Propuesta recibida en octubre de 2018 y aceptada para publicación en noviembre de 2019.

Estrategia multidimensional para la selección de candidatos de traducción automática para posedición

Multidimensional strategy for the selection of machine translation candidates for post-editing

Ona de Gibert
Universidad del País Vasco UPV/EHU
ona.degibert@ehu.eus

Nora Aranberri
Grupo IXA, Universidad del País Vasco UPV/EHU
nora.aranberri@ehu.eus

Resumen

Una integración eficiente de un sistema de traducción automática (TA) en un flujo de traducción conlleva la necesidad de distinguir entre oraciones que se benefician de la TA y las que no antes de que pasen a manos del traductor. En este trabajo, cuestionamos el uso por separado de las dimensiones de esfuerzo de posedición de Krings (2001) para clasificar oraciones en aptas para traducir o poseditar al entrenar modelos de predicción y abogamos por una estrategia multidimensional. A partir de una tarea de posedición en un escenario real, se recogen mediciones de los tres parámetros de esfuerzo, a saber, tiempo, tasa de palabras poseditadas, y percepción del esfuerzo, como representativos de las tres dimensiones (temporal, técnica y cognitiva). Los resultados muestran que, a pesar de que existen correlaciones entre las mediciones, los parámetros difieren en la clasificación de un número elevado de oraciones. Concluimos que la estrategia multidimensional es necesaria para estimar el esfuerzo real de posedición.

Palabras clave

traducción automática, esfuerzo de posedición, estimación de calidad

Abstract

An efficient integration of a machine translation (MT) system within a translation flow entails the need to distinguish between sentences that benefit from MT and those that do not before they are presented to the translator. In this work we question the use of **?** post-editing effort dimensions separately to classify sentences into suitable for translation or for post-editing when training predictions models and propose a multidimensional strategy instead. We collect measurements of three effort parameters, namely, time, number of post-edited words and perception of effort, as representative of the three dimensions (temporal, technical and cognitive) in a real post-editing task. The results show that, although there are correlations between the measurements, the effort parameters differ in the classification of a considerable number of sentences. We conclude that the multidimensional strategy is necessary to estimate the overall post-editing effort.

Keywords

machine translation, post-editing effort, quality estimation

1 Introducción

La integración de un sistema de traducción automática (TA) en un flujo de traducción de un proveedor de servicios lingüísticos conlleva la toma de varias decisiones. Por una parte, se encuentran aquellos aspectos de índole más técnica que abarcan desde la compatibilidad de herramientas y formatos, hasta cuestiones de rendimiento de las máquinas. Por otra parte, se han de abordar asuntos relacionados con la aplicación y uso que se quieran hacer de las propuestas de TA. Existen varias posibilidades a la hora de emplear dichas propuestas, entre otras:

- Presentar las propuestas de TA al traductor para todos los segmentos.
- Presentar las propuestas de TA al traductor siempre que se consideren de calidad adecuada para posedición.
- Presentar las propuestas de TA al traductor únicamente cuando no existan propuestas de una memoria de traducción (MT) a partir de un umbral preestablecido.
- Presentar las propuestas de TA al traductor únicamente cuando no existan propuestas de MT con un umbral preestablecido y se consideren de calidad adecuada para posedición.

La elección de una u otra opción dependerá de la calidad global del sistema de TA, así como de las características y los requisitos de los

DOI: 10.21814/lm.11.2.277

encargos de traducción. Claramente, la primera opción sería la más sencilla de implementar. Bastaría con traducir el texto completo con el sistema de TA y entregárselo al traductor para su posedición. Para ser eficiente, este escenario requeriría de propuestas de TA de calidad constante por encima de un umbral determinado, debido a que al no considerar la ayuda de segmentos de MT el traductor deberá trabajar continuamente con las propuestas de TA independientemente de su calidad.

Con respecto a las opciones de implementación descritas, aquellas que consideran la calidad de los segmentos que se proponen al traductor, independientemente de si proceden de una MT o de un sistema de TA, plantean un escenario óptimo para el uso eficiente de las tecnologías disponibles (Parra Escartín et al., 2017). Por una parte, se facilita la reutilización de segmentos ya traducidos y validados anteriormente, y por otra, se optimiza la tarea de realizar nuevas traducciones, ya sea con la ayuda de propuestas de TA en caso de que resulten adecuadas para una posedición eficiente, ya sea sin ellas, en cuyo caso el traductor formulará su propuesta sin necesidad de trabajar con propuestas de TA de mala calidad.

Sin embargo, este escenario requiere una implementación más compleja. En primer lugar, se debe establecer el umbral de coincidencia parcial para filtrar las propuestas de la MT. Aun siendo ésta una práctica habitual, para establecer el umbral idóneo en el escenario que nos ocupa, se deberá comparar la aportación de un segmento de MT con la de una propuesta de TA además de con la traducción manual (Forcada & Sánchez-Martínez, 2015). Por lo tanto, este umbral no será necesariamente el utilizado en un flujo sin TA. En segundo lugar, se debe establecer una manera de filtrar aquellas propuestas de TA que se presentarán al traductor y aquellas que se descartarán a favor de la traducción manual. Estos filtrados, y el segundo en particular, conllevan una tarea compleja, ya que requieren entrenar modelos automáticos de predicción que se han de desarrollar en una etapa previa al trabajo de traducción con datos reales del esfuerzo de posedición, e integrarlos en el flujo de traducción para que decidan automáticamente si una oración deberá presentarse al traductor para traducir o, junto con su traducción automática, para poseditar.

En este trabajo, nos centramos en esa etapa previa de entrenamiento de modelos automáticos de predicción, en concreto, en estudiar el tipo de información que se debería utilizar para entrenar dichos predictores. Tomando como base las tres dimensiones de esfuerzo de posedición identificadas por Krings (2001), partimos de la hipótesis de que el esfuerzo real de posedición no estará reflejado en su totalidad en dichos modelos si no se consideran las tres dimensiones, ya que por separado, las dimensiones pueden no alcanzar a medir parte del esfuerzo total incluido en la tarea. Por ello, proponemos el uso de una estrategia multidimensional que combine información referente a las tres dimensiones. Los datos de esfuerzo temporal, técnico y cognitivo se podrían integrar en el proceso de aprendizaje de los modelos de diversas maneras. Por una parte, se podrían incluir como características para entrenar un clasificador que prediga si una oración se debería poseditar o traducir. Por otra, se podrían utilizar como objetivos a predecir para (1) crear tres modelos que predigan cada dimensión de manera independiente, que se combinarían posteriormente para decidir si traducir o poseditar, (2) crear un único modelo que combine las tres dimensiones de esfuerzo como objetivos, o (3) una versión mixta en la que algunas dimensiones sean características y otras objetivos a predecir.

Dado que es posible recopilar información sobre las tres dimensiones durante un trabajo de posedición rutinario, realizamos un estudio preliminar que nos permite una primera aproximación al estudio de la relación entre las dimensiones citadas. Un primer análisis muestra que, tras clasificar manualmente las oraciones de un corpus, los conjuntos de decisiones (poseditar o traducir) obtenidos para cada dimensión varían. Este hecho parece indicar la necesidad de considerar las tres dimensiones para entrenar modelos de predicción, de lo contrario, podríamos no estar considerando el esfuerzo real de posedición.

