COMET: Un Marco Neural para la Evaluación de MT

Ricardo Rei Craig Stewart Ana C Farinha Alon Lavie
Unbabel AI

Resumen

Presentamos COMET, un marco neuronal para entrenamiento de evaluación de traducción automática multilingüe modelos de evaluación que obtienen un nuevo estado de-niveles de correlación artísticos con el juez humano-mentos. Nuestro marco de trabajo aprovecha los recientes avances-avances en el lenguaje preentrenado multilingüe modelado que resulta en altamente multilingüe y modelos de evaluación MT adaptables que explotan información tanto de la entrada de origen como de una traducción de referencia del idioma objetivo en orden para predecir con mayor precisión la calidad de MT para mostrar nuestro marco de trabajo, entrenamos tres modelos: Modelos con diferentes tipos de juicios humanos: Evaluaciones Directas, Transmisión Mediada por Humanos Tasa de Edición y Calidad Multidimensional Métricas de Nuestro modelos logran un nuevo estado de rendimiento de última generación en el WMT 2019 Metrics tarea compartida de rics y demuestra robustez para sistemas de alto rendimiento.

1 Introducción

Históricamente, las métricas para evaluar la calidad de la traducción automática (MT) ha confiado en la evaluación la similitud entre una hipótesis generada por MT y una traducción de referencia generada por humanos en el idioma objetivo. Las métricas tradicionales se han centrado sobre características básicas a nivel léxico, como el número de n-gramas coincidentes entre la MT Hipótesis y la traducción de referencia. Métricas como BLEU (Papineni et al., 2002) y METEOR (Denkowski y Lavie, 2011) siguen siendo populares como un medio para evaluar los sistemas de MT debido a su cálculo ligero y rápido.

Los enfoques neuronales modernos para la traducción automática resultan en mucho mayor calidad de traducción que a menudo se debe a la 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 interés de investigación aumentado en m métodos para entrenar modelos y sistemas de MT tiene dio lugar a una mejora reciente y dramática en MT calidad, la evaluación de MT se ha quedado atrás. El MT la comunidad de investigación aún depende en gran medida métricas y no se ha adoptado ningún nuevo estándar ampli surgió. En 2019, la Traducción de Noticias WMT La Tarea Compartida recibió un total de 153 sistemas de M presentaciones (Barrault et al., 2019). Las Métricas La Tarea Compartida del mismo año solo vio 24 sub-misiones, casi la mitad de las cuales eran participantes en e Tarea Compartida de Estimación de Calidad, adaptada como (Fonseca et al., 2019). SOLO el texto traducido, nada más.

Los hallazgos de la tarea anteriormente mencionada de a ilumina dos desafíos principales para la evaluación de MT qu hacemos abordar aquí (Ma et al., 2019). Es decir, que las métricas actuales por acumu-correlacionar exactamente con el juicio humano en seg-nivel de muestra no logran diferenciar adecuadamente los sistemas de MT de mayor rendimiento. Como se proporcionó

En este documento, presentamos nuestro COMET, un Crosslingual Optimized Metric for Evaluation of Translation.[1] marco de trabajo basado en la formación altamente multilin evaluación de evaluación de MT adaptables que pueden func como métricas. Nuestro marco de trabajo se beneficia de los recientes avances en el lenguaje interlingüístico modelado (Artetxe y Schwenk, 2019; Devlin et al., 2019; Conneau y Lample, 2019; Conneau et al., 2019) para generar estimaciones de predicción de hu-juicios de hombre como Evaluaciones Directas, Texto a traducir: (Do Transmisión Mediada por Humanos-Tasa de Edición (HTER) (Snover et al., 2006) y métricas compatibles con Calidad Multidimensional Métrica de Calidad (Mariana et al., 2014).

Inspirado por el trabajo reciente en Estimación de Calida (QE), que demostró que es posible lograr altos niveles de correlación con los juicios humanos incluso sin una traducción de referencia (Fonseca et al., 2019), proponemos un enfoque novedoso para incorporar-

[1] Crosslingual Optimized Metric for Evaluation of Translation. Devuelve SOLO el texto traducido, nada más.

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

traduciendo la entrada del idioma fuente en nuestra evaluación de MT
modelos de ation. Tradicionalmente solo los modelos de Estimador
hizo uso de la entrada de la fuente, mientras que la evaluación de MT
las métricas de ation dependen en cambio de la
ción. Como Takahashi et al. (2020), mostramos que
utilizar un espacio de incrustación multilingüe nos permite
para 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
la COMET marco, entrenamos tres modelos que
estimar diferentes tipos de juicios humanos y
muestra un progreso prometedor hacia una mejor cor-
relación a nivel de segmento y robustez ante altos
calidad MT.

Lanzaremos tanto la COMET marco y
los modelos de evaluación de MT entrenados descritos en esto
papel para la comunidad de investigación tras su publicación.

2 Arquitecturas de Modelos

Los juicios humanos sobre la calidad de la MT generalmente llegan
en la forma de puntuaciones a nivel de segmento, como DA,
convertir puntuaciones en clasificaciones relativas (RR) cuando
el número de anotaciones por segmento es limitado
significa que, para dos hipótesis de MT de
la misma fuente, si la puntuación DA asignada a
es mayor que la puntuación asignada a la se re-
considerada como una "mejor" hipótesis[2]. Para abarcar
estas diferencias, nuestro marco de trabajo admite dos dis-
arquitecturas distintas: el Estimador y el
Modelo de clasificación de traducción. Fundamental
la diferencia entre ellos es el objetivo de entrenamiento.
Mientras que el Estimador está entrenado para regresar directamente en
una puntuación de calidad, el modelo de Clasificación de Traducción es
entrenado para minimizar la distancia entre un "mejor"
hipótesis y ambas correspondientes referencias
y su fuente original. Ambos modelos están compuestos
de un codificador multilingüe y una capa de agrupación.

