COMET: Un Marco Neural para la Evaluación de MT

Ricardo Rei Craig Stewart Ana C Farinha Alon Lavie
Unbabel AI
{ricardo.rei, craig.stewart, catarina.farinha, alon.lavie}@unbabel.com

Resumen

Presentamos COMET, un marco neural para entrenar modelos de evaluación de traducción automática multilingüe que obtiene nuevos niveles de estado del arte. Nuestro marco aprovecha los recientes avances en el modelado de lenguaje preentrenado multilingüe resultando en modelos de evaluación de MT altamente multilingües y adaptables que explotan información tanto de la entrada de origen como de una traducción de referencia en el idioma objetivo con el fin de predecir con mayor precisión la calidad de MT. Para mostrar nuestro marco, entrenamos tres modelos con diferentes tipos de juicios humanos: Evaluaciones Directas, Tasa de Edición de Traducción Mediada por Humanos y Métricas de Calidad Multidimensional. Mediante el aprovechamiento de un nuevo rendimiento de estado del arte en la tarea compartida de la traducción automática (MT) en el WMT 2019 Metrics shared task and demonstrate robustness to high-performing systems.

1 Introduction

Historically, metrics for evaluating the quality of machine translation (MT) have relied on assessing la similitud entre una hipótesis generada por MT y una traducción de referencia generada por un humano en el idioma objetivo. Las métricas tradicionales se han centrado en características básicas a nivel léxico, como contar el número de n-gramas coincidentes entre la MT hipótesis y la traducción de referencia. Métricas como BLEU (Papineni et al., 2002) y METEOR (Lavie y Denkowski, 2009) siguen siendo populares como un medio para evaluar los sistemas de MT debido a su ligereza y rápida computación.

Los enfoques neuronales modernos para MT resultan en una calidad de traducción mucho más alta que a menudo se desvía de la transferencia léxica monótona entre idiomas. Por esta razón, se ha vuelto cada vez más evidente que ya no podemos confiar en métricas como BLEU para proporcionar una estimación precisa de la calidad de MT (Barrault et al., 2019).

Mientras que un creciente interés de investigación en mé para entrenar modelos y sistemas de MT ha resultado en una reciente y dramática mejora en la calidad la evaluación de MT se ha quedado atrás. La comunidad de investigación de MT todavía depende en gran medida de me y no ha surgido ningún nuevo estándar ampliamente adopta En 2019, la Tarea Compartida de Traducción de Noticias d recibió un total de 153 envíos de sistemas de MT (Barrault et al., 2019). La Tarea Compartida de M del mismo año vio solo 24 submisiones, casi la mitad de las cuales fueron participantes en Tarea Compartida de Estimación de Calidad, adaptada com (Ma et al., 2019).

Los hallazgos de la tarea mencionada anteriormente des tacan dos desafíos principales para la evaluación de MT que buscamos abordar aquí (Ma et al., 2019): Es decir, que las métricas actuales 1) luchan por correlacionarse con precisión con el juicio humano a nivel de segmento y 2) no logran diferenciar adecuadamente los sistemas de MT de mayor rendimiento.

En este trabajo, presentamos COMET[1], un marco basado en Py para entrenar modelos de evaluación de MT altamente mul y adaptables que pueden funcionar como métricas. Nuestro marco se beneficia de los recientes avances en el modelado de lenguaje cruza (XLM) (Artetxe y Schwenk, 2019; Devlin et al., 2019; Conneau y Lample, 2019; Conneau et al., 2019) para generar estimaciones de predicción de juicios humanos como Evaluaciones Directas (DA) (Graham et al., 2013), Tasa de Edición de Traducción Mediada (HTER) (Snover et al., 2006) y métricas compatibles como Marco de Métricas de Calidad Multidimensional (MQM) (Lommel et al., 2014).

Inspirados por trabajos recientes sobre Estimación de Ca (QE) que demostraron que es posible lograr altos niveles de correlación con juicios humanos incluso sin una traducción de referencia (Fonseca et al., 2019), proponemos un enfoque novedoso para incorporar-

[1] Cruzado Optimizado Métrico para Evaluación de Traducción.

Actas de la Conferencia 2020 sobre Métodos Empíricos en el Procesamiento del Lenguaje Natural, páginas 2685–2702,
16–20 de noviembre, 2020

ing la entrada en idioma fuente en nuestros modelos de evaluación MT. Tradicionalmente, solo los modelos QE han hecho uso de la entrada de la fuente, mientras que las métricas de evaluación MT basadas en cambio en la traducción de referencia. Como Takahashi et al. (2020), mostramos que el uso de un espacio de incrustación multilingüe con el objetivo de aprovechar la información de las tres entradas y demostrar el valor añadido por la fuente como entrada a nuestros modelos de evaluación MT.

Para ilustrar la eficacia y flexibilidad de el COMET marco, entrenamos tres modelos que estiman diferentes tipos de juicios humanos y muestran un progreso prometedor hacia una mejor correlación-lación a nivel de segmento y robustez a alta-calidad MT.

Vamos a liberar tanto el COMET marco y los modelos de evaluación MT entrenados descritos en este documento a la comunidad de investigación tras su publicación.

2 Arquitecturas de Modelos

Los juicios humanos sobre la calidad de MT suelen venir en forma de puntuaciones a nivel de segmento, como DA, MQM y HTER. Para DA, es práctica común convertir las puntuaciones en clasificaciones relativas (DARR), cuando el número de anotaciones por segmento es limitado (Bojar et al., 2017b; Ma et al., 2018, 2019). Esto significa que, para dos hipótesis h_i y h_j de la misma fuente s, si la puntuación DA asignada a h_i es mayor que la puntuación asignada a h_j, se considera como una hipótesis "mejor".[2] Para abarcar estas diferencias, nuestro marco admite dos distintas arquitecturas: El modelo Estimador y el modelo de Clasificación de Traducción. La fundamental diferencia entre ellos es el objetivo de entrenamiento. Mientras que el Estimador se entrena para regresar directamente en una puntuación de calidad, el modelo de Clasificación de Traducción se entrena para minimizar la distancia entre una hipótesis "mejor" y tanto su referencia correspondiente como su fuente original. Ambos modelos están compuestos por un codificador cruzado de lenguaje y una capa de agrupación.

2.1 Codificador Cruzado de Lenguaje

El bloque de construcción principal de todos los modelos en nuestro marco es un modelo cruzado de lenguaje BERT multilingüe (Devlin et al., 2019), XLM (Conneau y Lample, 2019) o XLM-RoBERTa (Conneau et al., 2019). Estos modelos contienen varias capas de codificador transformador que son entrenados para reconstruir tokens enmascarados al descubrir la relación entre esos tokens y los ejemplos de entrada. Cuando se entrenan con datos de múltiples idiomas, este objetivo preentrenado ha demostrado ser altamente efectivo en tareas interlingüísticas como la clasificación de documentos y la inferencia de lenguaje natural (Conneau et al., 2019), generalizando bien a idiomas y guiones no vistos (Pires et al., 2019). Para los experimentos en este documento, confiamos en XLM-RoBERTa (base) como nuestro codificador modelo.

Dada una secuencia de entrada $x = [x_0, x_1, ..., x_n]$, el codificador produce una incrustación para cada token x_j y cada capa $\ell \in \{0, 1, ..., k\}$. En nuestro marco, aplicamos este proceso a la fuente, hipótesis MT, y referencia para mapearlos a un espacio de características compartido.