2 Trabajos relacionados

Son diversos los trabajos que se han centrado en el desarrollo de modelos automáticos de estimación que tratan de predecir la forma más eficiente de editar una oración, bien traduciéndola bien poseditándola (He et al., 2010; Callison-Burch et al., 2012; Parra Escartín & Arcedillo, 2015; Bojar et al., 2017). En general, estos modelos se crean a partir de un proceso de aprendizaje automático en cuyo entrenamiento se utiliza información sobre esfuerzo de posedición real. Es precisamente este esfuerzo de posedición lo que determina si es más eficiente traducir o poseditar una oración.

Una de las mayores dificultades de este proceso es precisamente la medición del esfuerzo de posedición. En el estudio de referencia sobre posedición, Krings (2001) afirma que el esfuerzo de po-

sedición está compuesto por tres dimensiones, a saber, la dimensión temporal, la dimensión técnica y la dimensión cognitiva. Atendiendo en mayor o menor medida a esta teoría, estudios previos han utilizado diferentes parámetros para representar el esfuerzo de posedición (véase la Figura 1). Algunos trabajos se han basado en recopilar la percepción del esfuerzo como parámetro de la dimensión cognitiva, es decir, se ha recopilado la percepción de un evaluador sobre lo difícil que resultaría poseditar una propuesta de TA en una escala del 1 al 5 al analizar la propuesta de traducción y una versión de posedición realizada por un traductor, sin que tuviera que completar la posedición él mismo previamente (Specia et al., 2010; Felice & Specia, 2012; Shah et al., 2015). Moorkens et al. (2015), en cambio, han estudiado la posibilidad de utilizar el tiempo de fijación de la mirada como parámetro para esta dimensión.

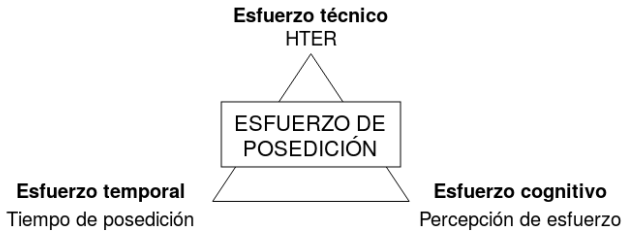

Figura 1: Las tres dimensiones del esfuerzo de posedición según Krings (2001) junto con posibles parámetros de medición

Otros trabajos consideran que el parámetro que se debería utilizar para medir el esfuerzo y clasificar las oraciones es el tiempo de ejecución (Parra Escartın & Arcedillo, 2015). En estos trabajos se llevan a cabo tests de productividad que comparan el ratio de segundos por palabra requeridos para traducir y poseditar una serie de oraciones (Plitt & Masselot, 2010). Esta información les permite concluir cuál de las dos aproximaciones, traducir o poseditar, es más rápida y clasificar las oraciones en la clase correspondiente. Si bien es cierto que en las tareas compartidas de estimación de calidad de 2013 y 2014 varios trabajos se centraron en crear modelos de estimación de tiempo, no se conoce ningún trabajo aplicado que recojan estos modelos y los utilice para desarrollar modelos de estimación binarios.

A pesar de que existen trabajos que consideran la percepción de esfuerzo y el tiempo de ejecución como posibles indicadores del esfuerzo de posedición, el parámetro más extendido y utilizado en las tareas compartidas de estimación de calidad (2012–2018) (Callison-Burch et al., 2012; Bojar et al., 2017) es la métrica HTER (Snover et al., 2009). Esta métrica es una variación de la tasa de edición de la traducción (TER, del inglés translation error rate) (Snover et al., 2006) que calcula de manera automática el número de ediciones (inserciones, eliminaciones, sustituciones y reordenaciones) que un traductor realiza en la propuesta de TA para que coincida con una traducción de referencia. En HTER, los traductores crean una nueva traducción de referencia o versión de posedición propia que sea fiel al original en términos de fluidez y significado realizando el número mínimo de ediciones. Después, se calcula el número de ediciones necesarias para transformar la propuesta de TA en la versión de posedición final.

Una rápida revisión de la literatura muestra que se ha experimentado con diferentes parámetros para medir el esfuerzo de posedición, si bien la métrica HTER es el estándar en el área, promovida por su supuesta correlación con la percepción de esfuerzo humana (Snover et al., 2006; Specia & Farzindar, 2010) y, sobre todo, por tratarse del parámetro más sencillo de obtener. Otra de las conclusiones que sacamos es que los estudios que tratan de medir o representar el esfuerzo de posedición se limitan al uso de un solo parámetro de esfuerzo, dando por sentado que la información de una única dimensión es suficiente para representar el esfuerzo real, si bien esto no está demostrado.

Si, como parece apuntar este trabajo, el resultado de la medición del esfuerzo de posedición varía dependiendo de la dimensión utilizada, aquellos flujos de traducción que únicamente consideren una de ellas podrían llevar a una toma de decisiones subóptimas. Por ejemplo, si una empresa utiliza información temporal para entrenar un modelo de clasificación binario, es muy probable que el modelo siga presentando a la traductora oraciones para poseditar que supongan gran esfuerzo técnico o cognitivo. Asimismo, si una empresa utiliza la dimensión técnica como criterio para clasificar oraciones el aptas para traducir o poseditar, podría estar agrupando oraciones que requieran un esfuerzo cognitivo o temporal diverso. Creemos que el uso combinado de la información representativa de las tres dimensiones de esfuerzo podría llevar a una estimación más precisa del esfuerzo real de posedición, ofreciendo la oportunidad de crear entornos de trabajo óptimos y tarifas más justas, entre otros.

3 Propuesta multidimensional

Tal y como refleja el análisis previo, el parámetro más común utilizado para clasificar una oración como adecuada para poseditar o idónea pa-

ra traducir es la medida HTER. Podemos argumentar que esta medida viene a reflejar, si bien de manera incompleta, la dimensión de esfuerzo técnico, ya que considera las ediciones necesarias para transformar la propuesta de TA en un segmento de calidad requerida. Decimos que recoge el esfuerzo técnico de manera parcial porque el recuento de ediciones llevado a cabo por la métrica no considera todas las ediciones realizadas durante el proceso de posedición de la oración sino que calcula las ediciones partiendo de la versión final de posedición, es decir, no tiene en cuenta las rectificaciones ni los cambios realizados antes de fijar la versión final de posedición.

