COMET: Un marco neuronal para la evaluación de MT

Ricardo Rei Craig Stewart Ana C Farinha Alon Lavie Unbabel IA

{ricardo.rei, craig.stewart, catarina.farinha, alon.lavie}@unbabel.com

Resumen

Presentamos COMET, un marco neural para e ntrenar modelos de evaluación de traducción a utomática multilingüe que obtiene nuevos nive les de correlación de vanguardia con los juicio s humanos. Nuestro marco aprovecha los recie ntes avances en el modelado de lenguaje preen trenado cruzado, lo que resulta en modelos de evaluación de TA altamente multilingües y ad aptables que explotan información tanto de la entrada de origen como de una traducción de r eferencia en el idioma de destino para predecir con mayor precisión la calidad de la TA. Para mostrar nuestro marco, entrenamos tres model os con diferentes tipos de juicios humanos: *Direct Assessments*, *Human-mediated Trans-*, *lation Edit Rate* y *Multidimensional Qual-*, *ity Metrics*. Nuestros modelos logran un rendi miento de vanguardia en la tarea compartida d e métricas WMT 2019 y demuestran robustez frente a sistemas de alto rendimiento.

1 Introducción

Históricamente, las métricas para evaluar la calida d de la traducción automática (TA) se han basado en evaluar la similitud entre una hipótesis generad a por TA y una traducción de referencia generada por humanos en el idioma de destino. Las métrica s tradicionales se han centrado en características b ásicas a nivel léxico, como contar el número de n-gramas coincidentes entre la hipótesis de TA y la traducción de referencia. Métricas como BLEU (P apineni et al., 2002) y METEOR (Lavie y Denko wski, 2009) siguen siendo populares como medio para evaluar sistemas de TA debido a su ligereza y rápida computación.

Los enfoques neuronales modernos para la trad ucción automática resultan en una calidad de tradu cción mucho más alta que a menudo se desvía de l a transferencia léxica monótona entre idiomas. Po r esta razón, se ha vuelto cada vez más evidente q ue ya no podemos confiar en métricas como BLE U para proporcionar una estimación precisa de la calidad de la traducción automática (Barrault et al. , 2019).

Mientras que un aumento en el interés de invest igación en métodos neuronales para entrenar mod elos y sistemas de MT ha resultado en una mejora dramática reciente en la calidad de MT, la evaluac ión de MT ha quedado rezagada. La comunidad de investigación en MT todavía depende en gran med ida de métricas obsoletas y no ha surgido ningún n uevo estándar ampliamente adoptado. En 2019, la Tarea Compartida de Traducción de Noticias de WMT recibió un total de 153 envíos de sistemas d e MT (Barrault et al., 2019). La Tarea Compartida de Métricas del mismo año solo vio 24 envíos, cas i la mitad de los cuales eran participantes de la Tar ea Compartida de Estimación de Calidad, adaptad a como métricas (Ma et al., 2019).

Los hallazgos de la tarea mencionada anteriorm ente destacan dos desafíos importantes para la eva luación de MT que buscamos abordar aquí (Ma et al., 2019). A saber, que las métricas actuales tiene n dificultades para correlacionarse con precisión c on el juicio humano a nivel de segmento y no logr an diferenciar adecuadamente los sistemas de MT de mejor rendimiento.

En este artículo, presentamos COMET[1], un mar co basado en PyTorch para entrenar modelos de e valuación de MT altamente multilingües y adapta bles que pueden funcionar como métricas. Nuestro marco aprovecha los recientes avances en modela do de lenguaje cruzado (Artetxe y Schwenk, 2019; Devlin et al., 2019; Conneau y Lample, 2019; Con neau et al., 2019) para generar estimaciones de pre dicción de juicios humanos como *Direct Assessments* (DA) (Graham et al., 2013), *Human-mediated Transla- tion Edit Rate* (HTER) (Snover et al., 2006) y métricas compatibles con el marco *Multidimensional Qual- ity Metric* (Lom mel et al., 2014).

Inspirados en trabajos recientes sobre Estimació n de Calidad (QE) que demostraron que es posible alcanzar altos niveles de correlación con los juicio s humanos incluso sin una traducción de referenci a (Fonseca et al., 2019), proponemos un enfoque n ovedoso para incorporar-

[1] **C**rosslingual **O**ptimized **M**etric for **E**valuation of **T**ranslation.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2685–2702,
November 16–20, 2020.

ingresar la entrada en el idioma fuente en nuestros modelos de evaluación de MT. Tradicionalmente, solo los modelos de QE han utilizado la entrada fu ente, mientras que las métricas de evaluación de MT se basan en la traducción de referencia. Como en (Takahashi et al., 2020), mostramos que el uso de un espacio de incrustación multilingüe nos per mite aprovechar la información de las tres entrada s y demostrar el valor añadido por la fuente como entrada a nuestros modelos de evaluación de MT.

Para ilustrar la efectividad y flexibilidad del ma rco COMET, entrenamos tres modelos que estima n diferentes tipos de juicios humanos y mostramos un progreso prometedor hacia una mejor correlaci ón a nivel de segmento y robustez ante traduccion es automáticas de alta calidad.

Publicaremos tanto el marco COMET como los modelos de evaluación de MT entrenados descrito s en este documento a la comunidad de investigaci ón tras la publicación.

2 Arquitecturas de Modelos

Los juicios humanos sobre la calidad de la MT su elen presentarse en forma de puntuaciones a nivel de segmento, como DA, MQM y HTER. Para DA, es una práctica común convertir las puntuaciones en clasificaciones relativas (DARR) cuando el nú mero de anotaciones por segmento es limitado (Bo jar et al., 2017b; Ma et al., 2018, 2019). Esto signi fica que, para dos hipótesis de MT h_i y h_j de la mi sma fuente s, si la puntuación DA asignada a h_i es mayor que la puntuación asignada a h_j, h_i se consi dera una hipótesis "mejor". Para abarcar estas dife rencias, nuestro marco admite dos arquitecturas di stintas: el modelo Estimador y el modelo de Clasif icación de Traducción. La diferencia fundamental entre ellos es el objetivo de entrenamiento. Mientr as que el Estimador se entrena para hacer una regr esión directa sobre una puntuación de calidad, el modelo de Clasificación de Traducción se entrena para minimizar la distancia entre una hipótesis "m ejor" y tanto su referencia correspondiente como s u fuente original. Ambos modelos están compuest os por un codificador multilingüe y una capa de ag rupamiento.

2.1 Codificador multilingüe

El bloque de construcción principal de todos los m odelos en nuestro marco es un modelo multilingüe preentrenado, como BERT multilingüe (Devlin et al., 2019), XLM (Conneau y Lample, 2019) o XL M-RoBERTa (Conneau et al., 2019). Estos model os contienen varias capas de codificadores transfo rmer que son entrenado para reconstruir tokens enmascarados al descubrir la relación entre esos tokens y los circun dantes. Cuando se entrena con datos de múltiples i diomas, se ha encontrado que este objetivo preentr enado es altamente efectivo en tareas multilingües como la clasificación de documentos y la inferenci a en lenguaje natural (Conneau et al., 2019), gener alizando bien a idiomas y escrituras no vistas (Pire s et al., 2019). Para los experimentos en este docu mento, nos basamos en XLM-RoBERTa (base) co mo nuestro modelo de codificador.

Dada una secuencia de entrada $x = [x_0, x_1, ..., x_n]$, el codificador produce un embeddi ng $\boldsymbol{e}_j^{(\ell)}$ para cada token x_j y cada capa $\ell \in \{0, 1, ..., k\}$. En nuestro marco, aplicamos este proceso a la fuente, la hipótesis de MT y la referencia para mapearlas en un espacio de características compar tido.

2.2 Capa de Agrupamiento

Las incrustaciones generadas por la última capa de los codificadores preentrenados se utilizan general mente para ajustar modelos a nuevas tareas. Sin e mbargo, (Tenney et al., 2019) mostraron que difer entes capas dentro de la red pueden capturar infor mación lingüística que es relevante para diferentes tareas posteriores. En el caso de la evaluación de MT, (Zhang et al., 2020) mostraron que diferentes capas pueden lograr diferentes niveles de correlaci ón y que utilizar solo la última capa a menudo res ulta en un rendimiento inferior. En este trabajo, uti lizamos el enfoque descrito en Peters et al. (2018) y agrupamos información de las capas de codifica dores más importantes en una sola incrustación pa ra cada token, $\boldsymbol{e}_j$, utilizando un mecanismo de ate nción por capas. Esta incrustación se calcula como :

$$\boldsymbol{e}_{x_j} = \mu E_{x_j}^{\top} \boldsymbol{\alpha} \quad (1)$$

donde μ es un coeficiente de peso entrenable, $\boldsymbol{E}_j = [\boldsymbol{e}_j^{(0)}, \boldsymbol{e}_j^{(1)}, \dots \boldsymbol{e}_j^{(k)}]$ corresponde al vector d e incrustaciones de capa para el token x_j, y $\boldsymbol{\alpha} = \text{softmax}([\alpha^{(1)}, \alpha^{(2)}, \dots, \alpha^{(k)}])$ es un vector corres pondiente a los pesos entrenables por capa. Para e vitar el sobreajuste a la información contenida en cualquier capa individual, utilizamos la deserción de capa (Kondratyuk y Straka, 2019), en la que co n una probabilidad p el peso $\alpha^{(i)}$ se establece en $-\infty$.