2.2 Capa de Agrupación

Los incrustaciones generadas por la última capa de los codificadores preentrenados suelen utilizarse para el ajuste de modelos a nuevas tareas. Sin embargo, Tenney et al. (2019) mostró que diferentes capas dentro de la red pueden capturar información lingüística que es relevante para diferentes tareas posteriores. En el caso de la evaluación MT, Zhang et al. (2020) demostró que diferentes capas pueden lograr diferentes niveles de correlación y que utilizar solo la última capa a menudo resulta en un rendimiento inferior. En este trabajo, nosotros utilizamos el enfoque descrito en Peters et al. (2018) y agrupamos información de las capas de codificación más una única incrustación para cada token x_j, utilizando un mecanismo de atención por capas. Esta incrustación se calcula entonces como:

$$e_{x_j} = \mu E_{\ell} \cdot \quad (1)$$

donde es un coeficiente de peso entrenable, $= [\ell^{(0)}, \ell^{(1)}, ..., \ell^{(k)}]$ corresponde al vector de incrustaciones de capa para el token x_j, y $= \text{softmax}([\cdot^{(1)}, \cdot^{(2)}, \ldots, \cdot^{(k)}])$ es un vector correspondiente a probabilidad con una probability el peso se establece en . Finalmente, como en Peters y Gurevych (2019), aplicamos un promedio de agrupación a la palabra resultante incrustaciones para derivar una incrustación de oración para cada segmento.

[2] En la Tarea de Métricas Compartidas WMT, si la diferencia entre las puntuaciones DA no es mayor a 25 puntos, esos segmentos son excluidos de la DA Datos RR.

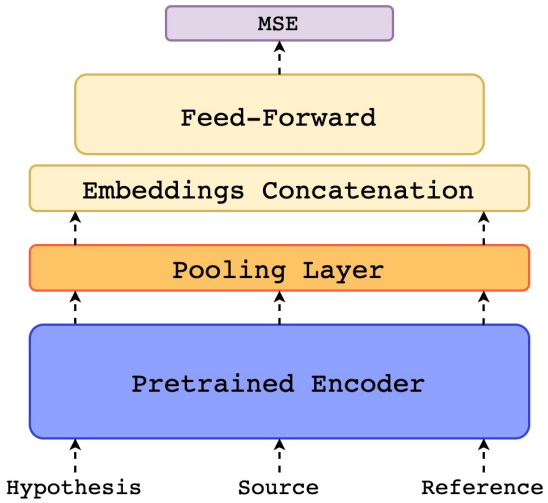

Figura 1: Arquitectura del modelo estimador. La fuente, la hipótesis y la referencia se codifican independientemente usando un codificador cruzado preentrenado. El resultado son incrustaciones de palabras que luego se pasan a través de una capa de agrupación para crear una incrustación de oración para cada segmento. Finalmente, las incrustaciones de oraciones resultantes se combinan y concatenan en un solo vector que se pasa a un regresor de avance. Todo el modelo es entrenado minimizando el Error Cuadrático Medio (MSE).

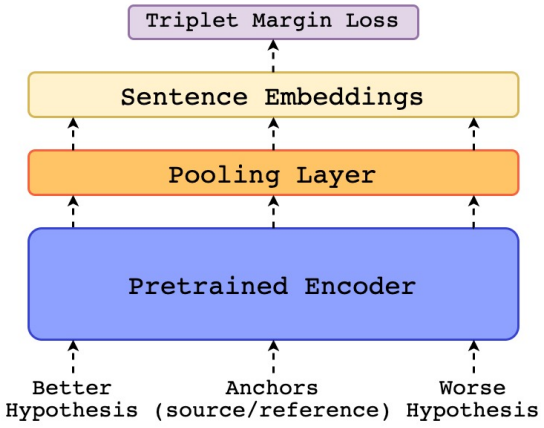

Figura 2: Arquitectura del modelo de Clasificación de Traducció Esta arquitectura recibe 4 segmentos: la fuente, la referencia, una hipótesis "mejor" y una "peor". Estos segmentos se codifican independientemente usando un co entrenado y una capa de agrupación en la parte superior. Finalmente, utilizando la pérdida de margen de Schroff et al. (2015) optimizamos el espacio de incrustación resultante para minimizar la distancia entre la hipótesis "mejor" y las "anclas" (fuente y referencia).

2.3 Modelo Estimador

Dado una d-incrustación de oración dimensional 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 por elemento: $s \cdot h$
- Producto de referencia elemento por elemento: $h \cdot r$
- Diferencia absoluta de la fuente elemento por elemento: $|h - s|$
- Diferencia absoluta de referencia elemento por elemento: $|h - r|$

Luego se concatenan estas características combinadas a la incrustación de referencia y la hipótesis embebida h en un solo vector $x = [h; r; h \cdot s; h \cdot r; |h - s|; |h - r|]$ que sirve como entrada para un regresor de avance rápido. La fuerza de estas características radica en destacar las diferencias entre las incrustaciones en el espacio de características semánticas.

Luego se entrena el modelo para minimizar la media del error cuadrado entre las puntuaciones predichas y las evaluaciones de calidad (DA, HTER o MQM). Figura 1 ilustra la arquitectura propuesta.

Nota que elegimos no incluir la fuente cruda incrustación (s) en nuestra entrada concatenada. Temprano la experimentación reveló que el valor añadido por la incrustación de la fuente como características de entrada regresor fue insignificante en el mejor de los casos. Una va nuestro modelo de estimador HTER entrenado con el vector $x = [h; s; r; h \cdot s; h \cdot r; |h - s|; |h - r|]$ como entrada para el avance solo logra impulsar- el rendimiento a nivel de segmento en 8 de los 18 pares de idiomas descritos en la sección 5 a continuación y el la mejora promedio en el Tau de Kendall en esos conjuntos fue de +0.0009. Como se señaló en Zhang et al. (2020), mientras que los modelos preentrenados multilingües son a a varios idiomas, el espacio de características entre los idiomas está mal alineado. Sobre esta base decidimos- a favor de excluir el incrustación de origen sobre la intuición de que la información más importante proviene de la incrustación de referencia y reduc- ir el espacio de características permitiría al modelo enfocarse más en la información relevante. Esto no sin embargo, niega el valor general de la fuente para nuestro modelo; donde incluimos características de combina tales como $s \cdot y$ $|h - s|$notamos ganancias en correlación como se explora más adelante en la sección 5.5.

2.4 Modelo de Clasificación de Traducción

Nuestro modelo de Clasificación de Traducción (Figura 2) recibe como entrada una tupla (s, h^+, h^-, r) donde h^+ denota una hipótesis que fue clasificada más alta que otra hipótesis h^-. Luego pasamos a través de nuestro codificador y capa de agrupación cross-lingual para tener un incrustación de oración para cada segmento en $\{s, h^+, h^-, r\}$. Finalmente, usando las incrustaciones, calculamos la pérdida de margen de triplete (Schroff et al., 2015) en relación con la fuente y la referencia:

$$L(\cdot) = L(s, h^+, h^-) + L(r, h^+, h^-) \quad (2)$$

donde:

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

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

$d(u, v)$ denota la distancia euclidiana entre u y v y $\cdot$ es un margen. Por lo tanto, durante el entrenamiento el modelo optimiza el espacio de incrustación para que la distancia entre los anclajes $(s$ y $r)$ y el "peor" hipótesis h^- es mayor al menos por $\cdot$ que la distancia entre los anclajes y la hipótesis "mejor" h^+.

Durante la inferencia, el modelo descrito recibe un trío (s, h, r) con solo una hipótesis. La puntuación de calidad asignada es la media armónica entre la distancia a la fuente $d(s, h)$ y la distancia a la referencia $d(r, h)$:

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

Finalmente, convertimos la distancia resultante en un puntaje de similitud limitado entre 0 y 1 como sigue:

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

3 Corpora

Para demostrar la efectividad de nuestras descritas arquitecturas de modelo (Sección 2), entrenamos tres MT modelos de evaluación donde cada modelo apunta a un diferente tipo de juicio humano. Para entrenar estos modelos, utilizamos datos de tres corpora diferentes: el corpus QT21, el DA RR del WMT Metrics tarea compartida (2017 a 2019) y un propietario Corpus anotado MQM.

3.1 El corpus QT21

El corpus QT21 es públicamente disponible[3] conjunto de datos que contiene oraciones generadas por la industria de ya sea un dominio de tecnología de la información o ciencias de la v (Specia et al., 2017). Este corpus contiene un total de 173K tuplas con oración fuente, respectiva referencia generada por humanos, hipótesis MT (ya sea de un MT estadístico basado en frases o de un neural MT), y MT post-editado (PE). Los pares de idiomas representados en este corpus son: inglés a alemán (en-de), letón (en-lt) y checo (en-cs), y alemán a inglés (de-en).