Parece innegable que esta manera de medir el esfuerzo es considerablemente limitada, ya que excluye la dimensión temporal así como la cognitiva. Sin embargo, en línea con las argumentaciones de algunos trabajos que tratan de estudiar las relaciones entre las dimensiones de esfuerzo, así como de los distintos parámetros que se utilizan para representarlos, se podría inferir que, de manera indirecta, la métrica también reconoce parte del esfuerzo temporal, puesto que cuanto mayor sea el número de cambios, más tiempo requiere la posedición. Sin embargo, este razonamiento sólo es cierto si todas las ediciones requieren el mismo tiempo y, tal y como apunta Temnikova (2010), ciertas ediciones suponen un esfuerzo cognitivo mayor, lo que hace suponer que requerirán mayor tiempo. De hecho, Koponen (2012) estudia la relación entre la dimensión técnica y cognitiva, y concluye que HTER y la percepción de esfuerzo pueden dar resultados diferentes. A partir de oraciones traducidas automáticamente del inglés al español, la autora obtiene puntuaciones sobre la percepción del esfuerzo de posedición para representar la dimensión de esfuerzo cognitivo, y calcula la métrica TER para atender a la dimensión de esfuerzo técnico. Los resultados no siempre son equiparables. Lo mismo ocurre en el estudio de Moorkens et al. (2015) en el cual se concluye que la percepción de esfuerzo no es equiparable al esfuerzo real determinado por el tiempo de posedición, el tiempo de fijación de mirada y TER. También podría ocurrir, asimismo, que el esfuerzo de diferentes tipos de ediciones se compense entre ellos en cada segmento y siga habiendo buena correlación.

Podemos extraer dos conclusiones de los trabajos que estudian la relación entre las dimensiones de esfuerzo. Primero, que los parámetros que utilizamos para abordar las diferentes dimensiones de esfuerzo, si bien están enfocadas a una de ellas de manera más explícita, podrían atender a las demás en cierto grado. Segundo, que los resultados de los distintos parámetros no siempre son equiparables. Por tanto, al utilizar únicamente un parámetro para representar el esfuerzo de posedición y al centrarse éste especialmente en una de las tres dimensiones, es posible que estemos descuidando parte del esfuerzo real. Es evidente que carecemos de resultados concluyentes y que es necesario seguir investigando en este campo. Curiosamente, los trabajos que se centran en desarrollar modelos de predicción para el esfuerzo de posedición no han abordado este aspecto, a excepción de Aranberri & Pascual (2018), quienes proponen la inclusión de varios parámetros de esfuerzo como características de entrenamiento para modelos binarios.

En este trabajo, exploramos y comparamos las mediciones de esfuerzo de las tres dimensiones por separado. En esta primera aproximación, recogemos información sobre el tiempo de posedición, la percepción de esfuerzo según el poseditor y HTER durante un proceso de posedición rutinaria, y tras estudiar los resultados, proponemos utilizar los datos derivados de las tres mediciones de manera unificada para entrenar modelos de clasificación automáticos que, una vez implementados en el flujo de traducción, separen las oraciones bien para poseditar o traducir.

4 Ejemplo de aplicación

En esta sección exponemos la metodología seguida para la obtención de una clasificación binaria (poseditar, traducir) manual para estudiar la relación entre los resultados obtenidos para las distintas dimensiones. En primer lugar, se describe el corpus, atendiendo a los textos incluidos. En segundo lugar, se describe el proceso de recopilación de los datos correspondientes a los parámetros de las tres dimensiones de esfuerzo, es decir, el tiempo de posedición, la percepción de esfuerzo y los valores de HTER. Finalmente se establecen los umbrales para crear los subconjuntos de oraciones óptimas para traducir o para poseditar para cada parámetro, y se analiza la relación entre dichos subconjuntos antes de presentar la selección final.

4.1 Compilación del corpus

Describimos en primer lugar los textos incluidos en el corpus, atendiendo a los tres factores que, según Bernth & Gdaniec (2001), determinan la calidad de una traducción automática y es, por lo tanto, imprescindible exponerlos para tener una visión completa del corpus: el sistema de TA, el par de lenguas y el dominio de los tex-

tos. Estos factores fueron predeterminados por el contexto real para el cual se realizó este estudio.

Los documentos recopilados son textos técnicos que pertenecen al área de la construcción. Específicamente, se trata de documentos de instalación y mantenimiento de equipamiento. Debido a la tipología textual, los textos incluyen un alto número de repeticiones, listas y enumeraciones, entre otros. El subcorpus anotado para este estudio consta de 7 textos con un total de 509 oraciones y 6.542 palabras (veáse el Cuadro 1).

Texto	# palabras	# frases	palabras/frase
1	360	25	14,40
2	663	74	8,72
3	444	53	8,38
4	726	70	10,37
5	1.198	95	12,61
6	246	17	14,47
7	2.905	174	16,70
Total	6.542	509	12,83

Cuadro 1: Descripción del corpus

Los textos están redactados originalmente en español y la empresa recibe el encargo de traducirlos al inglés (entre otras lenguas). Por lo tanto, a diferencia de estudios de investigación anteriores que se centran en el inglés como lengua de origen (Felice & Specia, 2012; Hardmeier, 2011), este trabajo se centra en el español como lengua de origen y el inglés como lengua de destino.

La traducción automática de los textos del español al inglés se obtuvo con el sistema de TA utilizado por la empresa dentro de su flujo de producción. Se trata de un sistema de TA neuronal (arquitectura de codificador–decodificador con mecanismos de atención (Bahdanau et al., 2014)) desarrollado específicamente para la empresa y personalizado para el respectivo cliente.

4.2 Compilación de parámetros de esfuerzo: percepción, tiempo y HTER

Los datos para los tres parámetros de esfuerzo de posedición se recogieron durante una tarea de posedición en la cual se extrajo información real de los parámetros que nos atañen. Es precisamente durante el trabajo de posedición cuando es posible medir el tiempo de ejecución. A su vez, la fiabilidad sobre la percepción del esfuerzo de posedición aumenta si la traductora realiza dicho trabajo, y no simplemente imagina el esfuerzo que conllevaría. Es necesario subrayar que esta técnica es dependiente de la capacidad de comunicación del evaluador, por lo tanto, en este trabajo hablaremos de percepción *comunicada* del esfuerzo. Finalmente, para calcular los valores de HTER es imprescindible contar con una versión poseditada de la propuesta de TA, por lo que se calcularon una vez terminada la tarea. En los siguientes párrafos se describe el proceso seguido para la obtención de las mediciones de cada uno de los parámetros de esfuerzo.

Es evidente que sería necesario contar con múltiples traductoras que completaran varias tareas extensas de forma paralela para poder recopilar datos suficientes con los que obtener resultados concluyentes. En este trabajo en concreto, sin embargo, contamos con recursos limitados. Tuvimos a nuestra disposición tres traductoras profesionales internas durante un tiempo limitado. Según la encuesta realizada, las tres traductoras son mujeres, nacieron entre 1987 y 1991 y han estado trabajando en la industria de la traducción de 2 a 4 años. Todas cuentan con estudios universitarios de traducción, con especialidad científica, técnica y literaria, y trabajan con inglés y español como lenguas de trabajo. Con respecto a la posedición, todas tenían experiencia en este campo. Sin embargo, su actitud hacia el uso de TA para posedición es algo negativa y afirman que la traducción manual es más efectiva, ya que resulta más fácil y más rápida.

Debido a la disponibilidad limitada de las traductoras, el trabajo de posedición se dividió en tres partes comparables teniendo en cuenta los límites y la longitud de los textos, y cada traductora poseditó una de ellas. Las tres partes contaban aproximadamente con el mismo número de palabras (∼2.180) (véase el Cuadro 3).