Finalmente, como en (Reimers y Gurevych, 20 19), aplicamos agrupamiento promedio a las incru staciones de palabras resultantes para derivar una i ncrustación de oración para cada segmento.

[2] In the WMT Metrics Shared Task, if the difference between the DA scores is not higher than 25 points, those segments are excluded from the DARR data.

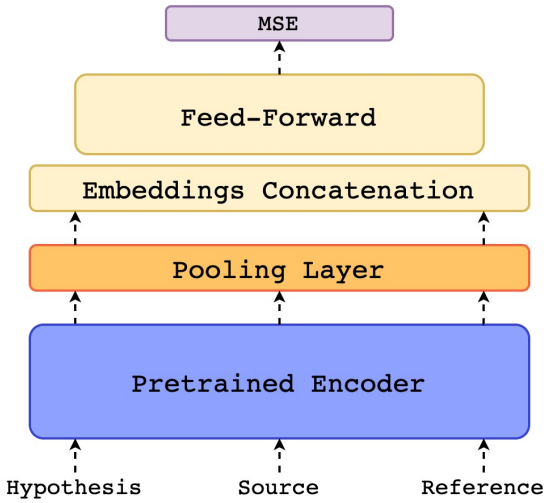

Figura 1: Arquitectura del modelo estimador. La fuente, la hipótesis y la referencia se codifican de manera independiente utilizando un codificador multilingüe preentrenado. Las incrustaciones de palabras resultantes se pasan a través de una capa de agrupamiento para crear una incrustación de oración para cada segmento. Finalmente, las incrustaciones de oración resultantes se combinan y concatenan en un solo vector que se pasa a un regresor de avance. Todo el modelo se entrena minimizando el Error Cuadrático Medio (MSE).

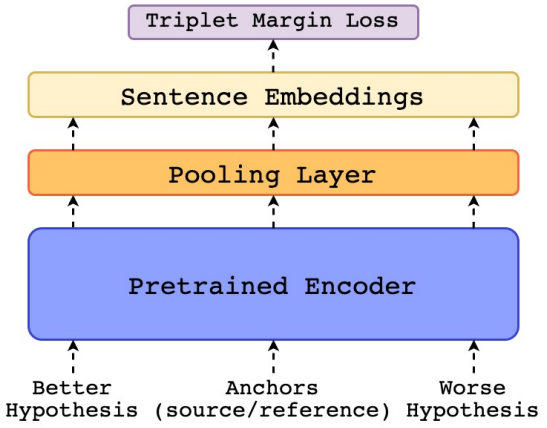

Figura 2: Arquitectura del modelo de clasificación de traducción. Esta arquitectura recibe 4 segmentos: el origen, la referencia, una hipótesis "mejor" y una "peor". Estos segmentos se codifican de manera independiente utilizando un codificador multilingüe preentrenado y una capa de agrupamiento en la parte superior. Finalmente, utilizando la pérdida de margen de tripletas (Schroff et al., 2015) optimizamos el espacio de incrustación resultante para minimizar la distancia entre la hipótesis "mejor" y los "anclajes" (origen y referencia).

2.3 Modelo de Estimador

Dada una incrustación de oración de d dimensiones para la fuente, la hipótesis y la referencia, adoptamos el enfoque propuesto en RUSE (Shimanaka et al., 2018) y extraemos las siguientes características combinadas:

- Producto fuente elemento a elemento: $\boldsymbol{h} \odot \boldsymbol{s}$
- Producto de referencia elemento a elemento: $\boldsymbol{h} \odot \boldsymbol{r}$
- Diferencia absoluta elemento por elemento de la fuente: $|\boldsymbol{h} - \boldsymbol{s}|$
- Diferencia de referencia absoluta elemento por elemento: $|\boldsymbol{h} - \boldsymbol{r}|$

Estas características combinadas se concatenan luego con la incrustación de referencia $\boldsymbol{r}$ y la incrustación de hipótesis $\boldsymbol{h}$ en un solo vector $\boldsymbol{x} = [\boldsymbol{h}; \boldsymbol{r}; \boldsymbol{h} \odot \boldsymbol{s}; \boldsymbol{h} \odot \boldsymbol{r}; |\boldsymbol{h} - \boldsymbol{s}|; |\boldsymbol{h} - \boldsymbol{r}|]$ que sirve como entrada para un regresor de avance. La fuerza de estas características radica en resaltar las diferencias entre las incrustaciones en el espacio de características semánticas.

El modelo se entrena para minimizar el error cuadrático medio entre las puntuaciones predichas y las evaluaciones de calidad (DA, HTER o MQM). La Figura 1 ilustra la arquitectura propuesta.

Tenga en cuenta que elegimos no incluir la incrustación de origen sin procesar ($\boldsymbol{s}$) en nuestra entrada concatenada. Experimentos iniciales revelaron que el valor añadido por la incrustación de origen como características de entrada adicionales para nuestro regresor era, en el mejor de los casos, insignificante. Una variación de nuestro modelo estimador HTER entrenado con el vector $\boldsymbol{x} = [\boldsymbol{h}; \boldsymbol{s}; \boldsymbol{r}; \boldsymbol{h} \odot \boldsymbol{s}; \boldsymbol{h} \odot \boldsymbol{r}; |\boldsymbol{h} - \boldsymbol{s}|; |\boldsymbol{h} - \boldsymbol{r}|]$ como entrada para el modelo de solo avance logró mejorar el rendimiento a nivel de segmento en 8 de los 18 pares de idiomas descritos en la sección 5 a continuación y la mejora promedio en Tau de Kendall en esos entornos fue +0.0009. Como se señala en Zhao et al. (2020), si bien los modelos preentrenados multilingües son adaptativos a múltiples idiomas, el espacio de características entre idiomas está mal alineado. Con base en esto, decidimos a favor de excluir la incrustación de origen con la intuición de que la información más importante proviene de la incrustación de referencia y reducir el espacio de características permitiría al modelo centrarse más en la información relevante. Sin embargo, esto no niega el valor general de la fuente para nuestro modelo; donde incluimos características de combinación como $\boldsymbol{h} \odot \boldsymbol{s}$ y $|\boldsymbol{h} - \boldsymbol{s}|$ notamos ganancias en la correlación, como se explora más a fondo en la sección 5.5 a continuación.

2.4 Modelo de Clasificación de Traducción

Nuestro modelo de clasificación de traducción (Fi gura 2) recibe como entrada una tupla $\chi = (s, h^+, h^-, r)$ donde h^+ denota una hipótesis que fue clasificada más alto que otra hipótesis h^-. Luego pasamos χ a través de nuestro codificador multilingüe y capa de agrupamiento para obtener una incrustación de oración para cada segmento e n el χ. Finalmente, utilizando las incrustaciones $\{\boldsymbol{s}, \boldsymbol{h}^+, \boldsymbol{h}^-, \boldsymbol{r}\}$, calculamos la pérdida de margen de tripletas (Schroff et al., 2015) en relación con l a fuente y la referencia:

$$L(\chi) = L(\boldsymbol{s}, \boldsymbol{h}^+, \boldsymbol{h}^-) + L(\boldsymbol{r}, \boldsymbol{h}^+, \boldsymbol{h}^-) \quad (2)$$

donde:

$$L(\boldsymbol{s}, \boldsymbol{h}^+, \boldsymbol{h}^-) = \max\{0, d(\boldsymbol{s}, \boldsymbol{h}^+) - d(\boldsymbol{s}, \boldsymbol{h}^-) + \epsilon\} \quad (3)$$

$$L(\boldsymbol{r}, \boldsymbol{h}^+, \boldsymbol{h}^-) = \max\{0, d(\boldsymbol{r}, \boldsymbol{h}^+) - d(\boldsymbol{r}, \boldsymbol{h}^-) + \epsilon\} \quad (4)$$

$d(\boldsymbol{u}, \boldsymbol{v})$ denota la distancia euclidiana entre $\boldsymbol{u}$ y $\boldsymbol{v}$ y ϵ es un margen. Así, durante el entrenamiento, el modelo optimiza el espacio de incrustación para q ue la distancia entre los anclajes ($\boldsymbol{s}$ y $\boldsymbol{r}$) y la hipóte sis "peor" $\boldsymbol{h}^-$ sea mayor en al menos ϵ que la dista ncia entre los anclajes y la hipótesis "mejor" $\boldsymbol{h}^+$.