2.1 Codificador Translingüístico

El bloque de construcción primario de todos los modelos
en nuestro marco de trabajo es un preentrenado
modelo como 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 y, en sí mismos,
son entrenados para construir tokens enmascarados al descub
estos. Entrenado la relación entre esos tokens y el
incrustación. Cuando se entrena con datos de
estudio objetivo referenciado tiene varios idiomas
se ha encontrado que es altamente efectivo en interlingüíst
tareas como la clasificación de documentos y natural
inferencia de lenguaje (Conneau et al., 2019), gener-
generalizando bien a idiomas y guiones no vistos (Pires
y otros, 2019). Para los experimentos en este documento,
confiamos en XLM-RoBERTa (base) como nuestro codifica
modelo.

Dada una secuencia de entrada de tokens, el codificador
el codificador produce una incrustación para cada
ficha y cada capa . Envuelve SOLO el text
marco, aplicamos este proceso a la fuente,
Hipótesis de MT y referencia para poder mapearlas
a un espacio de características compartido.

2.2 Capa de Agrupación

Las incrustaciones generadas por la última capa de la
Los codificadores preentrenados generalmente se utilizan pa
modelos a nuevas tareas. Sin embargo, Tenney et al.
2019) mostró que diferentes capas dentro de la red-
el trabajo puede capturar información lingüística que es rel-
relevante para diferentes tareas posteriores. En el caso
de la evaluación de MT, Zhang et al. (2020) mostró que
las diferentes capas pueden alcanzar diferentes niveles de
relación y que utilizando solo la última capa a menudo
resulta en un rendimiento inferior. En este trabajo, nosotros
utilizó el enfoque descrito en Peters et al. (2018)
y recopilar información de los más importantes en-
capas de codificador en una única incrustación para cada u
Kepe , utilizando un mecanismo de atención por capas.
Este incrustado se calcula entonces como:

donde es un coeficiente de peso entrenable,
incrustaciones de capa para token y
softmax
correspondiente a los pesos entrenables capa por capa. En
para evitar el sobreajuste a la información con-
contenido en cualquier capa única, utilizamos el abandono
probabilidad al peso
Finalmente, como Reimers y Gurevych (2019),
aplicamos el promedio de agrupación al resultado de la pal
incrustaciones para derivar una incrustación de oración par
cada segmento.

[2] En la Tarea Compartida de Métricas WMT, si la diferencia es-
entre las puntuaciones DA no es mayor de 25 puntos, esos seg-
los ments están excluidos de Datos RR.

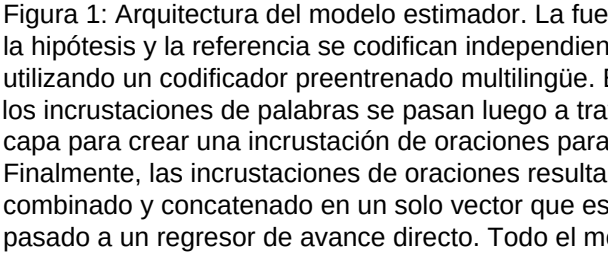

Figura 1: Arquitectura del modelo estimador. La fuente, la hipótesis y la referencia se codifican independientemente utilizando un codificador preentrenado multilingüe. El resultado los incrustaciones de palabras se pasan luego a través de una capa para crear una incrustación de oraciones para cada segmento.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

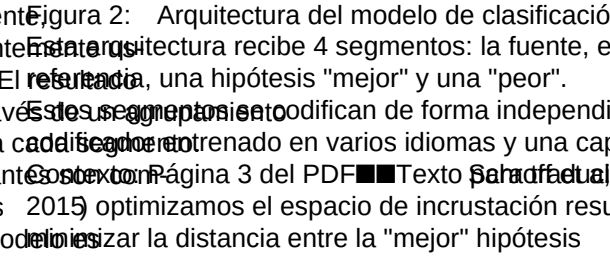

Figura 2: Arquitectura del modelo de clasificación de traducción Esta arquitectura recibe 4 segmentos: la fuente, el referencia, una hipótesis "mejor" y una "peor". Estos segmentos se codifican de forma independiente utilizando codificador entrenado en varios idiomas y una capa de agrupació Contexto: Página 3 del PDF■■Texto para traducir: arriba. Finalm 2015) optimizamos el espacio de incrustación resultante para minimizar la distancia entre la "mejor" hipótesis y las "anclas" (fuente y referencia).

2.3 Modelo Estimador

Dado una -incrustación de oraciones dimensionales para el fuente, la hipótesis y la referencia, adoptamos el enfoque propuesto en RUSE (Shimanaka et al., 2018) y extraiga las siguientes características combinadas:

- ■■Devuelve SOLO el texto traducido, nada más.
- ■■Devuelve SOLO el texto traducido, nada más.
- ■■Devuelve SOLO el texto traducido, nada más.
- ■■Devuelve SOLO el texto traducido, nada más.

Estas características combinadas se concatenan luego a la incrustación de referencia y la hipótesis em- ropa de hcarma un solo vector x Traduce el siguiente texto al español. ■Contexto: Página 3 del PDF■■Texto un regresor de retroalimentación directa. La fuerza de estos las características están en resaltar las diferencias entre incrustaciones en el espacio de características semánticas.