La puntuación HTER se obtiene calculando la tasa de edición de traducción (TER) (Snover et al., 2006) entre la hipótesis MT y el PE correspondiente. Finalmente, después de calcular el HTER para cada MT, construimos un conjunto de datos de entrenamiento $D = \{s_i, h_i^+, h_i^-, r_i\}_{n=1}^{N}$, donde s_i denota el texto fuente, h_i denota la MT hipótesis, r_i la traducción de referencia, y y_i la puntuación HTER para la hipótesis h_i. De esta manera buscamos aprender una regresión $f(s, h, r) \cdot$ que predice el esfuerzo humano requerido para corregir la hipótesis mirando la fuente, hipótesis, y referencia (pero no la hipótesis post-editada).

3.2 El WMT DA RR corpus

Desde 2017, los organizadores de la Tarea Compartida de Tarea Compartida de Traducción (Bojar et al., 2019) han recogido juicios humanos en forma de adecuaciones DA (Graham et al., 2013, 2014, 2017). Estos DAs se mapean entonces en rangos relativos-ings (DARR) (Ma et al., 2019). El resultado datos para cada año (2017-19) forman un conjunto de dato $D = \{s_i, h_i^+, h_i^-, r_i\}_{n=1}^{N}$ donde h_i^+ denota una hipótesis "mejor" y h_i^- denota una "peor". Aquí buscamos aprender una función $f(s, h, r)$ de tal manera que el puntaje asignado a h_i^+ es estrictamente mayor que el puntaje asignado 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 recursos como el chino al inglés (zh-en) y el inglés al gujarati (en-gu).

3.3 El corpus MQM

El corpus MQM es una base de datos interna propietaria de traducciones generadas por MT de soporte al cliente

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

[4] Los datos brutos para cada año de la tarea compartida de métricas están disponibles públicamente en la página de resultados (ejemplo de ample http://www.statmt.org/wmt19/results.html). Sin embargo, ten en cuenta que en los archivos se resalta que esto ocasionalmente requieren utilidades personalizadas que no están dispo 6.

mensajes de chat que fueron anotados de acuerdo con las pautas establecidas en Burchardt y Lommel (2014). Estos datos contienen un total de 12K tuplas, cubriendo 12 pares de idiomas del inglés al: alemán (en-de), español (en-es), español latinoamericano (en-es-latam), francés (en-fr), italiano (en-it), japonés (en-ja), holandés (en-nl), portugués (en-pt), portugués brasileño (en-pt-br), ruso (en-ru), sueco (en-sv), y turco (en-tr). Nota que en este corpus, el inglés siempre se ve como el idioma fuente, pero nunca como el idioma objetivo. Cada tupla consta de una oración fuente, una referencia generada por humanos, una hipótesis de MT, y su puntuación MQM, derivada de las anotaciones de error por uno (o más) anotadores entrenados. La métrica MQM a la que nos referimos a lo largo de este documento es una métrica interna definida de acuerdo con el marco MQM (Lommel et al., 2014) (MQM). Los errores se anotan bajo una tipología interna definida bajo tres tipos de errores principales: 'Precisión', 'Estilo', 'Fluidez' y 'Precisión'. Nuestros puntajes MQM varían de a 100 y se definen como:

$$\text{MQM} = 100 \cdot \left(1 - \frac{\text{Menor} \times 1 + \text{Mayor} \times 5 + \text{Crít.} \times 10}{\text{Longitud de la oración}} \right) \quad (7)$$

donde Yo_{Menor} denota el número de errores menores, Yo_{Mayor} el número de errores mayores y $\text{Yo}_{\text{Crít.}}$ el número de errores críticos.

Nuestra métrica MQM tiene en cuenta la severidad de los errores identificados en la hipótesis MT, lo que lleva a una métrica más detallada que HTER o DA. Cuando se usaron en nuestros experimentos, estos valores se dividieron por 100 y se truncaron a 0. Como en la sección 3.1, construimos 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 MT, r_i la traducción de referencia, y y_i la puntuación MQM para la hipótesis h_i.

4 Experimentos

Entrenamos dos versiones del modelo Estimador descrito en la sección 2.3: uno que regresa en HTER (COMET-HTER) entrenado con el corpus QT21, y otro que regresa en nuestra implementación propia de MQM (COMET-MQM) entrenado con nuestro corpus MQM interno. Para el modelo de Clasificación de Traducción, descrito en la sección 2.4, entrenamos con el WMTDA RR corpus de 2017 y 2018 (COMET-RANK). En esta sección, introducimos la configuración de entrenamiento para estos modelos y la correspondiente evaluación de la configuración.

4.1 Configuración de Entrenamiento

Las dos versiones de los Estimadores (COMET-HTER/MQM) comparten la misma configuración de entrena hiper-parámetros (los detalles están incluidos en los Apéndices). Para el entrenamiento, cargamos el codificador p e inicializamos tanto la capa de agrupación como el regresor de avance. Mientras que los escalares de capa p de la capa de agrupación se establecen inicialmente en 1/n, donde n es el número de capas, los pesos del avance se inicializan aleatoriamente. Durante el entrenamiento, dividimos los parámetros del modelo en dos grupos: los parámetros del c parámetros del modelo, que incluyen el modelo del codificad parámetros del modelo, que incluyen el modelo del codifica escalares de capa; y los parámetros del regresor, que incluyen los parámetros de la red de avance directo superio red. Aplicamos descongelamiento gradual y discriminativas tasas de aprendizaje (Howard y Ruder, 2018), lo que significa que el modelo de codificador está congelad época mientras que el avance directo se optimiza con una tasa de aprendizaje de 3e-5. Después de la primera época, el modelo completo se ajusta finamente pero la tasa de apren los parámetros del codificador se establece en 1e-5 para evitar el olvido cata En contraste con los dos Estimadores, para el COMET-RANK además, debido a que este modelo no añade ning nuevo parámetro encima de XLM-RoBERTa (base) aparte de los escalares de capa, usamos una tasa de aprendizaje de 1e-5 para todo el modelo.

Configuración de Evaluación. Utilizamos los datasets de prueba y la configuración de la (Ma et al., 2019) para comparar el COMET-RANK modelos con el mejor rendimiento en las presentaciones compartidas de la tarea y otras métricas de vanguardia como BERTSCORE y BLEURT.[5] El método de evaluación utilizado es la fórmula oficial de Kendall Tau, $\cdot$, del WMT 2019 Metrics Shared Task (Ma et al., 2019) definido como:

$$\cdot = \frac{\text{Concordante} - \text{Discordante}}{\text{Concordante} + \text{Discordante}} \quad (8)$$

donde Concordante es el número de veces que una métrica asigna una puntuación más alta a la hipótesis "mejor" y Discordante es el número de veces que una métrica asigna una puntuación más alta a la hipótesis "peor"

[5] Para facilitar la investigación futura, también proporcionaremos, den marco, instrucciones detalladas y scripts para ejecutar otras métricas como CHR F, BLEU, BERTSCORE, y BLEURT

Tabla 1: Tau de Kendall (τ) correlaciones en pares de idiomas con inglés como fuente para las Métricas DARR WMT19 corpus. Para BERTSCORE informamos los resultados con el modelo de codificador predeterminado para una comparación con XLM-RoBERTa (base) para ser justos con nuestros modelos. Los valores reportados para YiSi-1 se toman directamente del documento de la tarea compartida (Ma et al., 2019).

Métrica	en-cs	en-de	en-··	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(predeterminado)	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 hipótesis son las mismas.

Como se mencionó en los hallazgos de (Ma et al., 2019), las correlaciones a nivel de segmento de todas las métricas presentadas fueron frustrantemente bajas. Además, todas las métricas presentadas mostraron una dramática falta de capacidad para clasificar correctamente los sistemas de MT fuertes. Para evaluar si nuestros nuevos modelos de evaluación de MT abordan mejor este problema, seguimos la configuración de evaluación descrita utilizada en el análisis presentado en (Ma et al., 2019), donde se examinan los niveles de correlación para partes de los DARR datos que incluyen solo los 10, 8, 6 y 4 sistemas MT superiores.