Durante la inferencia, el modelo descrito recibe un triplete $(s, \hat{h}, r)$ con solo una hipótesis. La punt uación de calidad asignada a $\hat{h}$ es la media armóni ca entre la distancia a la fuente $d(\boldsymbol{s}, \hat{\boldsymbol{h}})$ y la distan cia a la referencia $d(\boldsymbol{r}, \hat{\boldsymbol{h}})$:

$$f(s, \hat{h}, r) = \frac{2 \times d(\boldsymbol{r}, \hat{\boldsymbol{h}}) \times d(\boldsymbol{s}, \hat{\boldsymbol{h}})}{d(\boldsymbol{r}, \hat{\boldsymbol{h}}) + d(\boldsymbol{s}, \hat{\boldsymbol{h}})} \quad (5)$$

Finalmente, convertimos la distancia resultante en una puntuación de similitud limitada entre 0 y 1 de la siguiente manera:

$$\hat{f}(s, \hat{h}, r) = \frac{1}{1 + f(s, \hat{h}, r)} \quad (6)$$

3 Corpora

Para demostrar la efectividad de nuestras arquitect uras de modelo descritas (sección 2), entrenamos t res modelos de evaluación de MT donde cada mo delo se dirige a un tipo diferente de juicio humano . Para entrenar estos modelos, utilizamos datos de tres corpora diferentes: el corpus QT21, el DARR de la tarea compartida de métricas WMT (2017 a 2019) y un corpus anotado MQM propietario.

3.1 El corpus QT21

El corpus QT21 es un conjunto de datos disponibl e públicamente [3] que contiene oraciones generada s por la industria de los dominios de tecnología de la información o ciencias de la vida (Specia et al., 2017). Este corpus contiene un total de 173K tupla s con la oración fuente, la referencia generada por humanos respectiva, la hipótesis de MT (ya sea de un MT estadístico basado en frases o de un MT ne uronal), y el MT post-editado (PE). Los pares de i diomas representados en este corpus son: inglés a alemán (en-de), letón (en-lt) y checo (en-cs), y ale mán a inglés (de-en).

La puntuación HTER se obtiene al calcular la ta sa de edición de traducción (TER) (Snover et al., 2 006) entre la hipótesis de MT y el PE correspondi ente. Finalmente, después de calcular el HTER par a cada MT, construimos un conjunto de datos de e ntrenamiento $D = \{s_i, h_i, r_i, y_i\}_{n=1}^N$, donde s_i de nota el texto fuente, h_i denota la hipótesis de MT, r_i la traducción de referencia y y_i la puntuación H TER para la hipótesis h_i. De esta manera, buscam os aprender una regresión $f(s, h, r) \rightarrow y$ que pred iga el esfuerzo humano requerido para corregir la hipótesis al observar el fuente, la hipótesis y la ref erencia (pero no la hipótesis post-editada).

3.2 El corpus DARR de WMT

Desde 2017, los organizadores de la Tarea Compa rtida de Traducción de Noticias WMT (Barrault et al., 2019) han recopilado juicios humanos en form a de DAs de adecuación (Graham et al., 2013, 201 4, 2017). Estos DAs se mapean en clasificaciones relativas (DARR) (Ma et al., 2019). Los datos resu ltantes para cada año (2017-19) forman un conjunt o de datos $D = \{s_i, h_i^+, h_i^-, r_i\}_{n=1}^N$ donde h_i^+ den ota una hipótesis "mejor" y h_i^- denota una "peor". Aquí buscamos aprender una función $r(s, h, r)$ tal que la puntuación asignada a h_i^+ sea estrictamente mayor que la puntuación asignada a h_i^- ($r(s_i, h_i^+, r_i) > r(s_i, h_i^-, r_i)$). Estos datos [4] contienen un total de 24 pares de idiomas de alto y bajo recurso, como chino a inglés (zh-en) e inglés a gujarati (en-gu).

3.3 El corpus MQM

El corpus MQM es una base de datos interna prop ietaria de traducciones generadas por MT de sopo rte al cliente.

[3] QT21 data: https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2390

[4] The raw data for each year of the WMT Metrics shared task is publicly available in the results page (2019 example: http://www.statmt.org/wmt19/results.html). Note, however, that in the README files it is highlighted that this data is not well documented and the scripts occasionally require custom utilities that are not available.

mensajes de chat que fueron anotados de acuerdo con las directrices establecidas en Burchardt y Lo mmel (2014). Estos datos contienen un total de 12 K tuplas, que cubren 12 pares de idiomas del inglé s a: alemán (en-de), español (en-es), español latino americano (en-es-latam), francés (en-fr), italiano (en-it), japonés (en-ja), neerlandés (en-nl), portugu és (en-pt), portugués brasileño (en-pt-br), ruso (en -ru), sueco (en-sv) y turco (en-tr). Tenga en cuenta que en este corpus el inglés siempre se considera c omo el idioma fuente, pero nunca como el idioma objetivo. Cada tupla consiste en una oración fuent e, una referencia generada por humanos, una hipót esis de MT y su puntuación MQM, derivada de an otaciones de errores por uno (o más) anotadores e ntrenados. La métrica MQM a la que se hace refer encia a lo largo de este documento es una métrica interna definida de acuerdo con el marco MQM (L ommel et al., 2014) (MQM). Los errores se anotan bajo una tipología interna definida en tres tipos pri ncipales de errores: 'Estilo', 'Fluidez' y 'Precisión'. Nuestras puntuaciones MQM varían de $-\infty$ a 100 y se definen como:

$$\text{MQM} = 100 - \frac{I_{\text{Minor}} + 5 \times I_{\text{Major}} + 10 \times I_{\text{Crit.}}}{\text{Sentence Length} \times 100} \tag{7}$$

donde I_{Minor} denota el número de errores menores, I_{Major} el número de errores mayores y $I_{\text{Crit.}}$ el núme ro de errores críticos.

Nuestra métrica MQM tiene en cuenta la grave dad de los errores identificados en la hipótesis de MT, lo que lleva a una métrica más detallada que HTER o DA. Cuando se usaron en nuestros experi mentos, estos valores se dividieron por 100 y se tr uncaron en 0. Como en la sección 3.1, construimo s un conjunto de datos de entrenamiento $D = \{s_i, h_i, r_i, y_i\}_{n=1}^{N}$, donde s_i denota el texto fuente, h_i denota la hipótesis de MT, r_i la traducci ón de referencia y y_i la puntuación MQM para la hipótesis h_i.

4 Experimentos

Entrenamos dos versiones del modelo Estimador d escrito en la sección 2.3: una que regresa en HTE R (COMET-HTER) entrenada con el corpus QT2 1, y otra que regresa en nuestra implementación pr opietaria de MQM (COMET-MQM) entrenada co n nuestro corpus interno de MQM. Para el modelo de Clasificación de Traducción, descrito en la secc ión 2.4, entrenamos con el corpus WMT DARR d e 2017 y 2018 (COMET-RANK). En esta sección, introducimos el entrenamiento. configuración para estos modelos y la configuraci ón de evaluación correspondiente.

4.1 Configuración del Entrenamiento

Las dos versiones de los Estimadores (COMET-H TER/MQM) comparten la misma configuración d e entrenamiento y hiperparámetros (los detalles se incluyen en los Apéndices). Para el entrenamiento , cargamos el codificador preentrenado e inicializa mos tanto la capa de agrupamiento como el regres or de avance. Mientras que los escalares por capa $\boldsymbol{\alpha}$ de la capa de agrupamiento se establecen inicial mente en cero, los pesos del avance se inicializan aleatoriamente. Durante el entrenamiento, dividim os los parámetros del modelo en dos grupos: los p arámetros del codificador, que incluyen el modelo del codificador y los escalares de $\boldsymbol{\alpha}$; y los parámet ros del regresor, que incluyen los parámetros de la red de avance superior. Aplicamos descongelamie nto gradual y tasas de aprendizaje discriminativas (Howard y Ruder, 2018), lo que significa que el m odelo del codificador está congelado durante una é poca mientras que el avance se optimiza con una t asa de aprendizaje de 3e$-$5. Después de la primer a época, todo el modelo se ajusta finamente, pero l a tasa de aprendizaje para los parámetros del codif icador se establece en 1e$-$5 para evitar el olvido c atastrófico.

En contraste con los dos Estimadores, para el m odelo COMET-RANK afinamos desde el principi o. Además, dado que este modelo no añade ningú n nuevo parámetro además de los escalares de cap a $\boldsymbol{\alpha}$ sobre XLM-RoBERTa (base), utilizamos una única tasa de aprendizaje de 1e$-$5 para todo el mo delo.