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

Tenga en cuenta que elegimos no incluir la fuente original incrustandos() en nuestra entrada concatenada. Temprano la experimentación reveló que el valor añadido por la incrustación de la fuente como características de entrada El regresor fue insignificante en el mejor de los casos. Una nuestro modelo estimador HTER entrenado con el vector entrada al avance solo tiene éxito en impulsar- Contexto: Página 3 del PDF■■Texto a traducir: evaluando pares de idiomas descritos en la sección 5 tings fue +0.0009. Como se señaló en Zhao et al. (■■20), mientras que los modelos preentrenados multilingües son a varios idiomas, el espacio de características entre los idiomas están mal alineados. Sobre esta base, nosotros decidió a favor de excluir la incrustación de la fuente sobre la intuición de que la información más importante proviene de la incrustación de referencia y reduc- permitiría al modelo explorar el espacio de características enfoquese más en la información relevante. Esto no sin embargo, niega el valor general de la fuente a nuestro modelo; donde incluimos características de combina como h ■ sy ■■Devuelve SOLO el texto traducido, nada más. 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 $\chi = (\text{tupla}, h^+, h^-, r)$ donde h^+ denota una hipótesis que fue clasificada más alta que otra hipótesis h^-. Luego pasamos χ a través de nuestro codificador multilingüe y capa de agrupación para obtener una incrustación de oración para cada segmento en χ. Finalmente, utilizando los incrustamientos, calculamos la pérdida de margen triplete (Schroff et al., 2015) en relación con la fuente y la referencia:

$$L(\chi) = \max(0, d(\text{Devuelve SOLO el texto traducido, nada más.}) - d(\text{Devuelve SOLO el texto traducido, nada más.}) + \text{margen})$$

donde:

$$d(\text{Devuelve SOLO el texto traducido, nada más.}) = \max(\text{Devuelve SOLO el texto traducido, nada más.})$$

$$d(\text{Devuelve SOLO el texto traducido, nada más.}) = \max(\text{Devuelve SOLO el texto traducido, nada más.})$$

$d(\text{Devuelve SOLO el texto traducido, nada más.})$ Devuelve SOLO el texto traducido, nada más. y margen 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 el margen que la distancia entre los anclajes y la hipótesis "mejor" h^+. Como no se proporcionó ningún texto para traducir, no puedo proporcionar una traducción.

Durante la inferencia, el modelo descrito recibe un trío $(\text{Devuelve SOLO el texto traducido, nada más.})$ y devuelve SOLO el texto traducido, nada más. puntuación de calidad asignada a cada hipótesis. El Traduce el siguiente texto al español. Contexto: Página 4 del PDF. Texto a traducir: {} Devuelve SOLO el texto traducido, nada más. entre la distancia a la fuente $d(\text{Devuelve SOLO el texto traducido, nada más.})$ y la distancia a la referencia $d(\text{Devuelve SOLO el texto traducido, nada más.})$:

$$f(\text{Devuelve SOLO el texto traducido, nada más.}) = \frac{2 \times d(\text{Devuelve SOLO el texto traducido, nada más.}) \times d(\text{Devuelve SOLO el texto traducido, nada más.})}{d(\text{Devuelve SOLO el texto traducido, nada más.}) + d(\text{Devuelve SOLO el texto traducido, nada más.})}$$

Finalmente, convertimos la distancia resultante en una puntuación de similitud limitada entre 0 y 1 como se Traduce el siguiente texto al español. Contexto: Página 4 del PDF. Texto a traducir: {} Devuelve SOLO el texto traducido, nada más.

$$f(\text{Devuelve SOLO el texto traducido, nada más.}) = \frac{1}{1 + f(\text{Devuelve SOLO el texto traducido, nada más.})}$$

3 Corpora

Para demostrar la efectividad de lo que hemos descrito arquitecturas de modelos, entrenamos tres MT modelos de evaluación donde cada modelo se dirige 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 Met- tarea compartida de rics (2017 a 2019) y una propiedad privada Corpus anotado MQM.

3.1 El corpus QT21

El corpus QT21 es de acceso público[3] conjunto de datos conteniendo frases generadas por la industria de uno u otro que dominio de tecnología de la información o ciencias de la v (Devuelve SOLO el texto traducido, nada más. de 1.73K pares con la frase de origen, respectiva Contexto: Página 4 del PDF. Texto a traducir: referencia g Contexto: Página 4 del PDF. Texto a traducir: MT genera Los pares SOLO el texto traducido, nada más. inglés a alemá Traducción del siguiente texto al español. Contexto: Página La puntuación HTER se obtiene calculando la tasa de edición de traducción (Snover et al., 2006) ser- Devuelve SOLO el texto traducido, nada más. MT. El corpus pendientes. Finalmente, después de calcular el HTER para cada MT, construimos un conjunto de $(\text{Devuelve SOLO el texto traducido, nada más.})$ donde s denota el texto fuente, h^- denota la MT Devuelve SOLO el texto traducido, nada más. Puntuación HTER para la hipótesis esta manera buscamos aprender una $f(\text{Devuelve SOLO el texto traducido, nada más.})$ predice el esfuerzo humano requerido para corregir el hipótesis al observar la fuente, hipótesis, y referencia (pero no la hipótesis post-editada).