El trabajo de posedición se realizó en la plataforma en línea Matecat (Federico et al., 2014), la cual nos permitió recopilar información para las tres dimensiones de manera sencilla y relativamente precisa, específicamente, el tiempo total de edición para cada oración, la percepción comunicada del esfuerzo y HTER.

Se prepararon las tareas para cada traductora por separado de manera que cada una accedía al texto original y a la traducción automática obtenida por el sistema de TA mencionado anteriormente para su parte del trabajo. Su trabajo, por lo tanto, consistió en poseditar la TA para lograr una traducción de buena calidad. Con el objetivo de garantizar un trabajo coherente y lo más homogéneo posible, se les proporcionó una guía de posedición con pautas específicas (ver Anexo I). No se restringió ni el tipo ni el número de ediciones posibles, pero se les pidió que modificaran la propuesta de TA lo mínimo posible.

Valor	Descripción	
1	Sin sentido	La traducción era totalmente incomprensible.
2	No utilizable	La traducción contenía tantos errores que claramente hubiera sido más rápido traducir.
3	Neutral	La traducción contenía bastantes errores, no está claro qué hubiera sido más rápido.
4	Utilizable	La traducción contenía algunos errores, pero sería más útil poseditar.
5	Muy buena	La traducción era correcta o casi correcta.

Cuadro 2: Escala de valoración de la percepción de esfuerzo de posedición

Además de poseditar su parte correspondiente, las traductoras calificaron el esfuerzo que les había supuesto el trabajo realizado con cada oración. Seguimos la metodología utilizada por Lacruz et al. (2014), quienes piden al evaluador que califique la idoneidad de un segmento para posedición después de haberlo editado, según una escala del 1 al 5 (véase el Cuadro 5). Ambas tareas se realizaron a la par para atenuar su dificultad y baja fiabilidad descrita en estudios anteriores (Callison-Burch et al., 2012). Se pidió a las traductoras que tras realizar la posedición añadieran al final de cada segmento la puntuación que considerasen oportuna. Esta opción ofrecía una manera sencilla de combinar las tareas y extraer dichas puntuaciones en un paso posterior. De esta manera conseguimos que la percepción de la dificultad se capturase inmediatamente después de editar cada segmento, lo más inalterada posible. Sin embargo, esta opción puede llegar a desvirtuar en cierto grado la medición del tiempo en caso de que la toma de decisión de las traductoras varíe de manera significativa para los distintos segmentos. Para minimizar este riesgo se pidió a las traductoras que limitaran el tiempo de decisión lo máximo posible. Es importante mencionar que el 1 % de las oraciones (5 oraciones en total) carecían de puntuación de esfuerzo. Estos valores fueron reemplazados por la mediana de los valores totales. Así, se les asignó un 5, ya que más de la mitad de las traducciones recibieron un 5.

La medición del tiempo fue sencilla desde el punto de vista técnico, ya que Matecat calcula el tiempo que la traductora emplea en cada segmento. Este cálculo se realiza acumulando el tiempo que cada segmento está activo, es decir, que el cursor se encuentra en la celda que corresponde a la traducción de un segmento concreto. Esta opción permite a la traductora retomar un segmento cuantas veces sea necesario y acumular el tiempo correspondiente. Es cierto que una traductora puede leer y considerar el trabajo realizado en un segmento mientras otro está activado, lo cual muestra una limitación de la herramienta para este tipo de mediciones. Sin embargo, las traductoras fueron informadas de la distorsión que este hecho podía introducir en los datos a fin de minimizarla. Por lo tanto, el tiempo dispuesto para la posedición de cada oración se obtuvo de manera automática tras la finalización de los trabajos.

A pesar de que existen herramientas que pueden capturar todas las ediciones realizadas durante el proceso de posedición (Aziz et al., 2012), y por consiguiente, recopilar el esfuerzo técnico de manera más precisa, decidimos centrar nuestro estudio en la métrica HTER, ya que es la más extendida tanto en trabajo de investigación como en la práctica industrial. Tal y como se ha indicado en la sección anterior, HTER es una métrica que calcula, de manera automática, el número de ediciones necesarias para transformar una propuesta de TA en una oración de calidad adecuada. Para ello, la métrica requiere el texto producido por el sistema de TA, así como una versión final de éste. En nuestro caso, calculamos el HTER tras haber completado el trabajo de posedición con la versión de TA presentada a las traductoras y las posediciones creadas por ellas. Este cálculo nos informa del número de ediciones necesario en cada una de las oraciones de la tarea.

Examinemos el trabajo de las traductoras individualmente. Podemos ver un resumen de su labor en el Cuadro 3. La traductora 1 trabajó con el conjunto que ofrecía mayor variación de oraciones, ya que incluía oraciones de cuatro textos diferentes. El tiempo medio empleado por oración fue de 31 segundos (0,57 min). Resultó ser la más rápida. Asimismo, observamos que es precisamente esta traductora quien modifica la traducción automática en menor proporción, ya que su promedio de HTER global es el más bajo con un valor de 6,83. Para ella, la valoración media de esfuerzo de posedición es de 4,66.

La traductora 2 abordó oraciones de dos textos diferentes. El tiempo medio empleado por oración es de 1 minuto. Su promedio de HTER es de 14,90 ediciones por segmento, lo cual la sitúa como la traductora que introduce el mayor número de cambios durante la posedición. La valoración de esfuerzo se sitúa en 4,56, aunque mínimamente, por debajo de la valoración asignada por sus compañeras a sus respectivos trabajos.

Traductora	# textos	# frases	# palabras	tiempo	seg/palabra	seg/frase	HTER	percepción comunicada
Traductora 1	4	222	2.168	02h:07m	3,5	34	6,83	4,66
Traductora 2	2	150	2.173	02h:34m	4,3	62	14,90	4,56
Traductora 3	1	137	2.198	04h:59m	8,2	131	10,14	4,66

Cuadro 3: Descripción de la labor de las traductoras en la tarea de posedición

La traductora 3 contó con oraciones de un solo texto. El tiempo medio empleado por oración es de 2,2 minutos. Se trata de la traductora más lenta. Su promedio de HTER es de 10,14. Casualmente, la valoración de esfuerzo es exactamente la misma que la de la Traductora 1: 4,66.

Los datos reportados dejan en evidencia que a pesar de las medidas tomadas para dividir la tarea de forma uniforme, la tarea y proceso de posición de cada traductora difiere, sobre todo en lo referente a la longitud de las oraciones poseditadas y la velocidad de trabajo. Curiosamente, podemos observar cómo, independientemente del tiempo empleado o del número de textos abordados, la percepción comunicada de esfuerzo es consistente entre las tres traductoras. Su valoración indica que el sistema de TA proporciona, en general, un número considerable de oraciones aptas para posedición. Estos resultados están en yuxtaposición directa con los resultados obtenidos en la encuesta. Al preguntarles respecto a la tarea específica de PE que llevaron a cabo, las traductoras consideraron que la TA no era útil, si bien una de ellas afirmó que era bastante precisa. De acuerdo con esto, podemos decir que la opinión de las traductoras sobre el uso de TA para posedición es claramente negativa, ya que aún habiendo valorado el esfuerzo de posedición como bajo (4,6 de media), siguen respondiendo que la tecnología no es provechosa.