4.2 Configuración de Evaluación

Utilizamos los datos de prueba y la configuración de la Tarea Compartida de Métricas WMT 2019 (Ma et al., 2019) para comparar los modelos COM ET con las mejores presentaciones de la tarea com partida y otras métricas recientes de vanguardia co mo BERTSCORE y BLEURT.[5] El método de eval uación utilizado es la formulación oficial similar a Tau de Kendall, τ, de la Tarea Compartida de Mét ricas WMT 2019 (Ma et al., 2019) definida como:

$$\tau = \frac{\textit{Concordant} - \textit{Discordant}}{\textit{Concordant} + \textit{Discordant}} \tag{8}$$

donde *Concordant* es el número de veces que una métrica asigna una puntuación más alta a la hipót esis "mejor" h^+ y *Discordant* es el número de vec es que una métrica asigna una puntuación más alt a a la hipótesis "peor"

[5]To ease future research we will also provide, within our framework, detailed instructions and scripts to run other metrics such as CHRF, BLEU, BERTSCORE, and BLEURT

Tabla 1: Correlaciones de Tau de Kendall (τ) en pares de idiomas con inglés como fuente para el corpus de métric as DARR WMT19. Para BERTSCORE, informamos los resultados con el modelo de codificador predeterminado para una comparación completa, pero también con XLM-RoBERTa (base) para la equidad con nuestros modelos. Los valores reportados para YiSi-1 se toman directamente del artículo de la tarea compartida (Ma et al., 2019).

Metric	**en-cs**	**en-de**	**en-fi**	**en-gu**	**en-kk**	**en-lt**	**en-ru**	**en-zh**
BLEU	0.364	0.248	0.395	0.463	0.363	0.333	0.469	0.235
CHRF	0.444	0.321	0.518	0.548	0.510	0.438	0.548	0.241
YISI-1	0.475	0.351	0.537	0.551	0.546	0.470	0.585	0.355
BERTSCORE (default)	0.500	0.363	0.527	0.568	0.540	0.464	0.585	0.356
BERTSCORE (xlmr-base)	0.503	0.369	0.553	0.584	0.536	0.514	0.599	0.317
COMET-HTER	0.524	0.383	0.560	0.552	0.508	0.577	0.539	0.380
COMET-MQM	0.537	0.398	0.567	0.564	0.534	0.574	**0.615**	0.378
COMET-RANK	**0.603**	**0.427**	**0.664**	**0.611**	**0.693**	**0.665**	0.580	**0.449**

h^- o las puntuaciones asignadas a ambas hipótesi s son las mismas.

Como se mencionó en los hallazgos de (Ma et a l., 2019), las correlaciones a nivel de segmento de todas las métricas enviadas fueron frustrantemente bajas. Además, todas las métricas enviadas mostra ron una falta dramática de capacidad para clasifica r correctamente los sistemas de MT fuertes. Para e valuar si nuestros nuevos modelos de evaluación d e MT abordan mejor este problema, seguimos el c onjunto de evaluación descrito utilizado en el anál isis presentado en (Ma et al., 2019), donde se exa minan los niveles de correlación para porciones de los datos de DARR que incluyen solo los 10, 8, 6 y 4 sistemas de MT principales.

5 Resultados

5.1 Del inglés al X

La Tabla 1 muestra los resultados para los ocho pa res de idiomas con inglés como fuente. Contrasta mos nuestros tres modelos COMET contra métric as de referencia como BLEU y CHRF, la métrica ganadora de la tarea de 2019 YISI-1, así como la más reciente BERTSCORE. Observamos que en g eneral nuestros tres modelos entrenados con el ma rco COMET superan, a menudo por márgenes sig nificativos, todas las demás métricas. Nuestro mo delo DARR Ranker supera a los dos Estimadores en siete de los ocho pares de idiomas. Además, au nque el Estimador MQM está entrenado solo con 12K segmentos anotados, rinde aproximadamente al mismo nivel que el Estimador HTER para la ma yoría de los pares de idiomas, y supera a todas las demás métricas en en-ru.

5.2 De X al inglés

La Tabla 2 muestra los resultados para los siete pa res de idiomas hacia el inglés. Nuevamente, contra stamos nuestros tres modelos COMET contra métr icas de referencia como BLEU y CHRF, la métric a ganadora de la tarea de 2019 YISI-1, como así como las métricas BERTSCORE y BLEURT p ublicadas recientemente. Como se muestra en la T abla 1, el modelo DARR muestra fuertes correlaci ones con los juicios humanos, superando la métric a BLEURT específica para inglés propuesta recien temente en cinco de siete pares de idiomas. Nueva mente, el Estimador MQM muestra resultados sor prendentemente fuertes a pesar de que este model o fue entrenado con datos que no incluían inglés c omo objetivo. Aunque el codificador utilizado en nuestros modelos entrenados es altamente multilin güe, hipotetizamos que este poderoso resultado "z ero-shot" se debe a la inclusión de la fuente en nue stros modelos.

5.3 Pares de idiomas que no involucran inglés

Los tres modelos COMET fueron entrenados con datos que involucran inglés (ya sea como fuente o como objetivo). Sin embargo, para demostrar que nuestras métricas se generalizan bien, las probamo s en los tres pares de idiomas de WMT 2019 que n o incluyen inglés ni en la fuente ni en el objetivo. Como se puede ver en la Tabla 3, nuestros resulta dos son consistentes con las observaciones en las Tablas 1 y 2.

5.4 Robustez ante MT de Alta Calidad

Para el análisis, utilizamos el corpus DARR de la t area compartida de 2019 y evaluamos en el subco njunto de los datos de los sistemas de MT de mejo r rendimiento para cada par de idiomas. Incluimos pares de idiomas para los cuales pudimos recupera r datos de al menos diez sistemas de MT diferente s (es decir, todos menos kk-en y gu-en). Contrasta mos con el fuerte BERTSCORE y BLEURT prop uestos recientemente, con BLEU como línea base. Los resultados se presentan en la Figura 3. Para lo s pares de idiomas donde el inglés es el objetivo, n uestros tres modelos son mejores o competitivos c on todos los demás; donde el inglés es la fuente, n otamos que en general nuestras métricas superan e l rendimiento de otros.

Tabla 2: Correlaciones de Tau de Kendall (τ) en pares de idiomas con inglés como objetivo para el corpus de métricas DARR de WMT19. En cuanto a BERTSCORE, para BLEURT informamos resultados para dos modelos: el modelo base, que es comparable en tamaño con el codificador que utilizamos, y el modelo grande que es el doble de tamaño.

Metric	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
BLEU	0.053	0.236	0.194	0.276	0.249	0.177	0.321
CHRF	0.123	0.292	0.240	0.323	0.304	0.115	0.371
YISI-1	0.164	0.347	0.312	**0.440**	0.376	0.217	0.426
BERTSCORE (default)	0.190	0.354	0.292	0.351	0.381	**0.221**	0.432
BERTSCORE (xlmr-base)	0.171	0.335	0.295	0.354	0.356	0.202	0.412
BLEURT (base-128)	0.171	0.372	0.302	0.383	0.387	0.218	0.417
BLEURT (large-512)	0.174	0.374	0.313	0.372	0.388	0.220	0.436
COMET-HTER	0.185	0.333	0.274	0.297	0.364	0.163	0.391
COMET-MQM	**0.207**	0.343	0.282	0.339	0.368	0.187	0.422
COMET-RANK	0.202	**0.399**	**0.341**	0.358	**0.407**	0.180	**0.445**

Tabla 3: Correlaciones de Tau de Kendall (τ) en pares de idiomas que no involucran inglés para el corpus de métricas DARR WMT19.

Metric	de-cs	de-fr	fr-de
BLEU	0.222	0.226	0.173
CHRF	0.341	0.287	0.274
YISI-1	0.376	0.349	0.310
BERTSCORE (default)	0.358	0.329	0.300
BERTSCORE (xlmr-base)	0.386	0.336	0.309
COMET-HTER	0.358	0.397	0.315
COMET-MQM	0.386	0.367	0.296
COMET-RANK	**0.389**	**0.444**	**0.331**

ers. Incluso el estimador MQM, entrenado con solo 12K segmentos, es competitivo, lo que destaca el poder de nuestro marco propuesto.