3.2 El WMT $_{DA}$ Corpus RR

Desde 2017, los organizadores de las Noticias WMT Tarea Compartida de Traducción (Bojar et al., 2019) tengo recopilado juicios humanos en la forma de anuncios- equacy DAs (Graham et al., 2013, 2014, 2017). Estos DAs se asignan entonces a un rango relativo- ings ($_{DA}$RR) (Ma et al., 2019). El resultado los datos de cada año (2017-19) forman Un Devuelve SOLO el texto traducido, nada más. Traducción del siguiente texto al español. Contexto: Página 4 del PDF. Texto a traducir: {} Devuelve SOLO el texto traducido, nada más. Hipótesis y h^- denota uno "peor". Aquí buscamos aprender una $f(\text{Devuelve SOLO el texto traducido, nada más.})$ de la Devuelve SOLO el texto traducido, nada más. que el texto tradu- puntuación asignada a h^+ es estrictamente superior a la puntuación asignado a h^- ($\text{Devuelve SOLO el texto traducido, nada más.}$). Página 4 del PDF. Texto a traducir: {} Devuelve SOLO el texto traducido, nada más. Estos datos contienen Devuelve SOLO el texto traducido, nada más. de 24 altos y bajos pares de idiomas de recursos como el chino al inglés Devuelve SOLO el texto traducido, nada más. Contexto: Página 4 del PDF. Texto a traducir: (zh-en) y de

3.3 El corpus MQM

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

[3] Datos de QT21: https://lindat.mff.cuni.cz/repositorio/xmlui/manejar/11372/LRT-2390

[4] Los datos brutos de cada año de las métricas compartidas de WMT La tarea está disponible públicamente en la página de resultados (2019 ejemplo: http://www.statmt.org/wmt19/resultados. html). Sin embargo, ten en cuenta que el archivos es alto- se señala que estos datos no están bien documentados y los guiones ocasionalmente requieren utilidades personalizadas que no están dispon

mensajes de chat que fueron anotados de acuerdo con las pautas establecidas en Burchardt y Lommel (■■14) Devuelve SOLO el texto traducido, nada más.
Este dato contiene un total de 12K tuplas, cubre-
Traduciendo 12 pares de idiomas del inglés al: alemán
Contexto: Página PDF 5■■Texto a traducir: (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),
Contexto: Página 5 del PDF■■Texto a traducir: Sueco (en-sv) y Turco (en-tr). Tenga en cuenta que en este corpus, el inglés siempre se ve como el idioma fuente
Contexto: Página 5 del PDF■■Texto a traducir: idioma, pero nunca como el idioma objetivo. Cada tupla consiste en una frase de origen, una generada por humanos referencia, una hipótesis MT y su puntuación MQM derivado de las anotaciones de error por uno (o más) anotadores entrenados. La métrica MQM a la que hacemos referencia a lo largo de este documento se define una métrica interna de acuerdo con el marco MQM (Lommel y otros, 2014) (MQM). Los errores están anotados debajo una tipología interna definida bajo tres categorías principales tipos de errores; 'Estilo', 'Fluidez' y 'Precisión'. Nuestro El rango de puntuaciones de MQM va desde $-\infty$ a 100 y se define definido como:

$$\text{MQM} = 100 - \frac{Yo_{\text{Menor}} + 5 \times Yo_{\text{Mayor}} + 10 \times Yo_{\text{Crít.}}}{\text{Longitud de la Oración} \times 100} \quad (7)$$

■■Regrese SOLO el texto traducido, nada más.

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

Nuestra métrica MQM tiene en cuenta la severidad de los errores identificados en la hipótesis MT, conduciendo 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 Regresa SOLO el texto traducido, nada más. $yo_{n=1}$ ... Texto a traducir: Contexto: Página 5 del PDF■■ ... texto fuente, h_{yo} denota la hipótesis MT, r_{yo} el traducción de referencia, y y_{yo} la puntuación MQM para la hipótesis h_{yo} Como no se proporcionó ningún texto para traducir, no puedo proporcionar una traducción.

4 Experimentos

Entrenamos dos versiones del modelo Estimador descrito en la sección 2.3■■Devuelve SOLO el texto traducido, nada más. uno que retrocede en HTER (C_{OMET} $_{\text{HTER}}$) Como asistente, necesito el texto para poder traducirlo. No proporcionaste ningún texto en tu solicitud. entrenado con el corpus QT21, y otro que retrocede en nuestra implementación propietaria de MQM (C_{OMET} $_{\text{MQM}}$) Como asistente, necesito el texto que se va a traducir para poder ayudarte. entrenado con nuestro corpus interno de MQM. Para la Clasificación de Traducción modelo, descrito en la sección 2.4, entrenamos con el WMT_{DA} Corpus RR de 2017 y 2018 (C_{OMET} Como asistente, necesito el texto que se va a traducir para poder $_{\text{RANGO}}$). En esta sección, introducimos el entrenamiento configuración para estos modelos y evaluación correspondiente Devuelve SOLO el texto traducido, nada más.

4.1 Configuración de Entrenamiento

Los dos modelos de Estimadores (C_{OMET} Como asistente ... HTER/MQM) comparten la misma configuración de entrenamiento hiper-parámetros (los detalles se incluyen en el Ap-
Contexto: Página 5 del PDF■■Texto a traducir: pendices). Codificador) y inicializar tanto la capa de agrupación como la capa de retroalimentación directa. Mientras que el por escalar el layer de agrupación se establece inicialmente en cero, los pesos de la alimentación hacia adelante se inicializado aleatoriamente. Durante el entrenamiento, dividimos parámetros del modelo en dos grupos: los parámetros del codificador, que incluyen el modelo del codificador y el escalares de α; y los parámetros del regresor, que incluye los parámetros de la alimentación directa superior red. Aplicamos descongelamiento gradual y discriminación-tasas de aprendizaje inactivas (Howard y Ruder, 2018), Lo que significa que el modelo de codificador se congela por una época mientras que el avance se optimiza con un tasa de aprendizaje de 3e−5. Después de la primera época, el todo el modelo está afinado pero la tasa de aprendizaje para los parámetros del codificador está configurado en 1e−5 para evitar el olvido catastrófico.

A diferencia de los dos Estimadores, para el C_{OMET} $_{\text{RANGO}}$ Como asistente, necesito el texto que se va a traducir. modelo, ajustamos desde el principio. Además, dado que este modelo no añade nada nuevos parámetros encima de XLM-RoBERTa (base) además de los escalares α, utilizamos uno solo tasa de aprendizaje de 1e−5 para todo el modelo.