5 Resultados

5.1 Del inglés al X

Tabla 1 muestra resultados para todos los ocho pares de idiomas con inglés como fuente. Contrastamos nuestros tres COMET modelos contra métricas de referencia como BLEU y CHRF, la métrica ganadora de la tarea 2019 YISI-1, así como la más reciente BERTSCORE. Observamos que en general nuestros tres modelos entrenados con el marco de trabajo COMET superan, a menudo por márgenes significativos, todas las demás métricas. Nuestro DARR Ranker supera a los dos Estimadores en siete de los ocho pares de idiomas. Además, incluso aunque el Estimador MQM se entrena solo en 12K segmentos anotados, se desempeña aproximadamente al mismo nivel que el Estimador HTER para la mayoría de los pares de idiomas, y supera todas las demás métricas en en-ru.

5.2 De X al inglés

Tabla 2 muestra resultados para los siete idiomas a inglés pares de idiomas. De nuevo, contrastamos nuestros tres COMET modelos contra métricas de referencia como BLEU y CHRF, la métrica ganadora de la tarea 2019 YISI-1, así como las métricas recientemente publicadas BERTSCORE y BLEURT. Como en la Tabla 1 el DARR El modelo RR muestra fuertes correlaciones con los juicios humanos, superando al recientemente propuesto específico para inglés BLEURT métrica en cinco de siete pares de idiomas. De nuevo, el Estimador MQM muestra sorprendentemente fuertes resultados a pesar de que este modelo fue entrenado con datos que no incluían inglés como objetivo. Aunque el codificador utilizado en nuestros modelos entrenados es altamente multilingüe, hipotetizamos que este poderoso resultado "zero-shot" se debe a la inclusión de la fuente en nuestros modelos.

5.3 Pares de idiomas que no involucran inglés

Nuestros 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 probamos en los tres pares de idiomas WMT 2019 que no incluyen inglés en ninguna fuente o objetivo. Como se puede ver en la Tabla 3, nuestros resultados son consistentes con las observaciones en las Tablas 1 y 2.

5.4 Robustez ante MT de Alta Calidad

Para el análisis, utilizamos el DARR corpus del WMT 2019 Tarea Compartida 2019 y evaluamos en el subconjunto de los datos de los sistemas MT de mejor rendimiento para cada par de idiomas. Incluimos pares de idiomas para los cuales pudimos recuperar datos de al menos diez diferentes sistemas MT (es decir, todos menos kk-en y gu-en). Contrastamos contra el fuerte recientemente propuesto BERTSCORE y BLEURT, con BLEU como base-línea. Los resultados se presentan en la Figura 3. Para los pares donde el inglés es el objetivo, nuestros tres modelos son mejores o competitivos con todos los demás; donde el inglés es la fuente, notamos que en general nuestras métricas superan el rendimiento de otros-

Tabla 2: Tau de Kendall ( ) correlaciones en pares de idiomas con inglés como objetivo para las Métricas DA RR WMT19 corpus. En cuanto a BERTSCORE, para BLEURT informamos los resultados de dos modelos: el modelo base, que es comparable en tamaño con el codificador que usamos y el modelo grande que es el doble de tamaño.

Métrica	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(predeterminado)	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: Tau de Kendall ( ) correlaciones en idioma pares no involucrados en inglés para las Métricas WMT19 DACorpus RR.

Métrica	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(predeterminado)	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

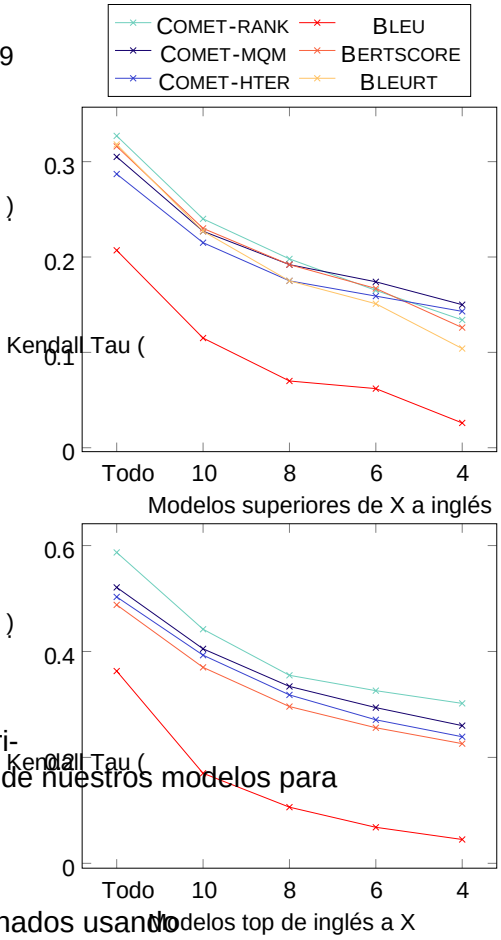

Figura 3: Rendimiento de las métricas en general y en los mejore 8, 6 y 4) sistemas MT.

ers. Incluso el Estimador MQM, entrenado solo con 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 del input del lenguaje fuente en la capacidad de nuestros modelos para aprender predicciones precisas, entrenamos dos versiones de nuestro modelo RR Ranker: uno que usa solo la referencia, y otro que usa tanto la referencia como la fuente. Ambos modelos fueron entrenados usando el corpus WMT 2017 que solo incluye pares de idiomas del inglés (en-de, en-cs, en-·, en-tr). En otras palabras, mientras que el inglés nunca fue observado como un idioma objetivo durante el entrenamiento para 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 ciones (con la excepción de en-cs porque cs-en no existe para WMT 2018). Los resultados en la Tabla 4 claramente muestra que para la arquitectura de clasificació la fuente mejora la correlación general con los juicios humanos. Además, la inclusión de la fuente expuso la segunda variante del modelo a los incrustaciones en inglés que es

Tabla 4: Comparación entre COMET-RANK (sección 2.4) y una versión solo de referencia de la misma en datos WMT18. Ambos modelos fueron entrenados con WMT17 lo que significa que el modelo solo de referencia nunca se expone al i durante el entrenamiento.

Métrica	en-cs	en-de	en-··	en-tr	cs-en	de-en	··-en	tr-en
COMET-RANK (solo ref.)	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
· ·	0.051	0.035	0.041	0.024	0.107	0.155	0.119	0.132

reflejado en un más alto para los pares de idiomas con inglés como objetivo.

6 Reproducibilidad

Vamos a liberar tanto la base de código de COMET como el marco y los modelos de evaluación de MT entrenados descritos en este documento a la comunidad de investigación tras la publicación, junto con los scripts detallados necesarios para ejecutar todas las líneas de base reportadas.[6] Todos los modelos reportados en este documento fueron entrenados en un único Tesla T4 (16GB) GPU. Además, nuestro marco de trabajo se basa en PyTorch Lightning (Falcon, 2019), un envoltorio ligero de PyTorch, que fue creado para la máxima flexibilidad y reproducibilidad.

7 Trabajo relacionado

Las métricas clásicas de evaluación de MT se caracterizan comúnmente como caracterizadas métricas de coincidencia de n-gramas que utilizando características hechas a mano, estiman la calidad de MT-calidad contando el número y la fracción de n-gramas que aparecen simultáneamente en una candidata hipótesis de traducción y una o más humanas-referencias. Métricas como BLEU (Papineni et al., 2002), METEOR (Lavie y Denkowski, 2009), y CHRF (Popović, 2015) han sido ampliamente estudiados y mejorados (Koehn et al., 2007; Popović, 2017; Denkowski y Lavie, 2011; Guo y Hu, 2019), pero, por diseño, generalmente fallan en reconocer y capturar la similitud semántica más allá del nivel léxico.