Tras completar la tarea de posedición y extraer la información necesaria, los datos recopilados para cada oración incluyen la oración original en español, la traducción automática al inglés, su versión poseditada en inglés, una valoración del esfuerzo de posedición, el tiempo de posedición y el valor de HTER. Es conveniente apuntar que debido a que el estudio se realizó en una empresa privada, no ha sido posible hacer público el conjunto de datos.

4.3 Correlación de los parámetros de esfuerzo

En esta sección se inspecciona la relación entre los resultados de los tres parámetros propuestos como representativos de las dimensiones del esfuerzo de posedición (percepción comunicada de esfuerzo, tiempo de posedición y valor de HTER) para cada segmento. Para ello, se ha calculado la correlación de Spearman. Es importante apuntar que, pese a las diferencias que emergieron respecto al trabajo de las traductoras, el análisis de correlación de parámetros de esfuerzo se realizó con la combinación de todos los datos recopilados. Es sabido que el proceso de traducción/posedición varía dependiendo del profesional que lo realiza aun tratándose del mismo conjunto de segmentos, y por consiguiente, un análisis siempre conlleva esta variabilidad en mayor o menor grado. Además, la unificación de los datos nos permite examinar un conjunto más representativo. No obstante, se realizó el mismo análisis para los datos de cada traductora por separado, lo cual confirmó que, sin bien los valores absolutos no coincidían, las tendencias eran las mismas, llevándonos a las mismas conclusiones. Los resultados se muestran en el Cuadro 4.

Parámetro de esfuerzo	(1)	(2)	(3)
(1) Percepción comunicada	1,000		
(2) Tiempo	-0,386*	1,000	
(3) HTER	-0,712*	0,443*	1,000

Cuadro 4: Correlación de Spearman de los parámetros de esfuerzo de posedición (*p<0.001)

Como podemos observar, la percepción comunicada de esfuerzo y HTER obtienen la correlación más alta, logrando una fuerte correlación. Esto podría indicar que, en cierta medida, el número de ediciones necesarias está relacionado con una percepción de mayor o menor de esfuerzo, es decir, a menor número de ediciones, más sencillo parece el trabajo. Sin embargo, la débil y moderada correlación entre el tiempo transcurrido durante la posedición, y la percepción de esfuerzo y HTER, respectivamente, no parece respaldar esta hipótesis.

En la Figura 2 podemos observar más claramente la correlación negativa entre la percepción de esfuerzo y HTER, es decir, cuanto más óptima parece una oración para posedición, menor es el valor HTER de dicha oración, menos ediciones necesita la propuesta de TA. La mayoría de las oraciones etiquetadas como 5, que indica que la oración es impecable y requiere un esfuerzo mínimo de posedición, tiene un HTER de 0, es decir, que no requiere ningún cambio.

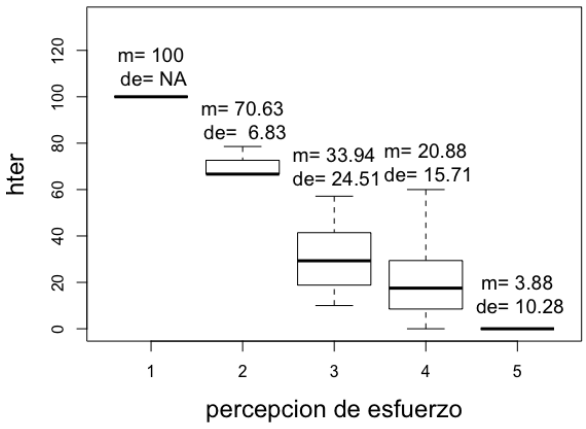

Figura 2: Correlación entre HTER y percepción comunicada de esfuerzo, donde *m* indica la media y *de* la desviación estándar

Se aprecia un gran salto entre los niveles 3 y 2. Para el nivel 2, el HTER es el segundo más alto. Esto significa que estas oraciones se han modificado en gran medida.

En la Figura 3 podemos observar la correlación negativa entre tiempo y percepción. Intuitivamente podríamos pensar que cuanto menor parece el esfuerzo requerido por una propuesta de TA, menor es el tiempo invertido en su posedición. Es interesante ver que esta correlación es débil. Por una parte, observamos que el tiempo medio necesario para poseditar las oraciones clasificadas en los niveles 2–5 no varía excesivamente. Si bien es cierto que la caja para el nivel 5 es comparativamente estrecha, lo cual sugiere que estas oraciones se poseditaron en un intervalo de tiempo similar, y que, por el contrario, la caja para el nivel 3 es comparativamente ancha, lo que indica que el tiempo de posedición varía más entre las traductoras en este caso, las medias se centran en torno a 3–8 segundos. Sin embargo, el tiempo medio necesario para poseditar las oraciones clasificadas como 1 se eleva a 36 segundos.

Por último, en la Figura 4, podemos ver cómo la correlación entre HTER y tiempo es positiva. Aunque la mayoría de los casos cuentan con un valor de HTER y un tiempo bajo, hay una tendencia que indica que cuanto más alto es el HTER, más alto es el tiempo de posedición. Esto parece indicar que cuantas más ediciones (inserciones, eliminaciones, sustituciones y reordenaciones, representadas por HTER) se hagan, más tiempo se necesita. Aun así, observamos que esta correlación es moderada, lo cual podría señalar la existencia de un tercer factor que distorsiona la relación directa entre HTER y el tiempo.

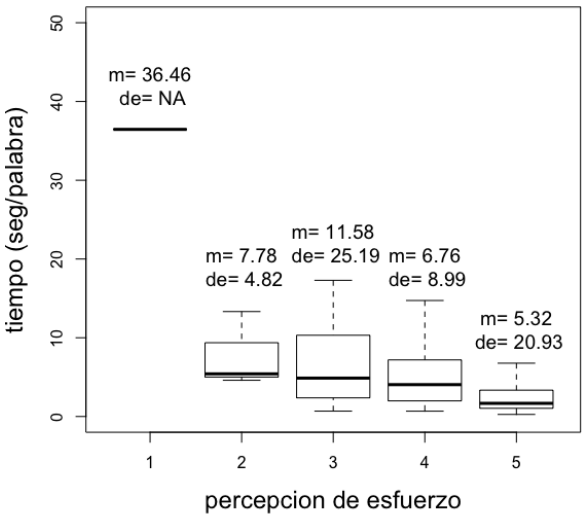

Figura 3: Correlación entre tiempo de posedición y percepción comunicada de esfuerzo, donde *m* indica la media y *de* la desviación estándar