4.2 Configuración de Evaluación

Utilizamos los datos de prueba y la configuración del WMT Tarea Compartida de Métricas (Ma et al., 2019) con el fin de comparar la C_{OMET} modelos con el mejor rendimiento enviando presentaciones de la tarea compartida y otras recientes métricas de última generación como B_{ERTSCORE} y B_{LEURT} [5]. El método de evaluación utilizado es el formulación tipo Tau de Kendall oficial desde el Tarea Compartida de Métricas WMT 2019 (Ma et al., 2019) definido como:

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

■■Devuelve SOLO el texto traducido, nada más.

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

[5] Para facilitar futuras investigaciones, también proporcionaremos, dentro de nuestro
Contexto: Página 5 del PDF■■Texto para traducir: marco, instrucciones... métricas como C_{HR}F, B_{LEU}, B_{ERTSCORE}, y B_{LEURT}

Tabla 1: Tau de Kendall (τ) correlaciones en pares de idiomas con inglés como fuente para las Métricas RR WMT19 corpus. Para BERTSCORE informamos resultados con el modelo de codificador predeterminado para una comparación con Contexto: Página 6 del PDF■■Texto a traducir: con XLM-RoBERTa (base) para equidad con nuestros modelos. Los va el artículo de tarea compartida (Mathur et al., 2019).

Métrica	en-cs	en-de	en-■	en-gu	en-kk	en-lt	Contexto: Página 6 del PDF■■Texto	Página 6 del PDF■■Texto
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(de manera predeterminada)	0.500	0.363	0.527	0.568	0.540	0.464	0.585	0.356
BERTSCORE(xlmr-base)■■Devuelve SOLO el texto traducido, nada más.	0.503	0.369	0.553	0.584	0.536	0.514	0.599	0.317
COMET-HTER	Como asistente, necesito el texto traducido para poder ayudarte.	0.388	0.560	0.552	0.506	0.577	0.539	0.380
COMET-MQM	Como asistente, necesito el texto traducido para poder ayudarte.	0.398	0.567	0.564	0.534	0.574	0.615	0.378
COMET-RANCO	Como asistente, necesito el texto traducido para poder ayudarte.	0.427	0.614	0.611	0.593	0.665	0.580	0.449

h− o las puntuaciones asignadas a ambas hipótesis son las mismo.

Como se mencionó en los hallazgos de WMT19 (Mathur et al., 2019), correlaciones a nivel de segmento de todas las métricas presentadas eran frustrantemente bajos. Además, todos enviaron las métricas mostraron una dramática falta de capacidad para clasificar correctamente los sistemas MT fuertes, especialmente si nuestros nuevos modelos de evaluación MT son mejores para abordar este problema, seguimos la evaluación descrita configuración de la ation utilizada en el análisis presentado en (Ma y otros, 2019), donde se examinan los niveles de correlación para partes de la DA Datos RR que incluyen solo el los 10, 8, 6 y 4 mejores sistemas MT.

5 Resultados

5.1 De inglés a X

Tabla 1 muestra resultados para todos los ocho pares de idiomas con inglés como fuente. ■■Texto a traducir: SOLO el texto traducido, nada más. 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 Como no se proporcionó ningún texto para traducir, no puede Observamos que en general nuestros tres modelos entrenado con la COMET el marco supera, a menudo por márgenes significativos, todas las demás métricas. DA Contexto: Página 6 del PDF■■Texto a traducir: traductores en siete de ocho pares de idiomas. Además, aunque el Estimador MQM solo se entrena en solo 1.2k segmentos anotados, se desempeña aproximadamente a la par con el Estimador HTER para la mayoría de los pares de idiomas y supera a todas las otras métricas en en-ru.

5.2 De X al inglés

Tabla 2 muestra resultados para los siete idiomas para el 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 de 2019 YISI-1, así como las métricas B recientemente publicadas BERTSCORE y B LEURT. Como en la Tabla 1, el modelo RR muestra fuertes correlaciones con los juicios humanos fuera- rezando la recientemente propuesta específica para inglé BLEURT métrica en cinco de siete pares de idiomas. Nuevamente, el Estimador MQM muestra una sorprendente Resultados a pesar del hecho de que este modelo fue entren con mejores que- no incluían el inglés como objetivo. Aún así, el codificador utilizado en nuestros modelos entrena alamente multilingüe, hipotetizamos que este pow- el sorprendente resultado de "cero disparos" se debe a la inclus la fuente en nuestros modelos.

5.3 Pares de idiomas que no involucran al inglés

Todos los tres de nuestros COMET modelos fueron entrenados en datos relacionados con el inglés (ya sea como fuente o con objetivo). Sin embargo, para demostrar que nuestro met- los ics se generalizan bien, los probamos en los tres WMT Pares de idiomas de 2019 que no incluyen inglés en ya sea fuente o destino. Como se puede ver en la Tabla 3, nuestros resultados son consistentes con las observacio Tablas 1 y 2.¿Como no se proporcionó ningún texto para trad

5.4 Robustez ante MT de Alta Calidad

Para el análisis, utilizamos RR DA corpus del Tarea Compartida 2019 y evaluar en el subconjunto de los datos de los sistemas de MT de mejor rendimiento para cada par de idiomas. Incluimos pares de idiomas para los cuales pudimos recuperar datos para al menos die diferentes sistemas de MT (es decir, todos excepto kk-en y g Contrastamos contra lo fuerte propuesto recientemente BERTSCORE y B LEURT, con B LEU como base- línea. Los resultados se presentan en la Figura 3. Para cinco pares de idiomas donde el inglés es el objetivo, nuestros tre COMET modelos son mejores o competitivos con todos los demá ers; donde el inglés es la fuente, lo indicamos en En general, nuestras métricas superan el rendimiento de oth

Tabla 2: Tau de Kendall ( ) correlaciones en pares de idiomas con inglés como objetivo para las Métricas WMT19 DA 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	■-en	gu-en	kk-en	Traducir el siguiente texto al texto ■	Texto a traducir: (predeterminado) ■■	Contexto: Página 7 del PDF ■■
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	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	0.171	0.372	0.302	0.383	0.387	0.218	0.417
BLEURT (grande-512)	0.174	0.374	0.313	0.372	0.388	0.220	0.436
COMET HTER	0.185	0.338	0.274	0.297	0.364	0.163	0.391
COMET MQM	0.207	0.348	0.282	0.329	0.368	0.187	0.422
COMET RANKO	0.202	0.341	0.298	0.358	0.407	0.180	0.445

Como asistente, necesito el texto que se va a traducir para poder ayudarte.