5.5 La Importancia de la Fuente

Para arrojar algo de luz sobre el valor real y la contribución de la entrada del idioma fuente en la capacidad de nuestros modelos para aprender predicciones precisas, entrenamos dos versiones de nuestro modelo DARR Ranker: una que utiliza solo la referencia y otra que utiliza tanto la referencia como la fuente. Ambos modelos fueron entrenados utilizando el corpus WMT 2017 que solo incluye pares de idiomas del inglés (en-de, en-cs, en-fi, en-tr). En otras palabras, aunque el inglés nunca se observó como un idioma objetivo durante el entrenamiento de ambas variantes del modelo, el entrenamiento de la segunda variante incluye incrustaciones de fuente en inglés. Luego probamos estas dos variantes del modelo en el corpus WMT 2018 para estos pares de idiomas y para las direcciones inversas (con la excepción de en-cs porque cs-en no existe para WMT 2018). Los resultados en la Tabla

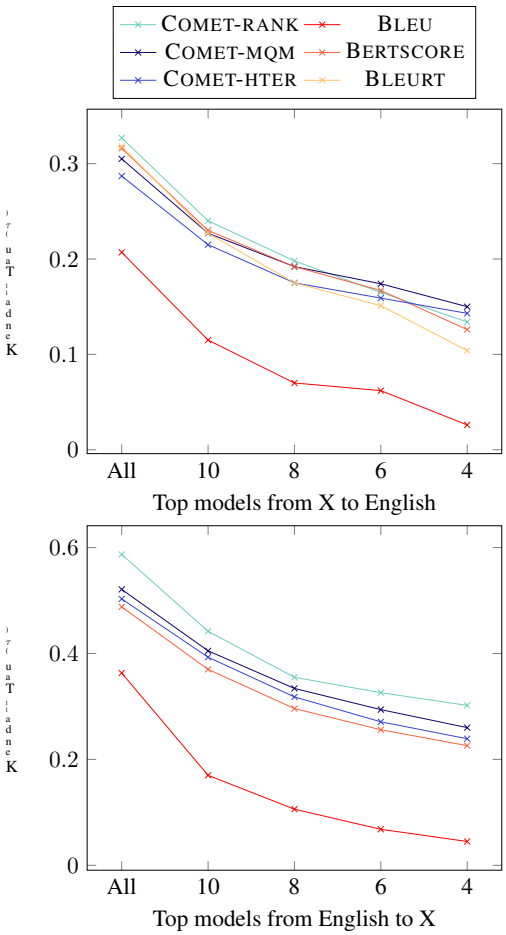

Figura 3: Rendimiento de métricas sobre todos y los mejores (10, 8, 6 y 4) sistemas de MT.

4 muestra claramente que para la arquitectura de clasificación de traducción, incluir la fuente mejora la correlación general con los juicios humanos. Además, la inclusión de la fuente expuso la segunda variante del modelo a los embeddings en inglés, que es

Tabla 4: Comparación entre COMET-RANK (sección 2.4) y una versión solo de referencia de este en los datos de WMT18. Ambos modelos fueron entrenados con WMT17, lo que significa que el modelo solo de referencia nunc a se expone al inglés durante el entrenamiento.

Metric	**en-cs**	**en-de**	**en-fi**	**en-tr**	**cs-en**	**de-en**	**fi-en**	**tr-en**
COMET-RANK (ref. only)	0.660	0.764	0.630	0.539	0.249	0.390	0.159	0.128
COMET-RANK	0.711	0.799	0.671	0.563	0.356	0.542	0.278	0.260
$\Delta\tau$	0.051	0.035	0.041	0.024	**0.107**	**0.155**	**0.119**	**0.132**

reflejado en un $\Delta\tau$ más alto para los pares de idio mas con inglés como objetivo.

6 Reproducibilidad

Publicaremos tanto la base de código del marco C OMET como los modelos de evaluación de MT en trenados descritos en este documento a la comunid ad de investigación tras la publicación, junto con l os scripts detallados necesarios para ejecutar todas las líneas base reportadas.[6] Todos los modelos rep ortados en este documento fueron entrenados en u na sola GPU Tesla T4 (16GB). Además, nuestro marco se basa en PyTorch Lightning (Falcon, 201 9), un envoltorio ligero de PyTorch, que fue cread o para una flexibilidad y reproducibilidad máxima s.

7 Trabajo Relacionado

Las métricas clásicas de evaluación de MT se cara cterizan comúnmente como métricas de coinciden cia de n-gramas porque, utilizando características elaboradas a mano, estiman la calidad de MT cont ando el número y la fracción de n-gramas que apa recen simultáneamente en una hipótesis de traducc ión candidata y una o más referencias humanas. M étricas como BLEU (Papineni et al., 2002), MET EOR (Lavie y Denkowski, 2009) y CHRF (Popov ić, 2015) han sido ampliamente estudiadas y mejo radas (Koehn et al., 2007; Popović, 2017; Denko wski y Lavie, 2011; Guo y Hu, 2019), pero, por di seño, generalmente no logran reconocer y capturar la similitud semántica más allá del nivel léxico.

En los últimos años, las incrustaciones de palab ras (Mikolov et al., 2013; Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2019) han surgid o como una alternativa comúnmente utilizada al e mparejamiento de n-gramas para capturar la simili tud semántica de las palabras. Métricas basadas en incrustaciones como METEOR-VECTOR (Servan et al., 2016), BLEU2VEC (Tättar y Fishel, 2017), YISI-1 (Lo, 2019), MOVERSCORE (Zhao et al., 2019) y BERTSCORE (Zhang et al., 2020) crean alineaciones suaves entre la referencia y la hipótes is.

[6]These will be hosted at: `https://github.com/Unbabel/COMET`

en un espacio de incrustación y luego calcular una puntuación que refleje la similitud semántica entre esos segmentos. Sin embargo, los juicios humanos como DA y MQM capturan mucho más que solo l a similitud semántica, lo que resulta en un límite s uperior de correlación entre los juicios humanos y las puntuaciones producidas por tales métricas.

Las métricas aprendibles (Shimanaka et al., 201 8; Mathur et al., 2019; Shimanaka et al., 2019) int entan optimizar directamente la correlación con lo s juicios humanos y han mostrado resultados prom etedores recientemente. BLEURT (Sellam et al., 2 020), una métrica aprendible basada en BERT (De vlin et al., 2019), afirma tener un rendimiento de v anguardia durante los últimos 3 años de la tarea co mpartida de métricas WMT. Debido a que BLEU RT se basa en English-BERT (Devlin et al., 2019) , solo se puede utilizar cuando el inglés es el idio ma objetivo, lo que limita su aplicabilidad. Adem á s, hasta donde sabemos, todas las métricas aprendi bles propuestas anteriormente se han centrado en optimizar DA, que, debido a la escasez de anotado res, puede resultar inherentemente ruidoso (Ma et al., 2019).

La evaluación de MT sin referencia, también co nocida como Estimación de Calidad (QE), histórica mente ha retrocedido a menudo en HTER para la evaluación a nivel de segmento (Bojar et al., 2013 , 2014, 2015, 2016, 2017a). Más recientemente, M QM se ha utilizado para la evaluación a nivel de d ocumento (Specia et al., 2018; Fonseca et al., 201 9). Al aprovechar codificadores multilingües pree ntrenados altamente efectivos como multilingual BERT (Devlin et al., 2019) y XLM (Conneau y L ampl, 2019), los sistemas de QE han mostrado cor relaciones auspiciosas con los juicios humanos (K epler et al., 2019a). Concurrentemente, el marco OpenKiwi (Kepler et al., 2019b) ha facilitado a lo s investigadores avanzar en el campo y construir modelos de QE más robustos.

8 Conclusiones y Trabajo Futuro

En este artículo presentamos COMET, un nuevo marco neuronal para entrenar modelos de evaluaci ón de MT que pueden servir como métricas autom áticas y ser fácilmente

adaptado y optimizado para diferentes tipos de jui cios humanos sobre la calidad de la MT.

Para demostrar la efectividad de nuestro marco, buscamos abordar los desafíos reportados en la Ta rea Compartida de Métricas WMT 2019 (Ma et al. , 2019). Entrenamos tres modelos distintos que log ran nuevos resultados de vanguardia para la correl ación a nivel de segmento con juicios humanos, y muestran una capacidad prometedora para diferen ciar mejor los sistemas de alto rendimiento.

Uno de los desafíos de aprovechar el poder de l os modelos preentrenados es el peso oneroso de lo s parámetros y el tiempo de inferencia. Una vía pri ncipal para el trabajo futuro en COMET examinar á el impacto de soluciones más compactas como DistilBERT (Sanh et al., 2019).

Además, aunque esbozamos la posible importa ncia del texto fuente arriba, señalamos que nuestro modelo COMET-RANK pondera la fuente y la ref erencia de manera diferente durante la inferencia, pero de manera equitativa en su función de pérdid a de entrenamiento. Trabajos futuros investigarán la optimalidad de esta formulación y examinarán más a fondo la interdependencia de las diferentes entradas.

Agradecimientos

Agradecemos a André Martins, Austin Matthews, Fabio Kepler, Daan Van Stigt, Miguel Vera y a lo s revisores, por sus valiosos comentarios y discusi ones. Este trabajo fue apoyado en parte por el Pro grama P2020 a través de los proyectos MAIA y U nba- bel4EU, supervisados por ANI bajo los núme ros de contrato 045909 y 042671, respectivamente .