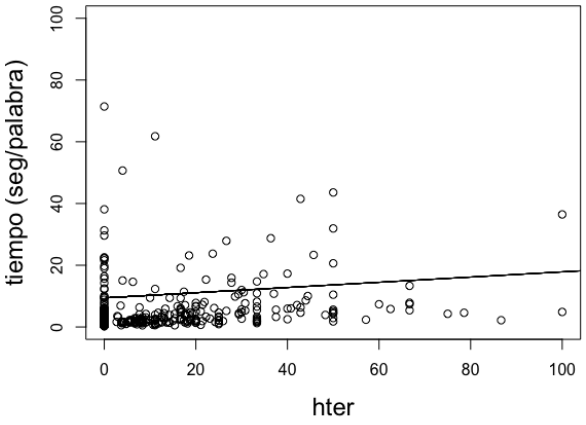

Figura 4: Correlación entre HTER y tiempo de posedición

4.4 Umbrales para los parámetros de esfuerzo

Como hemos mencionado anteriormente, el objetivo que perseguimos es estudiar la relación entre los resultados de los parámetros de las tres dimensiones para identificar qué parámetros de esfuerzo se deberían incluir en el entrenamiento de un modelo de predicción que sea capaz de recomendar si una oración nueva debería traducirse manualmente o si se debería poseditar. Sin embargo, los datos recopilados no responden a una división binaria; los parámetros de esfuerzo utilizados asignan a cada oración un rango de valores continuos (en el caso del tiempo y HTER) o dis-

cretos (en el caso de la percepción). Por lo tanto, una vez recopilados los datos, debemos establecer un umbral para clasificar las oraciones bien como óptimas para poseditar (PE) o para traducir (T) para cada uno de los parámetros de esfuerzo.

La literatura recoge varias aproximaciones a este paso. Una de las estrategias consiste en recopilar tareas de traducción y posedición para el mismo conjunto de oraciones y realizar una comparación directa de los resultados obtenidos para cada tarea (Aranberri & Pascual, 2018). Otros enfoques se basan en establecer valores de umbral específicos para cada parámetro de esfuerzo, como puede ser el de HTER (Parra Escartín & Arcedillo, 2015). Al carecer de datos de comparación directa (recordemos que las traductoras realizaron un trabajo de posedición pero no de traducción), proponemos umbrales para cada uno de los parámetros atendiendo al entorno experimental. Pese a que sería posible seleccionar umbrales distintos, que probablemente resultarían en una clasificación final distinta, nuestro objetivo principal no es obtener la selección final de las oraciones a traducir o poseditar sino mostrar la diferencia existente entre cada una de las dimensiones.

Percepción comunicada de esfuerzo > 3

El Cuadro 5 muestra la escala de percepción comunicada de esfuerzo de 1 a 5 según la cual las traductoras tuvieron que calificar cada oración de TA. Dada la definición asignada a cada valor, las valoraciones de 4 y 5 suponen que la calidad de las propuestas de TA es adecuada para posedición, mientras que en la valoración de 3 es dudosa, y para los valores 2 y 1 claramente no es adecuada. Teniendo en cuenta la opinión negativa de las traductoras hacia la traducción automática y el trabajo de posedición, decidimos excluir de la tarea de posedición todas aquellas oraciones que no facilitasen notablemente la posedición. Partiendo de esta decisión, establecemos el umbral para la percepción de esfuerzo en 3. De esta manera, las oraciones con una valoración de 4 y 5 se clasificarán para posedición y las oraciones con valoraciones de 1, 2 y 3 para traducción.

Tiempo < 11,5 seg/palabra

Durante la tarea de posedición, la plataforma Matecat registró el tiempo invertido por las traductoras en cada oración. Asumimos que cuanto menor es el tiempo empleado para poseditar una propuesta de TA, más sencilla resulta la tarea. El tiempo medio de posedición para las oraciones de nuestro conjunto de datos es de 6,1 segundos por palabra. El recuento del tiempo nos ofrece la posibilidad de ordenar las oraciones según el tiempo de trabajo requerido, normalizado por el número de palabras, pero aún así necesitamos establecer un umbral para poder obtener la clasificación binaria que buscamos. Para ello nos basamos en el tiempo medio asignado por la empresa para las tareas de traducción, que asciende a 313 palabras por hora y, por lo tanto, a 11,5 segundos por palabra. Así, aquellas oraciones que estén por debajo de ese tiempo se clasificarán como óptimas para poseditar y viceversa, las oraciones que superen los 11,5 segundos por palabra se clasificarán para traducir.

HTER < 33

HTER representa el número de ediciones necesarias para obtener una versión de traducción adecuada. Por lo tanto, cuanto más bajo sea el valor de HTER, mejor será la calidad de la propuesta de TA. Los pocos estudios que se han centrado en establecer el valor óptimo de HTER a partir del cual las oraciones se deberían poseditar o traducir, sugieren que éste se encuentra entre los 30–35 puntos (Parra Escartín & Arcedillo, 2015), por lo que nos centraremos en esa franja. Si observamos la Figura 5, vemos que para el nivel de percepción comunicada de esfuerzo 3, el HTER promedio de nuestros datos es del 33 %. Viendo que este valor se encuentra dentro del rango sugerido por los estudios previos, establecemos el umbral en el 33 % y asumimos que las oraciones con un HTER inferior al 33 % son de una calidad lo suficientemente buena como para ser seleccionadas para posedición.

4.5 Clasificación de las oraciones para los parámetros de esfuerzo

En resumen, atendiendo a los umbrales definidos, clasificamos las oraciones en óptimas para posedición (PE) o traducción (T), creando tres conjuntos, uno para cada dimensión tratada. Para dividir el conjunto según la percepción de esfuerzo, unimos las oraciones de los grupos 1, 2 y 3 asignándoles la categoría de T y las oraciones de los grupos 4 y 5 en PE. Para el conjunto de tiempo, aquellas oraciones que cuenten con un total de tiempo que suponga una media superior a 11 segundos por palabra formarán el subconjunto T, mientras que el formarán el subconjunto de PE. Finalmente, la división del conjunto según HTER se realizará asignando aquellas oraciones con un valor superior a 33 al subconjunto T y las oraciones con un valor inferior a 33 al subconjunto PE.

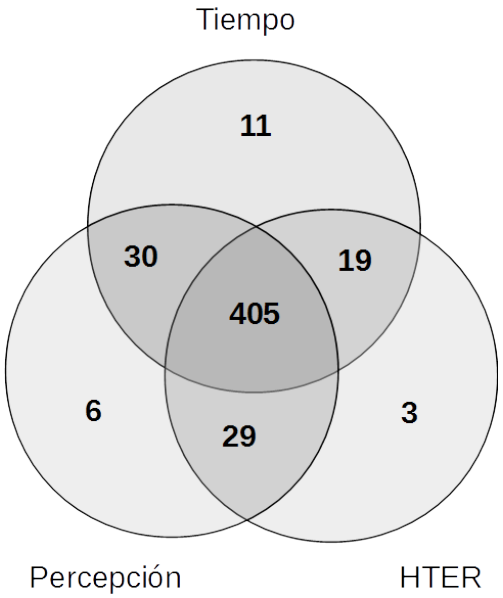

Figura 5: Número de elementos clasificados para poseditar comunes a las distintas dimensiones de esfuerzo de posedición.

El Cuadro 5 muestra el número de oraciones asignadas a cada clase para cada uno de los parámetros. A pesar de que el tamaño de ambas clases es muy similar independientemente del parámetro de esfuerzo utilizado, no se puede concluir que el comportamiento de los tres parámetros de esfuerzo es semejante. Esto se debe a que estos porcentajes no consideran que las oraciones incluidas en cada clase sean las mismas. Con el objetivo de estudiar la clasificación exacta de cada oración, calculamos cuáles de las asignadas para posedición son comunes a los tres parámetros de esfuerzo, cuáles a dos de ellos y cuáles únicamente a uno (ver Figura 5).