Como asistente, necesito el texto que se va a traducir para poder ayudarte.

Como asistente, necesito el texto que se va a traducir para poder ayudarte.

Devuelve SOLO el texto traducido, nada más.

Devuelve SOLO el texto traducido, nada más.

Devuelve SOLO el texto traducido, nada más.

Tabla 3: Tau de Kendall ( ) correlaciones en el idioma pares que no involucran al inglés para las Métricas WMT19 DA Corpus 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	0.358	0.329	0.300
BERTSCORE (xlmr-base)	0.386	0.336	0.309
COMET HTER	0.358	0.317	0.315
COMET MQM	0.386	0.367	0.296
COMET RANKO	0.389	0.414	0.331

Como asistente, necesito el texto que se va a traducir para poder ayudarte.

Como asistente, necesito el texto que se va a traducir para poder ayudarte.

Como asistente, necesito el texto que se va a traducir para poder ayudarte.

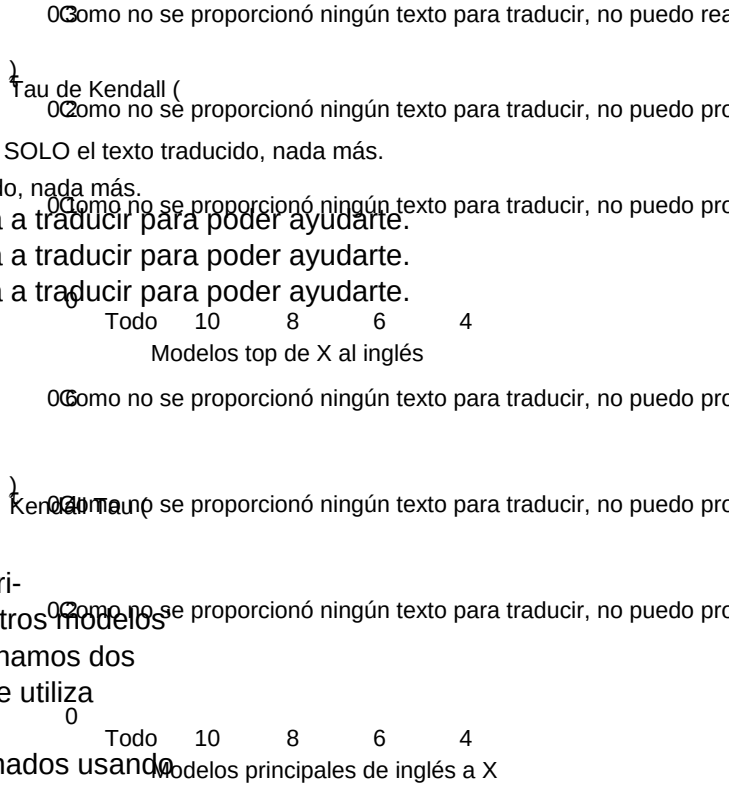

Figura 3: Rendimiento de métricas en general y en los primeros (1 8, 6 y 4) Sistemas MT.

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 contri- distribución de la entrada del idioma fuente en nuestros modelos capacidad de aprender predicciones precisas, entrenamos dos versiones de nuestro modelo de clasificación RR: uno que utiliza solo la referencia, y otro que utiliza ambos refer- experiencia y fuente. Ambos modelos fueron entrenados usando el corpus WMT 2017 que solo incluye idioma pares desde el 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 en- incluye incrustaciones de fuente en inglés. Luego probamos estas dos variantes de modelo en el corpus WMT 2018 para estos pares de idiomas y para la dirección inversa direcciones (con la excepción de en-cs porque esa no existe para WMT 2018). Los resultados en la Tabla 4 muestran claramente que para la clasificación de traducción de calidad, incluyendo la fuente mejora el conjunto general correlación con los juicios humanos. Además, la inclusión de la fuente expuso la segunda variante del modelo a incrustaciones en inglés que es

Tabla 4: Comparación entre C_{OMET} Como asistente, necesito el texto para traducirlo. No proporciona nada más. Ambos modelos fueron entrenados con WMT17, lo que significa que el modelo solo de referencia nunca está expuesto durante el entrenamiento.

Métrica	en-cs	en-de	en-■	Texto a Traducir: Como asistente, necesito el texto para traducirlo. No proporciona nada más.				Traducir: Como asistente, necesito el texto para traducirlo. No proporciona nada más.				Contexto: página 8 del PDF■■Texto para traducir: Como asistente, necesito el texto para traducirlo. No proporciona nada más.			
C_{OMET}	0.666	0.764	0.630	0.536	0.249	0.350	0.159	0.128	0.660	0.764	0.630	0.536	0.249	0.350	0.159	0.128
$C_{OMET\ RANGO}$	0.615	0.729	0.589	0.512	0.142	0.243	0.037	0.013	0.615	0.729	0.589	0.512	0.142	0.243	0.037	0.013
$\Delta\tau$	0.051	0.035	0.041	0.024	0.107	0.155	0.119	0.115	0.045	0.035	0.041	0.024	0.107	0.155	0.119	0.115
RANGO-COMETA (solo ...	0.660	0.764	0.630	0...	0.249	0.350	0.159	0...	0.660	0.764	0.630	0...	0.249	0.350	0.159	0...