En años recientes, los incrustaciones de palabras (Mikolov et al., 2013; Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2019) han surgido como una común alternativa utilizada para coincidencia para capturar la similitud semántica de las palabras. Las métricas basadas en métricas 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 referencia e hipótesis en un espacio de incrustación y luego calcular una puntuac que refleja la similitud semántica entre esos segmentos. Sin embargo, los juicios humanos como DA y MQM, capturan mucho más que solo semantic similarity, resultando en una correlación superior-límite entre los juicios humanos y las puntuaciones producidas por tales métricas.

Métricas aprendibles (Shimanaka et al., 2018; Mathur et al., 2019; Shimanaka et al., 2019) intentan optimizar directamente la correlación con juicios humanos, y han mostrado recientemente promising results. BLEURT (Sellam et al., 2020), una métrica aprendible basada en BERT (Devlin et al., 2019), reclama un rendimiento de vanguardia durante los últimos 3 años de la tarea compartida de métricas WMT. Porque BLEURT se construye sobre English-BERT (Devlin et al., 2019), solo puede ser utilizado cuando el inglés es el idioma objetivo que limita su aplicabilidad. Además, hasta donde sabemos, todos los anteriormente propuestos métricas aprendibles se han centrado en optimizar DA que, debido a la escasez de anotadores, puede resultar inherentemente ruidoso (Mathur et al., 2019).

Evaluación de MT sin referencia, también conocida como Estimación de Calidad (QE), históricamente a menudo regresado en HTER para la evaluación a nivel de segmento (Blatz et al., 2003, 2004, 2005, 2006, 2007a). Más recientemente, MQM ha sido utilizado para la evaluación a evaluación (Specia et al., 2018; Fonseca et al., 2019). Al aprovechar codificadores pre-entrenados altamente multilingües como BERT multilingüe (Devlin et al., 2019) y XLM (Conneau y Lample, 2019), los sistemas QE han estado mostrando auspiciosas correlaciones con los juicios humanos (Kepler et al., 2019a). Concurrentemente, el OpenKiwi marco (Kepler et al., 2019b) ha facilitado para los investigadores avanzar en el campo y construir modelos QE más fuertes.

8 Conclusiones y Trabajo Futuro

En este documento presentamos COMET, un nuevo marco neural para entrenar modelos de evaluación MT que pueden servir como métricas automáticas y ser fácilme

[6] Estos serán alojados en https://github.com/Unbabel/COMET

adaptado y optimizado para diferentes tipos de humanos juicios de calidad de MT.

Para mostrar la efectividad de nuestro marco, buscamos abordar los desafíos reportados en el 2019 WMT Metrics Shared Task (Ma et al., 2019). Entrenamos tres modelos distintos que logran nuevos resultados de vanguardia para la correlación a nivel de segmento con juicios humanos, y muestran una capacidad prometedora para diferenciar mejor los sistemas de alto rendimiento.

Uno de los desafíos de aprovechar el poder de los modelos preentrenados es el peso oneroso de parámetros y tiempo de inferencia. Un camino principal para el trabajo futuro en COMET se centrará en el impacto de soluciones más compactas como DistilBERT (Sanh et al., 2019).

Además, mientras esbozamos la posible importancia del texto fuente anterior, notamos que nuestro COMET-RANK modelo pondera la fuente y la referencia de manera diferente durante la inferencia pero igual en su función de pérdida de entrenamiento. El trabajo futuro optimalidad de esta formulación y examinará más a fondo la interdependencia de las diferentes entradas.

Agradecimientos

Estamos agradecidos a Andr´e Martins, Chrysoula Zerva, Daan van Stigt, Miguel Vera, los revisores, por sus valiosos comentarios y discusiones. Este trabajo fue apoyado en parte por el Programa P2020 a través de los proyectos MAIA y Unbabel4EU, supervisados por ANI bajo los números de contrato 045909 y 042671, respectivamente.

Referencias

Mikel Artetxe y Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 7:597–610.

Lo¨ıc Barrault, Ond·rej Bojar, Marta R. Costa-juss`a, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias M¨uller, Santanu Pal, Matt Post, y Marcos Zampieri. 2019. Hallazgos de la conferencia 2019 sobre traducción automática (WMT19). En Actas de la Cuarta Conferencia sobre Traducción Automática (Volumen 2: Documentos de Tareas Compartidas, Día 1), páginas 1–61, Florencia, Italia. Asociación para la Lingüística Computacional.

Ond·rej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, y Lucia Specia. 2013. Hallazgos del Taller 2013 sobre Traducción Automática Estadística. En Actas del Octavo Taller sobre Traducción Automática Estadística, páginas 1–44, Sofía, Bulgaria. Asociación para la Lingüística Computacional.

Ond·rej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, y Ale·s Tamchyna. 2014. Hallazgos del taller de 2014 sobre traducción automática estadística. En Actas del Noveno Taller sobre Traducción Automática, páginas 12–58, Baltimore, Maryland, EE. UU. Asociación para la Lingüística Computacional.

Ond·rej Bojar, 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, y Marco Turchi. 2017a. Hallazgos de la conferencia de 2017 sobre la traducción automática (WMT17). En Actas de la Segunda Conferencia sobre Traducción Automática, páginas 169–214, Copenhague, Dinamarca. Asociación para la Computación Lingüística.

Ond·rej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aur´elie N´ev´eol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, y Marcos Zampieri. 2016. Hallazgos de la conferencia de 2016 sobre traducción automática. En Actas de la Primera Conferencia sobre Traducción Automática: Volumen 2, Documentos de Tareas Compartidas, páginas 131–198, Berlín, Alemania. Asociación para la Computación Lingüística.

Ond·rej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, y Marco Turchi. 2015. Hallazgos del taller de 2015 sobre traducción automática estadística. En Actas del Décimo Taller sobre Traducción Automática Estadística, páginas 1–46, Lisboa, Portugal. Asociación para la Computación Lingüística.

Ond·rej Bojar, Yvette Graham, y Amir Kamran. 2017b. Resultados de la tarea compartida de métricas WMT17. En Actas de la Segunda Conferencia sobre Traducción Automática, páginas 489–513, Copenhague, Dinamarca. Asociación para la Lingüística Computacional.

Aljoscha Burchardt y Arle Lommel. 2014. Prácticas Directrices para el Uso de MQM en la Investigación Científica y Profesional relacionada con 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 Zettlemoyer, y Veselin Stoyanov. 2019. Aprendizaje no supervisado de representación cruzada de idiomas a gran escala. arXiv preimpresión arXiv:1911.02116.

Alexis Conneau y Guillaume Lample. 2019. Preentrenamiento del modelo de lenguaje multilingüe. En H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, y R. Garnett, editores, Avances en los Sistemas de Procesamiento de Información Neural 32, páginas 7059–7069. Curran Associates, Inc.

Jeremy Howard y Sebastian Ruder. 2018. Ajuste fino de modelos de lenguaje universal para la clasificación de texto. En Actas de la 56ª Reunión Anual de la Asociación para la Lingüística Computacional (Volumen 1: Documentos largos), páginas 328-339, Melbourne, Australia. Asociación para la Lingüística Computacional.

Meteor 1.3: Métrica automática para la optimización de sistemas de traducción. En Actas del Sexto Taller sobre Traducción Automática, páginas 85–91, Edimburgo, Escocia, Reino Unido.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, Ant´onio G´ois, M. Amin Farajian, Ant´onio V. Lopes, y Andr´e F. T. Martins. 2019a. Participación de Unbabel en la tarea compartida de estimación de calidad de traducción. En Actas de la Cuarta Conferencia sobre Traducción Automática (Volumen 3: Documentos de Tarea Compartida, Día 2), páginas 78–84, Florence, Italia. Asociación para la Lingüística Computacional.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, y Kristina Toutanova. 2019. BERT: Preentrenamiento de transformadores bidireccionales profundos para la comprensión del lenguaje. En Actas de la Conferencia 2019 del Capítulo Norteamericano de la Asociación para la Lingüística Computacional: Tecnologías del Lenguaje Humano, Volumen 1 (Papers Largos y Cortos), páginas 4171–4186, Minneapolis, Minnesota. Asociación para la Lingüística Computacional.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, y Andr´e F. T. Martins. 2019b. OpenKiwi: Un marco de trabajo de código abierto para la estimación de calidad. En Actas de la 57ª Reunión Anual de la Asociación para la Lingüística Computacional: Demostraciones, páginas 117–122, Florencia, Italia. Asociación para la Lingüística Computacional.