Referencias

Mikel Artetxe y Holger Schwenk. 2019. Embeddings d e oraciones masivamente multilingües para transferenc ia cruzada sin entrenamiento y más allá. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Chr istian Federmann, Mark Fishel, Yvette Gra- ham, Barr y Haddow, Matthias Huck, Philipp Koehn, Shervin Ma lmasi, Christof Monz, Mathias Müller, Santanu Pal, M att Post, y Marcos Zampieri. 2019. Hallazgos de la con ferencia de 2019 sobre traducción automática (WMT1 9). En *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, páginas 1–61, Florencia, Italia. A sociación de Lingüística Computacional.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, C hristian Federmann, Barry Haddow, Philipp Koehn, C hristof Monz, Matt Post, Radu Soricut, y Lucia Specia. 2013. Resultados del Taller de Traduc ción Automática Estadística 2013. En *Proceedings of the Eighth Workshop on Statistical Machine Translation*, páginas 1–44, Sofía, Bulgaria. Asociaci ón de Lingüística Computacional.

Ondřej Bojar, Christian Buck, Christian Federmann, B arry Haddow, Philipp Koehn, Johannes Leveling, Chri stof Monz, Pavel Pecina, Matt Post, Herve Saint-Aman d, Radu Soricut, Lucia Specia y Aleš Tamchyna. 2014. Resultados del taller de 2014 sobre traducción automát ica estadística. En *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pá ginas 12–58, Baltimore, Maryland, EE. UU. Asociació n de Lingüística Computacional.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matth ias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva , Christof Monz, Matteo Negri, Matt Post, Raphael Ru bino, Lucia Specia y Marco Turchi. 2017a. Hallazgos de la conferencia de 2017 sobre traducción automática (WMT17). En *Proceedings of the Second Conference on Machine Translation*, páginas 169–214, Copenhague, Dinamarca. Asociación de Lingüísti ca Computacional.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Anton io Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Marian a Neves, Martin Popel, Matt Post, Raphael Rubino, Ca rolina Scarton, Lucia Spe- cia, Marco Turchi, Karin Ve rspoor, y Marcos Zampieri. 2016. Hallazgos de la conf erencia de 2016 sobre traducción automática. En *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, páginas 131–198, Berlín, Alem ania. Asociación de Lingüística Computacional.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Ne gri, Matt Post, Carolina Scarton, Lucia Specia y Marco Turchi. 2015. Resultados del taller de 2015 sobre tradu cción automática estadística. En *Proceedings of the Tenth Workshop on Statistical Machine Translation*, páginas 1–46, Lisboa, Portugal. Asociación de Lingüística Computacional.

Ondřej Bojar, Yvette Graham y Amir Kamran. 2017b. Resultados de la tarea compartida de métricas WMT17 . En *Proceedings of the Second Conference on Machine Translation*, páginas 489–513, Copenhague, Dinamarca. Asociación de Lingüística Computacional.

Aljoscha Burchardt y Arle Lommel. 2014. Directrices prácticas para el uso de MQM en la investigación cient ífica sobre la calidad de la traducción. (fecha de acceso : 2020-05-26).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettle-mo yer y Veselin Stoyanov. 2019. Aprendizaje de repre sentación cruzada no supervisado a gran escala. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau y Guillaume Lample. 2019. Preentren amiento de modelos de lenguaje multilingües. En H. W allach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox y R. Garnett, editores, *Advances in Neural Information Processing Systems 32*, páginas 7059–7069. Curran Associates, Inc.

Michael Denkowski y Alon Lavie. 2011. Meteor 1.3: Métrica automática para la optimización y evaluación confiables de sistemas de traducción automática. En *Proceedings of the Sixth Workshop on Statistical Machine Translation*, páginas 85–91, Edimburgo, Escocia. Asoc iación de Lingüística Computacional.

Jacob Devlin, Ming-Wei Chang, Kenton Lee y Kristin a Toutanova. 2019. BERT: Pre-entrenamiento de transf ormadores bidireccionales profundos para la comprens ión del lenguaje. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, p áginas 4171–4186, Minneapolis, Minnesota. Asociació n de Lingüística Computacional.

WA Falcon. 2019. PyTorch Lightning: El envoltorio li gero de PyTorch para la investigación de IA de alto ren dimiento. *GitHub*.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins , Mark Fishel y Christian Federmann. 2019. Resultados de las tareas compartidas de WMT 2019 sobre estimaci ón de calidad. En *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, páginas 1–10, Florencia, Italia. Asociación de Lingüística Computacional.

Yvette Graham, Timothy Baldwin, Alistair Moffat y Ju stin Zobel. 2013. Escalas de medición continuas en la e valuación humana de la traducción automática. En *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, páginas 33–41, S ofía, Bulgaria. Asociación de Lingüística Computacion al.

Yvette Graham, Timothy Baldwin, Alistair Moffat y Ju stin Zobel. 2014. ¿Está mejorando la traducción autom ática con el tiempo? En *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 443–451, Gotemb urgo, Suecia. Asociación de Lingüística Computaciona l.

Yvette Graham, Timothy Baldwin, Alistair Moffat y Ju stin Zobel. 2017. ¿Pueden los sistemas de traducción a utomática ser evaluados solo por la multitud? *Natural Lan- guage Engineering*, 23(1):330.

Yinuo Guo y Junfeng Hu. 2019. Meteor++ 2.0: Adopta r el conocimiento de paráfrasis a nivel sintáctico en la evaluación de traducción automática. En *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, páginas 501–506, Fl orencia, Italia. Asociación de Lingüística Computaci onal.

Jeremy Howard y Sebastian Ruder. 2018. Ajuste fino d e un modelo de lenguaje universal para la clasificación de texto. En *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 328–339, Melbourne, Australia. Asociación de Lingüística Computacional.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M. Amin Farajian, António V. Lo pes, y André F. T. Martins. 2019a. La participación de Unbabel en la tarea compartida de estimación de calida d de traducción WMT19. En *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, páginas 78–84, Flore ncia, Italia. Asociación de Lingüística Computacional.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera y André F. T. Martins. 2019b. OpenKiwi: Un ma rco de código abierto para la estimación de calidad. En *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, páginas 117–122, Florencia, Italia. As ociación de Lingüística Computacional.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris C allison-Burch, Marcello Federico, Nicola Bertoldi, Bro oke Cowan, Wade Shen, Christine Moran, Richard Zen s, Chris Dyer, Ondřej Bojar, Alexandra Constantin y E van Herbst. 2007. Moses: herramienta de código abiert o para la traducción automática estadística. En *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, páginas 177–180, Praga, República Checa. Asoc iación de Lingüística Computacional.

Dan Kondratyuk y Milan Straka. 2019. 75 idiomas, 1 modelo: Analizando dependencias universales de mane ra universal. En *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, páginas 2779–2795, Hong Kong, China. Aso ciación de Lingüística Computacional.

Alon Lavie y Michael Denkowski. 2009. La métrica m eteor para la evaluación automática de la traducción au tomática. *Machine Translation*, 23:105–115.

Chi-kiu Lo. 2019. YiSi - una métrica unificada de eval uación y estimación de calidad de MT semántica para l enguas con diferentes niveles de recursos disponibles. En *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, páginas 507–513, Florencia, Italia. Asociación de Lingüística Computacional.

Arle Lommel, Aljoscha Burchardt y Hans Uszkoreit. 2 014. Métricas de calidad multidimensional (MQM): A

marco para declarar y describir métricas de calidad de traducción. *Tradumtica: tecnologies de la traducci*, 0:455–463.

Qingsong Ma, Ondřej Bojar y Yvette Graham. 2018. Resultados de la tarea compartida de métricas WMT18: Tanto los caracteres como los embeddings logran un buen rendimiento. En *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, páginas 671–688, Bélgica, Bruselas. Asociación de Lingüística Computacional.

Qingsong Ma, Johnny Wei, Ondřej Bojar y Yvette Graham. 2019. Resultados de la tarea compartida de métricas WMT19: Los sistemas de MT a nivel de segmento y fuertes plantean grandes desafíos. En *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, páginas 62–90, Florencia, Italia. Asociación de Lingüística Computacional.

Nitika Mathur, Timothy Baldwin y Trevor Cohn. 2019. Poniendo la evaluación en contexto: Las incrustaciones contextuales mejoran la evaluación de la traducción automática. En *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, páginas 2799–2808, Florencia, Italia. Asociación de Lingüística Computacional.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado y Jeff Dean. 2013. Representaciones distribuidas de palabras y frases y su composicionalidad. En *Advances in Neural Information Processing Systems 26*, páginas 3111–3119. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward y Wei-Jing Zhu. 2002. Bleu: un método para la evaluación automática de la traducción automática. En *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, páginas 311–318, Filadelfia, Pensilvania, EE. UU. Asociación de Lingüística Computacional.