Métrica	en-cs	en-de	en-■	Texto a Traducir: Como asistente, necesito el texto para traducirlo. No proporciona nada más.				Traducir: Como asistente, necesito el texto para traducirlo. No proporciona nada más.				Contexto: página 8 del PDF■■Texto para traducir: Como asistente, necesito el texto para traducirlo. No proporciona nada más.			
C_{OMET}	0.666	0.764	0.630	0.536	0.249	0.350	0.159	0.128	0.660	0.764	0.630	0.536	0.249	0.350	0.159	0.128
$C_{OMET\ RANGO}$	0.615	0.729	0.589	0.512	0.142	0.243	0.037	0.013	0.615	0.729	0.589	0.512	0.142	0.243	0.037	0.013
$\Delta\tau$	0.051	0.035	0.041	0.024	0.107	0.155	0.119	0.115	0.045	0.035	0.041	0.024	0.107	0.155	0.119	0.115
RANGO-COMETA (solo ...	0.660	0.764	0.630	0...	0.249	0.350	0.159	0...	0.660	0.764	0.630	0...	0.249	0.350	0.159	0...

reflejado en un nivel superior para los pares de idiomas con Inglés como objetivo.

6 Reproducibilidad

Lanzaremos tanto la base de código de la C_{OMET} marco de trabajo y los modelos de evaluación de MT entrenados descrito en este documento para la comunidad de investigación al publicarse, junto con los guiones detallados se requiere para ejecutar todas las líneas de base informadas.[6] Todos los modelos reportados en este documento fueron entrenados en un Contexto: página 8 del PDF■■Texto para traducir: una sola GPU Tesla T4 (16GB). Además, nuestro marco el trabajo se basa en PyTorch Lightning (Falcón, 2019), un envoltorio ligero de PyTorch, que fue creado para máxima flexibilidad y reproducibilidad.

7 Trabajo Relacionado

Las métricas clásicas de evaluación de MT son comúnmente caracterizado como métricas de coincidencia de n-gramas utilizando características hechas a mano, estimando la calidad de MT-idad al contar el número y la fracción de n-gramas que aparecen simultáneamente en un candidato Hipótesis de traducción y una o más humanas-referencias. Métricas como B_{LEU} (■■Devuelve SOLO el texto traducido, nada más. 2002), M_{ETEOR} (■■Devuelve SOLO el texto traducido, nada más. y $_{CHR}F$ (Popović, 2015) han sido ampliamente estudiado y mejorado (Koehn et al., 2007; Popović, 2017; Denkowski y Lavie, 2011; Guo y Hu, 2019), pero, por diseño, generalmente no logran reconocer y capturar la similitud semántica más allá de lo léxico nivel.

En los últimos años, las incrustaciones de palabras (Mikolov y otros, 2013; Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2019) han surgido como un com- alternativa comúnmente utilizada a n-gramas para capturando la similitud semántica de las palabras. métricas basadas como M_{ETEOR}-W_{ECTOR} (■■Devuelve SOLO el texto traducido, nada más. y otros, 2016), $B_{LEU}2_{VEC}$ (■■Devuelve SOLO el texto traducido, nada más. 2017), $Y_{YO}S_{-}1$ (Lo, 2019), $M_{CONTRNNACLES}$ (■■Devuelve SOLO el texto traducido, nada más. 2019), y $B_{ERTSCORE}$ (■■Devuelve SOLO el texto traducido, nada más. 2020) crean alineaciones suaves entre referencia e hipótesis en un espacio de incrustación y luego calcular una puntuación que refleja la similitud semántica entre esos segmentos. Sin embargo, juicios humanos como DA y MQM, capturan mucho más que solo semántica similitud, resultando en una correlación superior- vinculado entre los juicios humanos y las puntuaciones producido por tales métricas.

Métricas aprendibles (■■Devuelve SOLO el texto traducido, nada más. 2018; Mathur et al., 2019; Shimanaka et al., 2019) entrenado para optimizar directamente la correlación con juicios del hombre, y recientemente han mostrado promesa resultados de ing- $_{LEURT}$ (■■Devuelve SOLO el texto traducido, nada más. 2020), un texto métrica capaz basada en BERT (Devlin et al., 2019), afirma un rendimiento de última generación para los últimos años de la tarea compartida de métricas WMT. Porque B_{LEURT} se construye sobre English-BERT (Devlin y otros, 2019), solo puede usarse cuando el inglés es el idioma objetivo que limita su aplicabilidad. También, según nuestro mejor conocimiento, todo lo anteriormente Las métricas aprendibles propuestas se han centrado en o minimizando DA que, debido a la escasez de anotadores, puede demostrar ser inherentemente ruidoso (Mathur et al., 2019).

Evaluación de MT sin referencias también conocido como Estimación de Calidad (QE), históricamente a menudo ha re progresado en HTER para la evaluación a nivel de segmento jar et al., 2013, 2014, 2015, 2016, 2017a). Más recientemente, MQM ha sido utilizado para el nivel de docu evaluación (Specia et al., 2018; Fonseca et al., 2019). Al aprovechar pre- altamente multilingüe codificadores entrenados como BERT multilingüe (Devlin et al., 2019) y XLM (Conneau y Lample, 2019), los sistemas QE han estado mostrando correlaciones sospechosas con juicios humanos (Kepler et al., 2019a). Concurrentemente, el OpenKiwi marco (Kepler et al., 2019b) ha facilitado para que los investigadores avancen en el campo y construy modelos QE más fuertes.