WA Falcon. 2019. PyTorch Lightning: El envoltorio ligero de PyTorch para la investigación de IA de alto rendimiento. GitHub.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondˇrej Bojar, Alexandra Constantin, y Evan Herbst. 2007. Moses: Kit de herramientas de código abierto para la traducción automática. En Actas de la 45ª Reunión Anual de la Asociación para la Lingüística Computacional Sesiones de Demostración, páginas 177–180, Praga, República Checa. Asociación para la Lingüística Computacional.

WA Falcon. 2019. PyTorch Lightning: The lightweight PyTorch wrapper for high-performance AI research. GitHub.

Erick Fonseca, Lisa Yankovskaya, Andr´e F. T. Martins, Mark Fishel y Christian Federmann. 2019. Hallazgos de las tareas compartidas de WMT 2019 sobre estimación de calidad. En Actas de la Cuarta Conferencia sobre Traducción Automática (Volumen 3: Documentos de Tarea Compartida, Día 2), páginas 1–10, Florencia, Italia. Asociación para la Lingüística Computacional.

Dan Kondratyuk y Milan Straka. 2019. 75 lenguas, 1 modelo: Análisis de dependencias universales parseando cada idioma. En Actas de la Conferencia sobre Métodos Empíricos en el Procesamiento de Lenguaje Natural y la 9na Conferencia Conjunta Internacional sobre Procesamiento de Lenguaje Natural (EMNLP-IJCNLP), páginas 2779–2795, Hong Kong, China. Asociación para la Lingüística Computacional.

Yvette Graham, Timothy Baldwin, Alistair Moffat y Justin Zobel. 2013. Escalas de medición continua en la evaluación humana de la traducción automática. En Actas del 7º Taller de Anotación Lingüística e Interoperabilidad con el Discurso, páginas 33–41, Sofía, Bulgaria. Asociación para la Lingüística Computacional.

Alon Lavie y Michael Denkowski. 2009. El meteor métrico para la evaluación de la traducción automática. Traducción Automática, 23(2-3):105–115.

Yvette Graham, Timothy Baldwin, Alistair Moffat y Justin Zobel. 2014. ¿Está mejorando la traducción automática con el tiempo? En Actas de la 14ª Conferencia del Capítulo Europeo de la Asociación para la Lingüística Computacional, páginas 443–451, Gotemburgo, Suecia. Asociación para la Lingüística Computacional.

Chi-kiu Lo. 2019. YiSi - una métrica unificada de calidad de traducción para evaluación y estimación multilingües. En Actas de la Cuarta Conferencia sobre Traducción Automática (Volumen 2: Documentos de Pista Compartida), páginas 507–513, Florencia, Italia. Asociación para la Lingüística Computacional.

Yvette Graham, Timothy Baldwin, Alistair Moffat y Justin Zobel. 2017. ¿Pueden los sistemas de traducción automática ser evaluados solo por la multitud? Ingeniería del Lenguaje Natural, 23(1):330–341.

Arle Lommel, Aljoscha Burchardt y Hans Uszkoreit. 2014. Métricas de calidad multidimensional (MQM). En Actas de la Cuarta Conferencia sobre Traducción Automática (Volumen 1: Documentos de Investigación), páginas 455–463, Ginebra, Suiza. Asociación para la Lingüística Computacional.

Yinuo Guo y Junfeng Hu. 2019. Meteor++ 2.0: Adoptar conocimientos de paráfrasis a nivel sintáctico para una mejor evaluación de traducción automática. En Actas de la Cuarta Conferencia sobre Traducción Automática (Volumen 2: Documentos de Tarea Compartida, Día 1), páginas 501–506, Florence, Italia. Asociación para la Lingüística Computacional.

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

Qingsong Ma, Ond·rej Bojar, y Yvette Graham. 2018. Resultados de la tarea compartida de métricas WMT18: Ambos caracteres y incrustaciones logran un buen rendimiento. En Actas de la Tercera Conferencia sobre Traducción Automática: Documentos de Tarea Compartida, páginas 671–688, Bélgica, Bruselas. Asociación para la Computación Lingüística.

Qingsong Ma, Johnny Wei, Ond·rej Bojar, y Yvette Graham. 2019. Resultados de las métricas WMT19 tarea compartida: los sistemas de MT a nivel de segmento y fuertes presentan grandes desafíos. En Actas de la Cuarta Conferencia sobre Traducción Automática (Volumen 2: Documentos de Tarea Compartida), páginas 62–90, Florencia, Italia. Asociación para la Computación Lingüística.

Nitika Mathur, Timothy Baldwin, y Trevor Cohn. 2019. Poniendo la evaluación en contexto: Los incrustaciones contextuales mejoran la evaluación de la traducción. En Actas de la 57ª Reunión Anual de la Asociación para la Lingüística Computacional, páginas 2799–2808, Florencia, Italia. Asociación para la Computación Lingüística.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, y Jeff Dean. 2013. Representaciones distribuidas de palabras y frases y su composicionalidad. En Avances en Sistemas de Procesamiento de Información Neural 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 Actas de la 40ª Reunión Anual de la Asociación para la Computación Lingüística, páginas 311-318, Filadelfia, Pensilvania, EE. UU. Asociación para la Computación Lingüística.

Jeffrey Pennington, Richard Socher, y Christopher Manning. 2014. Glove: Vectores globales para la representación de palabras. En Actas de la Conferencia 2014 sobre Métodos Empíricos en el Procesamiento del Lenguaje Natural (EMNLP), páginas 1532-1543, Doha, Qatar. Asociación para la Computación Lingüística.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, y Luke Zettlemoyer. 2018. Representaciones de palabras contextualizadas profundamente. En Actas de la Conferencia 2018 del Capítulo Norteamericano de la Asociación para la Computación Lingüística: Tecnologías del Lenguaje Humano, Volumen 1 (Documentos largos), páginas 2227-2237, Nueva Orleans, Luisiana. Asociación para la Computación Lingüística.

Telmo Pires, Eva Schlinger, y Dan Garrette. 2019. ¿Qué tan multilingüe es BERT multilingüe? En Procedimientos de la 57ª Reunión Anual de la Asociación para la Computación Lingüística, páginas 4996-5001, Florencia, Italia. Asociación para la Computación Lingüística.

Maja Popovi·. 2015. chrF: puntuación f de n-gramas de caracteres para la evaluación automática de MT. En Actas del Décimo Taller de Traducción Automática Estadística, páginas 392–395, Lisboa, Portugal. Asociación para la Lingüística Computacional.

Maja Popovi´c. 2017. chrF++: palabras ayudando a los caracteres n-gramas. En Actas de la Segunda Conferencia de Traducción Automática, páginas 612–618, Copenhague, Dinamarca. Asociación para la Lingüística Computacional.

Nils Reimers e Iryna Gurevych. 2019. Sentence-BERT: Incrustaciones de frases usando redes Siamese BERT-networks. En Actas de la Conferencia 2019 sobre Métodos Empíricos en Procesamiento del Lenguaje Natural y la 9na Conferencia Internacional Conjunta sobre Procesamiento del Lenguaje Natural (EMNLP-IJCNLP), páginas 3982-3992, Hong Kong, China. Asociación para la 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. preimpresión arXiv:1910.01108

Schroff, D. Kalenichenko, y J. Philbin. 2015. Facenet: Una incrustación unificada para el reconocimiento y agrupación de caras. En Conferencia IEEE 2015 sobre Visión por Computadora y Reconocimiento de Patrones (CVPR), páginas 815–823.

Thibault Sellam, Dipanjan Das, y Ankur Parikh. 2020. BLEURT: Aprendiendo métricas robustas para la generación de textos. En Actas de la 58va Reunión Anual de la Asociación para la Lingüística Computacional, páginas 7881–7892, Online. Asociación para la Computación Lingüística.