Jeffrey Pennington, Richard Socher y Christopher Manning. 2014. Glove: vectores globales para la representación de palabras. En *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 1532–1543, Doha, Catar. Asociación de Lingüística Computacional.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee y Luke Zettlemoyer. 2018. Representaciones de palabras contextualizadas profundas. En *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, páginas 2227–2237, Nueva Orleans, Luisiana. Asociación de Lingüística Computacional.

Telmo Pires, Eva Schlinger y Dan Garrette. 2019. ¿Qué tan multilingüe es BERT multilingüe? En *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, páginas 4996–5001, Florencia, Italia. Asociación de Lingüística Computacional.

Maja Popović. 2015. chrF: puntuación f de n-gramas de caracteres para la evaluación automática de MT. En *Proceedings of the Tenth Workshop on Statistical Machine Translation*, páginas 392–395, Lisboa, Portugal. Asociación de Lingüística Computacional.

Maja Popović. 2017. chrF++: palabras que ayudan a los n-grams de caracteres. En *Proceedings of the Second Conference on Machine Translation*, páginas 612–618, Copenhague, Dinamarca. Asociación de Lingüística Computacional.

Nils Reimers e Iryna Gurevych. 2019. Sentence-BERT: incrustaciones de oraciones utilizando redes Siamese BERT. En *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, páginas 3982–3992, Hong Kong, China. Asociación de Lingüística Computacional.

Victor Sanh, Lysandre Debut, Julien Chaumond y Thomas Wolf. 2019. Distilbert, una versión destilada de BERT: más pequeña, más rápida, más barata y más ligera. *arXiv preprint arXiv:1910.01108*.

F. Schroff, D. Kalenichenko y J. Philbin. 2015. Facenet: Un embedding unificado para el reconocimiento y agrupamiento de rostros. En *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 815–823.

Thibault Sellam, Dipanjan Das y Ankur Parikh. 2020. BLEURT: Aprendiendo métricas robustas para la generación de texto. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, páginas 7881–7892, En línea. Asociación de Lingüística Computacional.

Christophe Servan, Alexandre Bérard, Zied Elloumi, Hervé Blanchon y Laurent Besacier. 2016. Word2Vec vs DBnary: ¿Aumentando METEOR utilizando representaciones vectoriales o recursos léxicos? En *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, páginas 1159–1168, Osaka, Japón. El Comité Organizador de COLING 2016.

Hiroki Shimanaka, Tomoyuki Kajiwara y Mamoru Komachi. 2018. RUSE: Regresor utilizando incrustaciones de oraciones para la evaluación automática de traducción automática. En *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, páginas 751–758, Bélgica, Bruselas. Asociación de Lingüística Computacional.

Hiroki Shimanaka, Tomoyuki Kajiwara y Mamoru Komachi. 2019. Evaluación de Traducción Automática con Regresor BERT. *arXiv preprint arXiv:1907.12679*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Lin nea Micciulla y John Makhoul. 2006. Un estudio de la tasa de edición de traducción con anotación humana di rigida. En *In Proceedings of Association for Machine Translation in the Americas*, páginas 223–231.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ra món Astudillo y André F. T. Martins. 2018. Resultados de la tarea compartida de WMT 2018 sobre estimación de calidad. En *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, páginas 689–709, Bélgica, Bruselas. Asociación de Lingüística Computacional.

Lucia Specia, Kim Harris, Frédéric Blain, Aljoscha Bu rchardt, Viviven Macketanz, Inguna Skadina, Matteo Negri, y Marco Turchi. 2017. Calidad de traducción y productividad: Un estudio sobre lenguas con morfologí a rica. En *Machine Translation Summit XVI*, páginas 5 5–71, Nagoya, Japón.

Kosuke Takahashi, Katsuhito Sudoh y Satoshi Nakam ura. 2020. Evaluación automática de traducción autom ática utilizando entradas en el idioma fuente y un mode lo de lenguaje multilingüe. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, páginas 3553–3558, En línea. Asociación de Lingüística Computacional.

Andre Tättar y Mark Fishel. 2017. bleu2vec: la métrica dolorosamente familiar en esteroides de espacio vector ial continuo. En *Proceedings of the Second Conference on Machine Translation*, páginas 619–622, Copenhag ue, Dinamarca. Asociación de Lingüística Computacio nal.

Ian Tenney, Dipanjan Das y Ellie Pavlick. 2019. BER T redescubre la pipeline clásica de PLN. En *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, páginas 4593–4 601, Florencia, Italia. Asociación de Lingüística Comp utacional.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. W einberger y Yoav Artzi. 2020. Bertscore: Evaluando la generación de texto con bert. En *International Conference on Learning Representations*.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West y Steffen Eger. 2020. Sobre las limitacion es de los codificadores multilingües según lo expuesto por la evaluación de traducción automática sin referenc ia. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, páginas 1656–1 671, En línea. Asociación de Lingüística Computacion al.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christ ian M. Meyer y Steffen Eger. 2019. MoverScore: Eval uación de generación de texto con incrustaciones conte xtualizadas y distancia de movimiento de tierra. En *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, páginas 563–578, Hong Kong, China. Asociación de Lingüística Computaci onal.

A Apéndices

En la Tabla 5 enumeramos los hiperparámetros uti lizados para entrenar nuestros modelos. Antes de i nicializar estos modelos, se estableció una semilla aleatoria en 3 en todas las bibliotecas que realizan operaciones "aleatorias" (torch, numpy, random y cuda).

Tabla 5: Hiperparámetros utilizados en nuestro marco COMET para entrenar los modelos presentados.

Hyper-parameter	COMET(**Est-HTER/MQM)**	COMET-RANK
Encoder Model	XLM-RoBERTa (base)	XLM-RoBERTa (base)
Optimizer	Adam (default parameters)	Adam (default parameters)
n frozen epochs	1	0
Learning rate	3e-05 and 1e-05	1e-05
Batch size	16	16
Loss function	MSE	Triplet Margin ($\epsilon = 1.0$)
Layer-wise dropout	0.1	0.1
FP precision	32	32
Feed-Forward hidden units	2304,1152	–
Feed-Forward activations	Tanh	–
Feed-Forward dropout	0.1	–

Tabla 6: Estadísticas para el corpus QT21.

	en-de	**en-cs**	**en-lv**	**de-en**
Total tuples	54000	42000	35474	41998
Avg. tokens (reference)	17.80	15.56	16.42	17.71
Avg. tokens (source)	16.70	17.37	18.39	17.18
Avg. tokens (MT)	17.65	15.64	16.42	17.78

Tabla 7: Estadísticas para el corpus DARR WMT 2017.

	en-cs	**en-de**	**en-fi**	**en-lv**	**en-tr**
Total tuples	32810	6454	3270	3456	247
Avg. tokens (reference)	19.70	22.15	15.59	21.42	17.57
Avg. tokens (source)	22.37	23.41	21.73	26.08	22.51
Avg. tokens (MT)	19.45	22.58	16.06	22.18	17.25

Table 8: Statistics for the WMT 2019 DARR into-English language pairs.

Type	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en
Total	85365	32179	20110	9728	21862	39852	31070
Avg. tokens (reference)	20.29	18.55	17.64	20.36	26.55	21.74	42.89
Avg. tokens (source)	18.44	12.49	21.92	16.32	20.32	18.00	7.57
Avg. tokens (MT)	20.22	17.76	17.02	19.68	25.25	21.80	39.70

Table 9: Statistics for the WMT 2019 DARR from-English and non-English language pairs.

Type	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh	fr-de	de-cs	de-fr
Total	27178	99840	31820	11355	18172	17401	24334	18658	1369	23194	4862
Avg. tokens (reference)	22.92	25.56	20.12	33.32	18.89	21.00	24.79	9.25	22.86	22.27	27.32
Avg. tokens (source)	24.98	24.97	25.23	24.32	23.78	24.46	24.45	24.39	28.60	25.22	21.36
Avg. tokens (MT)	22.60	24.98	19.69	32.97	19.92	20.97	23.37	6.83	23.36	21.89	25.68

Table 10. MQM corpus (section 3.3) statistics.

Total pairs	en-nl	en-sv	en-ja	en-de	en-ru	en-es	en-fr	en-it	en-pt-br	en-tr	en-pt	en-es-latam
	2447	970	1590	2756	1043	259	1474	812	504	370	91	6
Avg. tokens (reference)	14.10	14.24	20.32	13.78	13.37	10.90	13.75	13.61	12.48	7.95	12.18	10.33
Avg. tokens (source)	14.23	15.31	13.69	13.76	13.94	11.23	12.85	14.22	12.46	10.36	13.45	12.33
Avg. tokens (MT)	13.66	13.91	17.84	13.41	13.19	10.88	13.59	13.02	12.19	7.99	12.21	10.17

Table 11. Statistics for the WMT 2018 DARR language pairs.

