. En Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut y Lucia Specia (Eds.), *Proceedings of the 7th Workshop on Statistical Machine Translation* (pp. 71-75). Association for Computational Linguistics. https://aclanthology.org/W12-3106/

Popović, Maja (2015). chrF: character n-gram F-score for automatic MT evaluation. En Ondřej Bojar, Rajen Chatterjee, Christian Federman, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva y Pavel Pecina (Eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation* (pp. 392-395). Association for Computational Linguistics. https://aclanthology.org/W15-3049

Popović, Maja (2017). chrF++: words helping character n-grams. En Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federman, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Julia Kreutzer (Eds.), *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers* (pp. 612-618). Association for Computational Linguistics. https://aclanthology.org/W17-4770

Yoo-Jean Lee (2021). Still taboo? Using machine translation for low-level EFL writers, *ELT Journal, 75*(4), 432-441. https://doi.org/10.1093/elt/ccab018

Apellidos, Nombre; Apellidos, Nombre, y Apellidos, Nombre (año). Título del artículo: subtítulo. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr01 (1-19). http://doi.org/10.1344/teisel.v2.39307

 | Recepción: 22/03/2022 | Revisión: 04/07/2022| Aceptación: 19/09/2022| Publicación: 20/09/2022