Christophe Servan, Alexandre B´erard, Zied Elloumi, Herv´e Blanchon, y Laurent Besacier. 2016. Word2Vec vs DBnary: ¿Aumentando METEOR usando representaciones vectoriales o recursos léxicos? En Actas de COLING 2016, la 26ª Conferencia Internacional sobre Lingüística Computacional: Documentos Técnicos, páginas 1159–1168, Osaka, Japón. El Comité Organizador de COLING 2016.

Hiroki Shimanaka, Tomoyuki Kajiwara, y Mamoru Komachi. 2018. RUSE: Regresor utilizando oraciones incrustaciones para la evaluación automática de la traducción d. En Actas de la Tercera Conferencia sobre Traducción Automática: Documentos de Tarea Compartida, páginas 751–758, Bélgica, Bruselas. Asociación para la Computación Lingüística.

Hiroki Shimanaka, Tomoyuki Kajiwara, y Mamoru Komachi. 2019. Evaluación de la Traducción de Máquinas con BERT Regresor. preimpresión de arXiv arXiv:1907.12679

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla y John Makhoul. 2006. Un estudio de la tasa de edición de traducción con anotación humanación En En las Actas de la Asociación para la Traducción Automática en , páginas 223·231.

Lucia Specia, Fr´ed´eric Blain, Varvara Logacheva, Ram´on Astudillo y Andr´e F. T. Martins. 2018. Hallazgos de la tarea compartida WMT 2018 sobre la calidad estimación En Actas de la Tercera Conferencia sobre Traducción Automática: Documento , páginas Tareas Compartidas 689·709, Bélgica, Bruselas. Asociación para la Computación Lingüística.

Lucia Specia, Kim Harris, Fr´ed´eric Blain, Aljoscha Burchardt, Viviven Macketanz, Inguna Skadina, Matteo Negri, , y Marco Turchi. 2017. Traducción de calidad y productividad: Un estudio sobre idiomas ricos en morfología . En Cumbre de Traducción Automática XVI, páginas 55·71, Nagoya, Japón.

Kosuke Takahashi, Katsuhito Sudoh y Satoshi Nakamura. 2020. Evaluación automática de la traducción utilizando entradas en el idioma fuente y modelo de idioma cruzado lingüístico . En Actas de la 58ª Reunión Anual de la Asociación para la Computación Lingüística , páginas 3553·3558, En línea. Asociación para la Computación Lingüística.

Andre T¨attar y Mark Fishel. 2017. bleu2vec: la métrica dolorosamente familiar en el espacio vectorial continuo esteroides En Actas de la Segunda Conferencia sobre Traducción Automática , páginas 619·622, Copenhague, Dinamarca. Asociación para la Computación Lingüística.

Ian Tenney, Dipanjan Das y Ellie Pavlick. 2019. BERT redescubre el clásico pipeline de NLP En Actas de la 57ª Reunión Anual de la Asociación para la Lingüística Computacional , páginas 4593-4601, Florencia, Italia. Asociación para la Lingüística Computacional.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, y Yoav Artzi. 2020. Bertscore: Evaluando la generación de texto con En Conferencia Internacional sobre Representaciones de Aprendizaje . Wei Zhao, Goran Glava·s, Maxime Peyrard, Yang Gao,

Robert West, y Steffen Eger. 2020. Sobre las limitaciones de los codificadores cruzados como se expone por la evaluación de la traducción automática sin referencia . En Actas de la 58ª Reunión Anual de la Asociación para la Lingüística Computacional , páginas 1656-1671, En línea. Asociación para la Lingüística Computacional.
Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, y Steffen Eger. 2019. MoverScore: Evaluación de la generación de texto con incrustaciones contextualizadas y la dis . En Actas de la Conferencia 2019 sobre Métodos Empíricos en Procesamiento del Lenguaje Natural y la 9ª Conferencia Internacional Conjunta sobre Procesamiento del Lenguaje Natural (EMNLP-IJCNLP) , páginas 563-578, Hong Kong, China. Asociación para la Lingüística Computacional.

A Apéndices

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

Tabla 5: Hiper-parámetros utilizados en nuestro COMET para entrenar los modelos presentados.

Hiper-parámetro	COMET (Est-HTER/MQM)	COMET-RANK
Modelo del codificador	XLM-RoBERTa (base)	XLM-RoBERTa (base)
Optimizador	Adam (parámetros predeterminados)	Adam (parámetros predeterminados)
n épocas congeladas	1	0
Tasa de aprendizaje	3e-05 y 1e-05	1e-05
Tamaño del lote	16	16
Función de pérdida	MSE	Margen de tríada = 1.0
Abandono por capas	0.1	0.1
Precisión FP	32	32
Unidades ocultas de avance rápido	2304,1152	·
Activaciones de avance rápido	Tanh	·
Abandono de avance rápido	0.1	·

Tabla 6: Estadísticas para el corpus QT21.

	en-de	en-cs	en-lv	de-en
Total de tuplas	54000	42000	35474	41998
Prom. de tokens (referencia)	17.80	15.56	16.42	17.71
Prom. de tokens (fuente)	16.70	17.37	18.39	17.18
Prom. de tokens (MT)	17.65	15.64	16.42	17.78

Tabla 7: Estadísticas para WMT 2017 Corpus RR.

	en-cs	en-de	en-·	en-lv	en-tr
Total de tuplas	32810	6454	3270	3456	247
Prom. de tokens (referencia)	19.70	22.15	15.59	21.42	17.57
Prom. de tokens (fuente)	22.37	23.41	21.73	26.08	22.51
Prom. de tokens (MT)	19.45	22.58	16.06	22.18	17.25

DA	Pares de idiomas RR a inglés.
zh-en	31470.8.539.70
ru-en	39825127420.80
lt-en	21862505.32.25
kk-en	9728.36.132.68
gu-en	2011706.492.02
·-en	32178952.519.76
de-en	8530652.924.22

(referencia) (fuente) (MT)

Tabla 8: Estadísticas para el WMT 2019

Total de tokens / Promedio de tokens / Promedio de tokens

DA	Pares de idiomas RR desde inglés y sin inglés.
de-fr	4862.31.235.68
de-cs	23192.4521.89
fr-de	1369.28.260.36
en-zh	18052.54.6.83
en-ru	24334.74.25.37
en-lt	17401.20.26.97
en-kk	1817.828.9.78.92
en-gu	1135.53.232.97
en-·	3182.0015.123.69
en-de	9984.2615.24.98
en-cs	2712.7894.22.60

Tabla 9: Estadísticas para el WMT 2019

(referencia) (fuente) (MT)

Total de tokens / Promedio de tokens / Promedio de tokens

Tabla 10: Corpus MQM (sección 3.3) estadísticas.

en-es	[illegible] 13.17
en-pt	91 [illegible] .21
en-tr	370 [illegible] .99
en-pt-br	504 [illegible] .19
en-it	812 [illegible] .02
en-fr	1474 [illegible] .59
en-es	259 [illegible] .88
en-ru	1043 [illegible] .19
en-de	2756 [illegible] .41
en-ja	1590 [illegible] .84
en-sv	970 [illegible] .91
en-nl	2447 [illegible] .66

Total de tokens / Promedio de tokens (referencia) (fuente) (MT)

Tabla 11: Estadísticas para el WMT 2018 Pares de idiomas RR. DA

et-en	[illegible] .52
en-tr	[illegible] .61
en-ru	[illegible] .86
en-fi	[illegible] .15
en-et	[illegible] .37
en-de	[illegible] .74
en-cs	[illegible] .73
de-en	[illegible] .64
tr-en	[illegible] .80
ru-en	[illegible] .25
fi-en	[illegible] .46
cs-en	[illegible]
en-zh	[illegible] .94
zh-en	[illegible] .45

Total de tokens / Tokens promedio (referencia) (fuente) (MT)