Clase	Percepción	Tiempo	HTER
PE	470 (92.34 %)	465 (91.36 %)	456 (89.59 %)
T	39 (7.66 %)	44 (8.64 %)	53 (10.41 %)

Cuadro 5: Número de oraciones asignadas a las clases PE y T para cada parámetro

Los resultados muestran que un total de 405 oraciones pertenecen a la intersección de los tres parámetros de esfuerzo, es decir, que los tres parámetros las clasifican para posedición. De la misma manera, en la Figura 5 también se puede observar el número de oraciones que son comunes únicamente a dos de los parámetros (por ejemplo, 30 en el caso del tiempo y la percepción). Finalmente, también se observa el número de oraciones que son comunes a un único parámetro de esfuerzo (por ejemplo, 3 en el caso de HTER).

Estos resultados muestran que si bien los tres parámetros de esfuerzo han asignado un porcentaje elevado de oraciones a la misma clase, la clasificación de alrededor de un 15 % difiere. Estos resultados parecen indicar que los parámetros de esfuerzo evaluados no son capaces de representar el esfuerzo global de posedición por separado, y por lo tanto, sugieren que se deberían utilizar de manera conjunta para una medición más exacta del esfuerzo de posedición. Es posible que al limitarnos al uso de los parámetros por separado, como es común en el desarrollo de modelos automáticos de estimación, no estemos proporcionando a dichos modelos datos completos para un entrenamiento óptimo.

A la vista de la discordancia en la clasificación de oraciones dependiendo del parámetro de esfuerzo, en este trabajo adoptamos una visión conservadora y proponemos clasificar como oraciones óptimas para posedición únicamente aquellas que son comunes a las tres dimensiones, que en este caso supondría el 79,5 % de las oraciones totales. Esta decisión se basa en la visión general negativa sobre la TA por parte de las traductoras. Con los datos recopilados y la clasificación realizada, se proseguiría al entrenamiento de un modelo automático de estimación que clasificase las nuevas oraciones en óptimas para posedición o traducción. Éste propondría para poseditar aquellas oraciones que fueran óptimas desde el punto de vista de las tres dimensiones. De esta manera, pretendemos que las traductoras tengan una experiencia lo más positiva posible durante las tareas de posedición y comiencen a vincular la TA con unos hábitos de trabajo satisfactorios.

5 Conclusiones

Este trabajo cuestiona el uso por separado de información referente a las dimensiones de esfuerzo propuestas por Krings (2001) (temporal, cognitiva y técnica) a la hora de medir el esfuerzo del trabajo de posedición y aboga por la inclusión de información de las tres dimensiones de manera conjunta. Se propone una estrategia multidimensional para la selección de información incluida en los modelos automáticos de selección de candidatos de TA para poseditar. Proponemos combinar parámetros que representan las tres dimensiones para establecer un umbral más preciso para clasificar oraciones en óptimas para traducir o para poseditar. Específicamente, en este trabajo se estudia la posibilidad de utilizar el tiempo como medida de la dimensión temporal, la percepción comunicada de esfuerzo de posedición como medida del esfuerzo cognitivo, y HTER como medida del esfuerzo técnico.

En el trabajo se presenta la recopilación de datos de un entorno real, los cuales se analizan siguiendo la propuesta multidimensional en preparación a utilizarlos posteriormente para el entrenamiento de modelos automáticos de clasificación que, una vez optimizados, podrían incluirse en el flujo de traducción para seleccionar las propuestas de traducción automática que se deberían presentar a las traductoras y excluir aquellas que perjudicarían la producción.

El análisis de las clasificaciones de este pequeño conjunto de datos reales utilizado a modo de ejemplo parece indicar que distintos parámetros de esfuerzo de posedición (percepción, tiempo y HTER) que atienden a las tres dimensiones en grados diferentes no valoran de igual manera si es más eficiente traducir o poseditar una oración. Los resultados obtenidos parecen sugerir que las dimensiones por separado no son capaces de describir el esfuerzo real de posedición y que, por lo tanto, el uso de un único parámetro de esfuerzo que se centre particularmente en una de las dimensiones no es adecuado para calcular el esfuerzo de posedición real.

Debido a la naturaleza preliminar de este trabajo, recordamos que los resultados deben tomarse con cautela. El conjunto de datos utilizado a modo de ejemplo cuenta con limitaciones tanto de extensión como de participantes, de modo que la unificación de los datos podría no ser completamente representativa del trabajo de posedición para el tipo de texto y sistema de traducción automática estudiados. Así, apuntamos como trabajo futuro replicar este análisis con un mayor número de traductoras, que realicen tareas paralelas, e incluyan conjuntos más extensos, para poder obtener datos más sólidos respecto a las correlaciones entre los parámetros de medición de las distintas dimensiones. Además, convendría extender el estudio de distintas combinaciones de umbrales.

Si bien parece conveniente combinar las tres dimensiones de esfuerzo, sería necesario continuar investigando qué parámetros representan las distintas dimensiones de manera más precisa y completa. Este trabajo ha explorado el uso de HTER como parámetro de esfuerzo técnico. Sin embargo, como ya se ha mencionado anteriormente, sería interesante utilizar el número total de ediciones realizadas por las traductoras durante el proceso de edición. Durante el proceso de traducción y posedición, es práctica común reformular una y otra vez distintas partes de una oración hasta conseguir una versión final, lo cual probablemente vaya ligado a la dificultad que entraña el segmento en cuestión. El uso de HTER podría ocultar el esfuerzo técnico real llevado a cabo en segmentos complejos.

Asimismo, cabría explorar otras mediciones más objetivas para el esfuerzo cognitivo, es decir, que no se basen en la opinión de las traductoras, quizá en la línea de Koponen et al. (2012) y Temnikova (2010), o incluso Moorkens et al. (2015). En este punto es necesario subrayar la importancia de la metodología seguida y las herramientas seleccionadas para recopilar los datos. Por ejemplo, en este trabajo, se ha querido primar la obtención del esfuerzo cognitivo de manera inmediata y para cada uno de los segmentos. Como el contexto del experimento exigía realizar la recopilación en línea, las aplicaciones disponibles no eran lo suficientemente flexibles como para realizar la medición temporal y de percepción completamente por separado. Si bien, como en este trabajo, se pueden tomar medidas adicionales para evitar que los datos se desvirtúen, consideramos que estrategias más rigurosas aportarán una mayor solidez de las conclusiones.

Finalmente, sería de gran interés analizar la relación existente entre las tres dimensiones de posedición para estudiar la aportación específica de cada dimensión al esfuerzo de posedición. Este estudio podría aportar mayor información a la hora de seleccionar o definir parámetros de esfuerzo para cada dimensión.

Agradecimientos

Las autoras quieren agradecer a los revisores, cuyos comentarios han contribuido a mejorar la versión original del trabajo. Este trabajo ha sido financiado parcialmente por el proyecto Modena (KK-2018/00087) del Departamento de Desarrollo Económico e Infraestructuras del Gobierno Vasco, el proyecto UnsupNMT (TIN2017-91692-EX) del Ministerio de Economía, Industria y Competitividad de España y el proyecto DOMINO (PGC2018-102041-B-I00, MCIU/AEI/FEDER, UE).