Total pairs	zh-en	en-zh	cs-en	fi-en	ru-en	tr-en	de-en	en-cs	en-de	en-et	en-fi	en-ru	en-tr	et-en
	33357	28602	5110	15648	10404	8525	77811	5413	19711	32202	9809	22181	1358	56721
Avg. tokens (reference)	28.86	24.04	21.98	21.13	24.97	23.25	23.29	19.50	23.54	18.21	16.32	21.81	20.15	23.40
Avg. tokens (source)	23.86	28.27	18.67	15.03	21.37	18.80	21.95	22.67	24.82	23.47	22.82	25.24	24.37	18.15
Avg. tokens (MT)	27.45	14.94	21.79	20.46	25.25	22.80	22.64	19.73	23.74	18.37	17.15	21.86	19.61	23.52

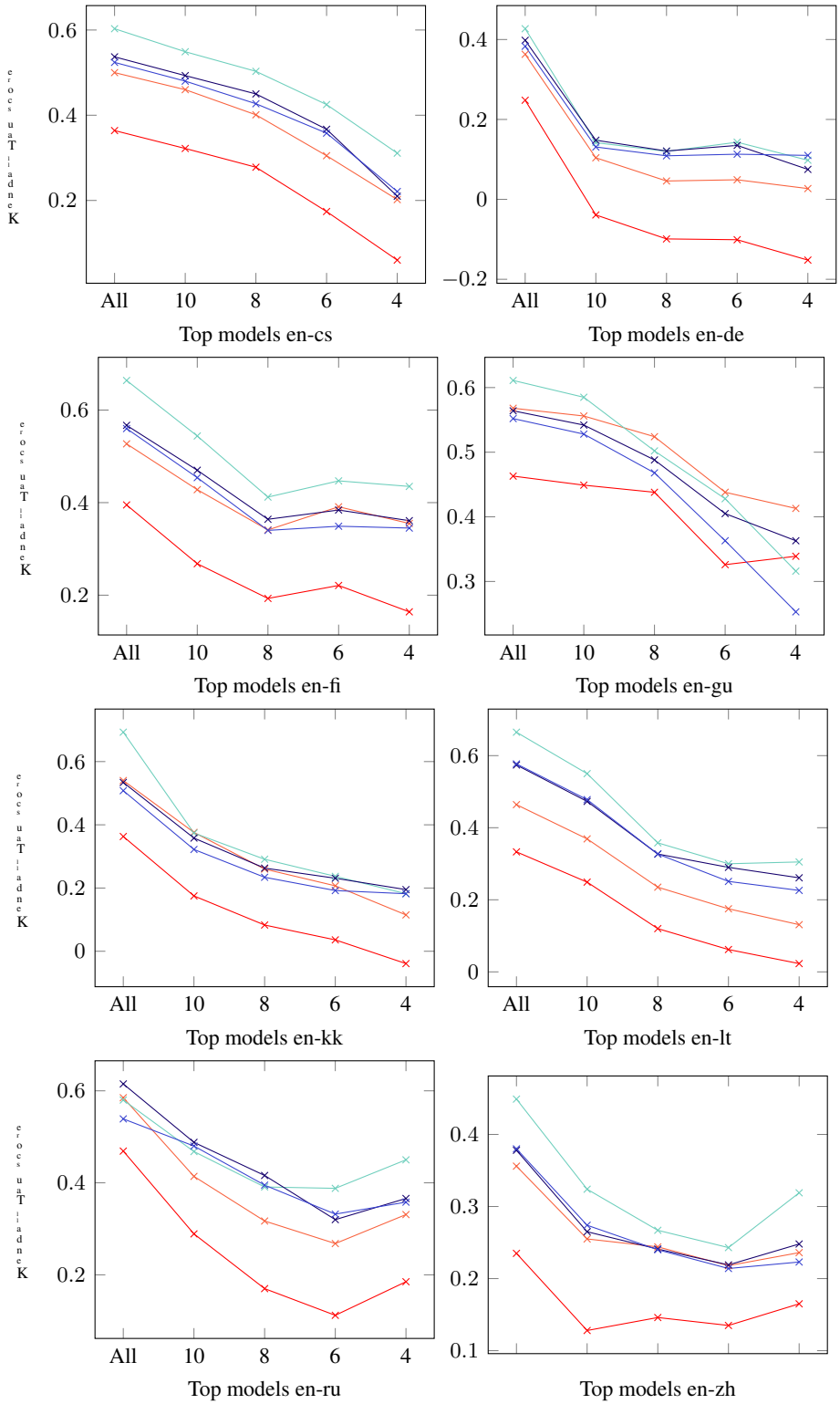

Tabla 12: Rendimiento de métricas sobre todos y los mejores (10, 8, 6 y 4) sistemas de MT para todos los pares de idiomas de inglés. El esquema de colores es el siguiente: COMET-RANK, COMET-HTER, COMET-MQM, BLEU, BERTSCORE

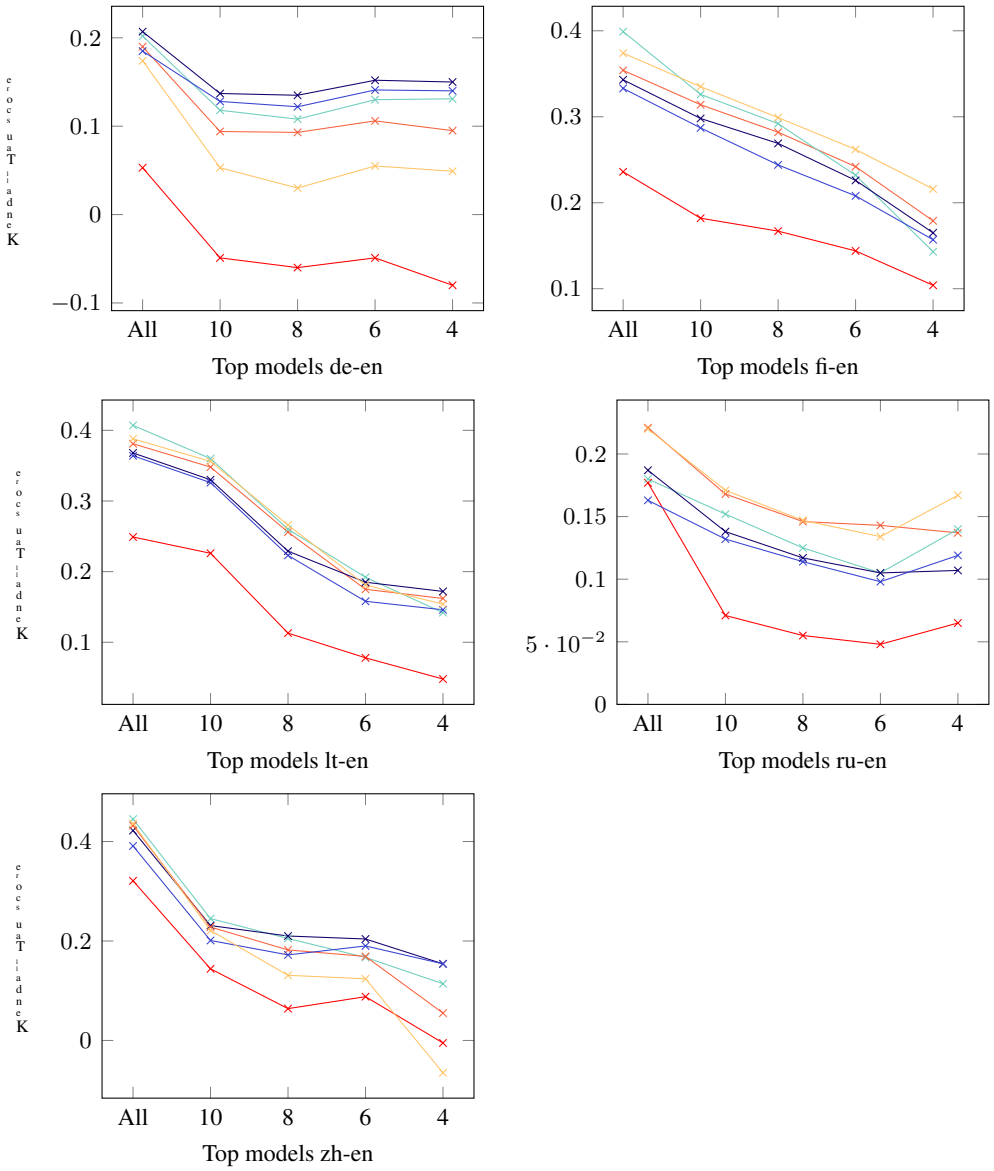

Tabla 13: Rendimiento de métricas sobre todos y los mejores (10, 8, 6 y 4) sistemas de MT para todos los pares de idiomas hacia el inglés. El esquema de colores es el siguiente: COMET-RANK, COMET-HTER, COMET-MQM, BLEU, BERTSCORE, BLEURT