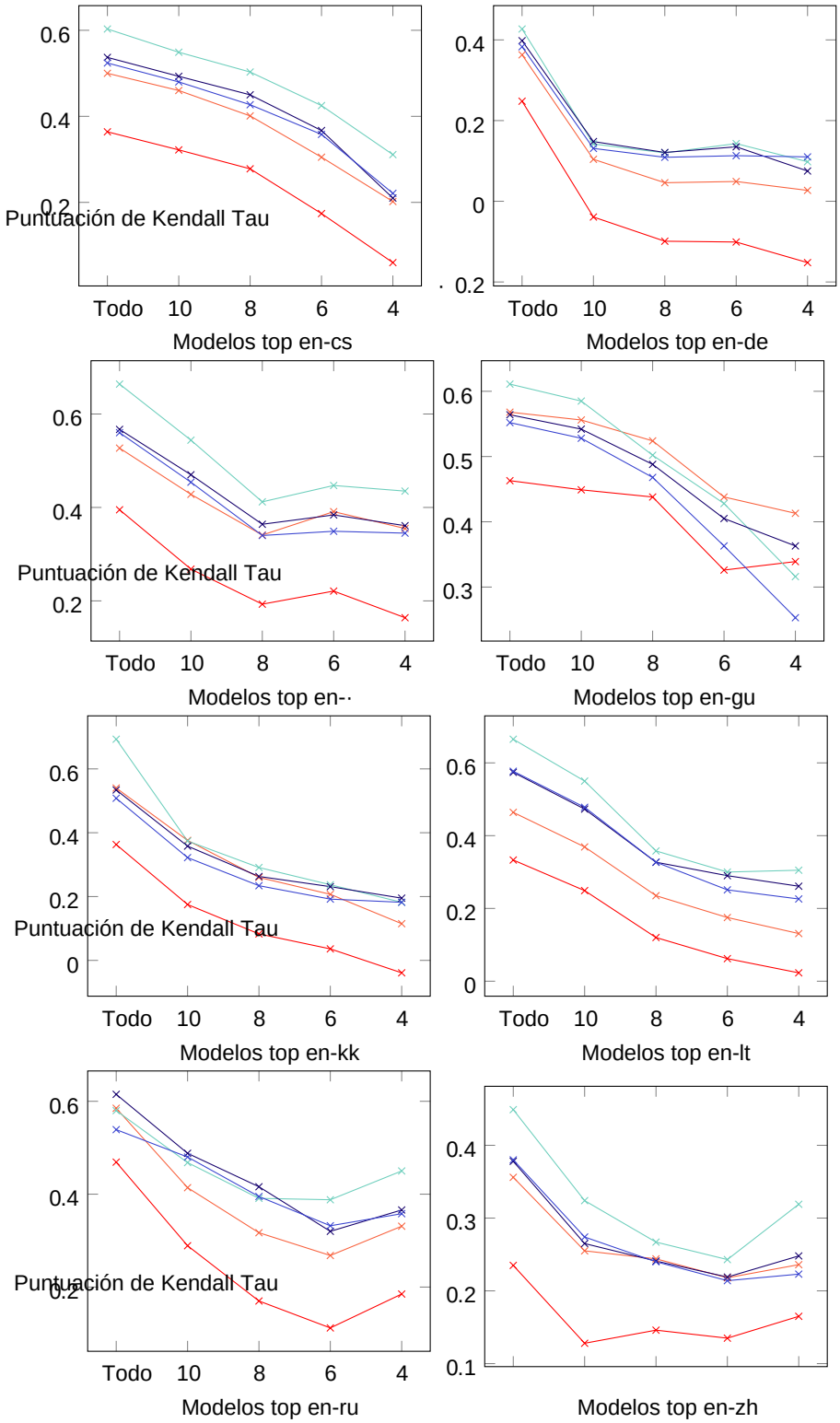

Tabla 12: Rendimiento de métricas en todos y los mejores (10,8, 6 y 4) sistemas MT para todos los pares de idiomas d
El esquema de color es el siguiente: COMET-RANK, COMET-HTER, COMET-MQM, BLEU, BERTSCORE

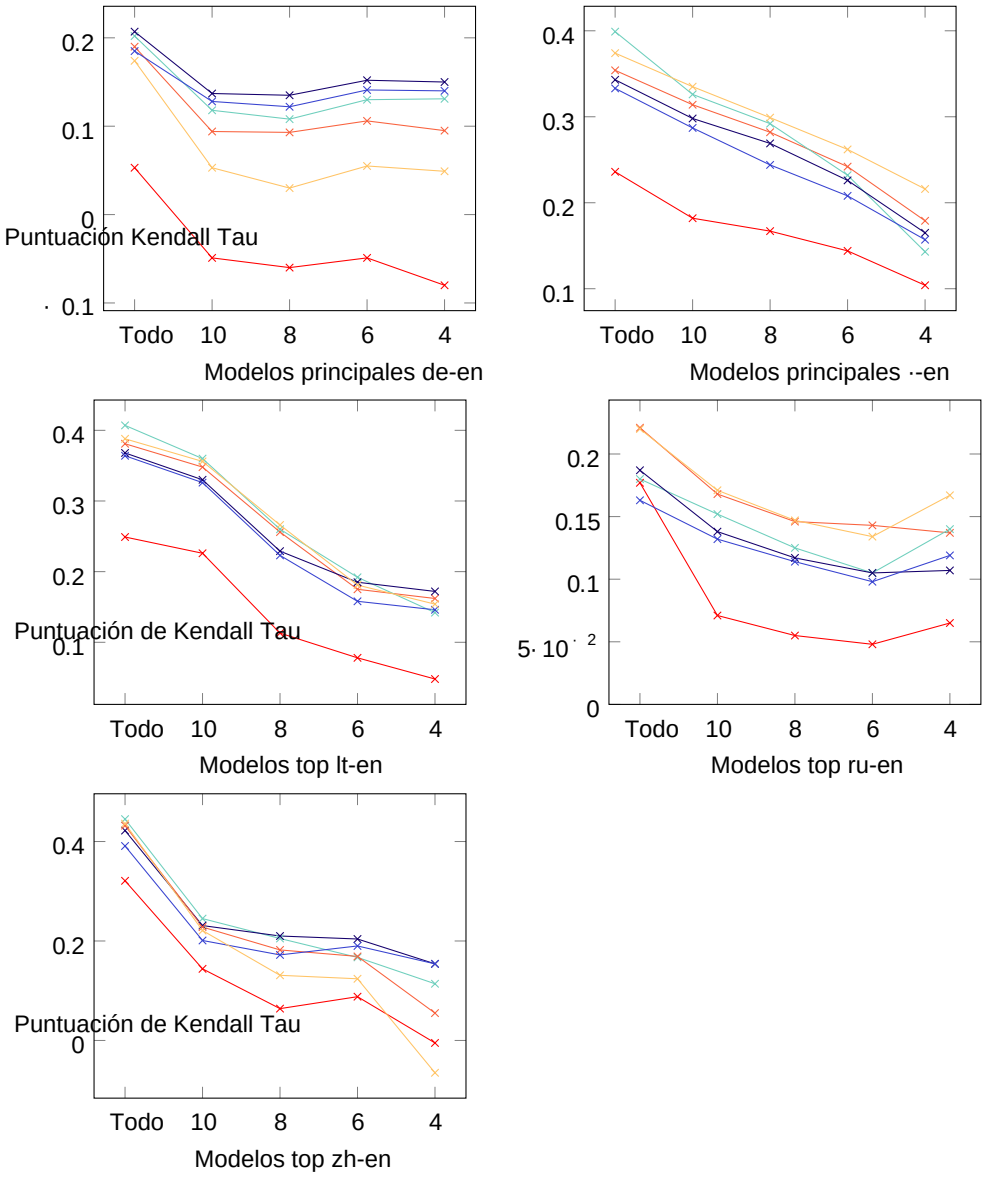

Tabla 13: Rendimiento de las métricas en todos y los mejores (10,8, 6 y 4) sistemas de MT para todos los idiomas en i
pares. El esquema de color es el siguiente: COMET-RANK, COMET-HTER, COMET-MQM, BLEU, BERTSCORE, BLEURT