Referencias

Aranberri, Nora & Jose A Pascual. 2018. Towards a post-editing recommendation system for Spanish–Basque machine translation. En *21st Annual Conference of the European Association for Machine Translation (EAMT)*, 21–30.

Aziz, Wilker, Sheila Castilho Monteiro de Sousa & Lucia Specia. 2012. PET: a tool for post-editing and assessing machine translation. En

8th Language Resources and Evaluation Conference (LREC), 3982–3987.

Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint,* arXiv:1409.0473.

Bernth, Arendse & Claudia Gdaniec. 2001. MTranslatability. *Machine Translation* 16(3). 175–218. 10.1023/A:1019867030786.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia & Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT). En *2nd Conference on Machine Translation, Volume 2: Shared Task Papers*, 169–214. 10.18653/v1/W17-4717.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut & Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. En *7th Workshop on Statistical Machine Translation*, 10–51.

Federico, Marcello, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines et al. 2014. The MateCat tool. En *25th International Conference on Computational Linguistics: System Demonstrations (COLING)*, 129–132.

Felice, Mariano & Lucia Specia. 2012. Linguistic features for quality estimation. En *7th Workshop on Statistical Machine Translation*, 96–103.

Forcada, Mikel L. & Felipe Sánchez-Martínez. 2015. A general framework for minimizing translation effort: towards a principled combination of translation technologies in computer-aided translation. En *18th Annual Conference of the European Association for Machine Translation (EAMT)*, 27–34.

Hardmeier, Christian. 2011. Improving machine translation quality prediction with syntactic tree kernels. En *15th Annual Conference of the European Association for Machine Translation (EAMT)*, 233–240.

He, Yifan, Yanjun Ma, Josef van Genabith & Andy Way. 2010. Bridging SMT and TM with translation recommendation. En *48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 622–630.

Koponen, Maarit. 2012. Comparing human perceptions of post-editing effort with post-editing operations. En *7th Workshop on Statistical Machine Translation*, 181–190.

Koponen, Maarit, Wilker Aziz, Luciana Ramos & Lucia Specia. 2012. Post-editing time as a measure of cognitive effort. *Workshop on Post-Editing Technology and Practice* 11–20.

Krings, Hans P. 2001. *Repairing texts: empirical investigations of machine translation post-editing processes*, vol. 5. Kent State University Press.

Lacruz, Isabel, Michael Denkowski & Alon Lavie. 2014. Cognitive demand and cognitive effort in post-editing. En *3rd Workshop on Post-Editing Technology and Practice*, 73–84.

Moorkens, Joss, Sharon O'Brien, Igor da Silva, Norma de Lima Fonseca & Fabio Alves. 2015. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation* 29(3–4). 267–284. 10.1007/s10590-015-9175-2.

Parra Escartín, Carla & Manuel Arcedillo. 2015. Living on the edge: productivity gain thresholds in machine translation evaluation metrics. En *4th Workshop on Post-Editing Technology and Practice (WPTP)*, 46–56.

Parra Escartın, Carla & Manuel Arcedillo. 2015. Machine translation evaluation made fuzzier: A study on post-editing productivity and evaluation metrics in commercial settings, 131–144.

Parra Escartín, Carla, Hanna Béchara & Constantin Orăsan. 2017. Questing for quality estimation: A user study. *The Prague Bulletin of Mathematical Linguistics* 108(1). 343–354. 10.1515/pralin-2017-0032.

Plitt, Mirko & François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics* 93. 7–16. 10.2478/v10108-010-0010-x.

Shah, Kashif, Trevor Cohn & Lucia Specia. 2015. A bayesian non-linear method for feature selection in machine translation quality estimation. *Machine Translation* 29(2). 101–125. 10.1007/s10590-014-9164-x.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla & John Makhoul. 2006. A study of translation edit rate with targeted human annotation. En *Association*

for Machine Translation in the Americas, 223–231.

Snover, Matthew, Nitin Madnani, Bonnie J. Dorr & Richard Schwartz. 2009. Fluency, adequacy, or HTER?: exploring different human judgments with a tunable MT metric. En *4th Workshop on Statistical Machine Translation*, 259–268.

Specia, Lucia & Atefeh Farzindar. 2010. Estimating machine translation post-editing effort with HTER. En *Workshop Bringing MT to the User: MT Research and the Translation Industry (AMTA)*, 33–41.

Specia, Lucia, Dhwaj Raj & Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation* 24(1). 39–50. doi 10.1007/s10590-010-9077-2.

Temnikova, Irina P. 2010. Cognitive evaluation approach for a controlled language post–editing experiment. En *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 3485–3490.

ANEXO I: Instrucciones de posedición

You will be presented machine translated (MT) sentences in English, together with their source sentence in Spanish. You need to modify (postedit) the machine translated sentence as little as possible so that it bares the same meaning as its source. This postedition may include substitutions, deletions, insertions or reorderings. For each postedition, three things will be recorded:

- Your postedited version of the sentence
- A number provided by you indicating the quality of the machine translated sentence
- The time you take to postedit (this will be recorded automatically)

Here you can find an example of the task you'll need to perform:

	Ejemplo
Source	Es necesario tapar esos agujeros ya que las patas traseras de la armadura apoyan sobre ellos.
MT	These holes have to be covered as the frame rear feet support them.
Postedited version	These holes need to be covered as the frame back legs support on them.
Quality	1– 2 – 3 – **4** –5

The tool: Matecat

The platform we will be using is called Matecat. Matecat is an open source online CAT tool. If you want more information, check this link: https://www.matecat.com/about/.

To perform this task, you will be given two links, one for each task (practice task and real task). In one, you will have 5 sentences to familiarize yourself with the environment. In the other, you will have between 130 and 230 sentences to postedit, around 2180 words. When opening the link, you will be directed to the environment in which you will work. This is the following:

The platform shows the source sentence on the left side and its machine translation on the right side. This machine translated proposal is taken from a translation memory that contains all segments translated automatically. You need to post-edit the Machine Translate sentence in the right window taking into account the source sentence shown on the left window.

Quality rating

To rate the quality of the Machine Translation, please follow the scale proposed by Lacruz, Denkowski & Lavie (2014):

Valor	Descripción	
1	Sin sentido	La traducción era totalmente incomprensible.
2	No utilizable	La traducción contenía tantos errores que claramente hubiera sido más rápido traducir.
3	Neutral	La traducción contenía bastantes errores, no está claro qué hubiera sido más rápido.
4	Utilizable	La traducción contenía algunos errores, pero sería más útil poseditar.
5	Muy buena	La traducción era correcta o casi correcta.

After post-editing each sentence, you need to add a numberat the end of each sentence indicating the quality of the machine translated sentence following the aforementioned scale (1-5). Then, hit the button TRANSLATED to finish.

Please keep in mind that you need to write this number for each sentence.

If you feel that a certain machine translated sentence is perfectand needs no post-editing, just add a numberfrom 1 to 5 at the end and hit the TRANSLATED button directly.