8 Conclusiones y Trabajo Futuro

En este documento presentamos $_{OMET}$, una nueva neu- Marco general para la formación de modelos de evaluación que pueden servir como métricas automáticas y ser fácilme

[6] Estos se alojarán en: https://github.com/Unbabel/COMET

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

Para demostrar la eficacia de nuestro marco, buscamos abordar los desafíos reportados en el Tarea Compartida de Métricas WMT 2019 (Ma et al., 2019). Entrenamos tres modelos distintos que logran algo nuevo 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 abrumador de parámetros y tiempo de inferencia. Una vía principal para trabajo futuro en COMET examinará el impacto de soluciones más compactas como DistilBERT (Sanh et al., 2019).

Además, mientras esbozamos el potencial importancia del texto fuente anterior, notamos que nuestro COMET RANKER modelo de primera clase aún necesita el texto que se va a traducir para poder ayudarte. No de manera diferente durante la inferencia pero igualmente en su entrenamiento. Como el modelo de referencia, el función de pérdida ing. El trabajo futuro investigará el optimalidad de esta formulación y examinar aún más la interdependencia de las diferentes entradas.

Reconocimientos

Estamos agradecidos a Matthew Snover, Fabio Kepler, Daan Van Stigt, Miguel Vera, y 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, supervisado por ANI bajo el contrato número números 045909 y 042671, respectivamente.

Referencias

Mikel Artetxe y Holger Schwenk. 2019. Más-incrustaciones de frases multilingües de manera intensiva para cero-transferencia cruzada de idiomas a codificadores y más rápida. Transactions de la Asociación para la Lingüística Computacional-tics, 7:597–610.

Lo¨ıc Barrault, Ondrej Bojar, Marta R. Costa-juss`a, Christian Federmann, Mark Fishel, Yvette Grajamón, 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-ción (WMT19) En Actas de la Cuarta Conferencia sobre Traducción Automática (Volumen 2: Documentos de Tarea Compartida, páginas 1–61, Florencia, Italia. Asociación para la Lingüística Computacional.

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

Ondrej 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 Aleš Tamchyna. 2014. Hallazgos del taller de 2014 sobre traducción automática estadística. En Actas de la Noveno Taller de Traducción Automática Estadística páginas 12–58, Baltimore, Maryland, EE. UU. Asociación para la Lingüística Computacional.

Ondrej 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 máquina Contexto: Página 9 del PDF. Texto a traducir: Segunda Conferencia sobre Traducción Automática, páginas 169–214, Copenhague, Dinamarca. Asociación para Computacional Lingüística.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aur´elie Nev´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 2016 sobre la traducción automática. En Actas de la Primera Conferencia sobre Traducción Automática: Volumen 2, Documentos de Tarea Compartida, páginas 131–198, Berlín, Alemuchos. Asociación para la Lingüística Computacional.

Ondrej 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 de la Taller de 2015 sobre traducción automática. En Actas del Décimo Taller sobre Estadística Traducción Automática, páginas 1–46, Lisboa, Portugal. Asociación para la Lingüística Computacional.

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

Aljoscha Burchardt y Arle Lommel. 2014. Practi-Directrices de Uso de MQM en Investigación Científica búsqueda 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 Zettlemoyer, y Veselin Stoyanov. 2019. No supervisado aprendizaje de representación entre idiomas. CoRR, preprint arXiv:1911.02116. Como no se proporcionó ningún texto para traducir, no puedo proporcionar una traducción.

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

Michael Denkowski y Alon Lavie. 2011. Métrica automática para optimización confiable y eva- Evaluación de sistemas de traducción. En Actas del Sexto Taller sobre Máquina Estadística Traducción, páginas 85–91, Edimburgo, Escocia. Asociación para la Lingüística Computacional.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, y Kristina Toutanova. 2019. BERT: Pre-entrenamiento de transformadores bidireccionales profundos para el lenguaje de pie. En Actas de la Conferencia 2019 de la Sección Norteamericana de la Asociación para la Lingüística Computacional: Lenguaje Humano Tecnologías, Volumen 1 (Documentos Largos y Cortos), páginas 4171–4186, Minneapolis, Minnesota. Asociación para la Lingüística Computacional.

WA Falcon. 2019. PyTorch Lightning: El liviano Envoltorio de PyTorch para investigación de IA. GitHub. Como no se proporcionó ningún texto para traducir, no puedo proporcionar una traducción.

Erick Fonseca, Lisa Yankovskaya, Andr´e F. T. Martins, Mark Fishel, y Christian Federmann. 2019. Buscar-resultados de las tareas compartidas del WMT 2019 sobre la 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 Lingüística Computacional.

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

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

Yvette Graham, Timothy Baldwin, Alistair Moffat, y Justin Zobel. 2017. ¿Puede el sistema de traducción automática que los sistemas sean evaluados de manera confiable? Ingeniería del Lenguaje Natural, 23(1):3–30.

Yinuo Guo y Junfeng Hu. 2019. Meteor++ 2.0: Adopte el conocimiento de paráfrasis a nivel sintáctico y evaluación de la traducción. En Actas de la Cuarta Conferencia sobre Traducción Automática (Volumen 2: Documentos de Tarea Compartida, Día 1), páginas 501–506, Florencia, Italia. Asociación para la Lingüística Computacional. Como no se proporcionó ningún texto para traducir, no puedo proporcionar una traducción.

Jeremy Howard y Sebastian Ruder. 2018. Universal ajuste fino del texto del lenguaje 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.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, Ant´onio G´ois, M. Amin Farajian, Ant´onio V. Lopes, y André F. T. Martins. 2019a. Unbabel's participación en la tarea compartida de estimación de calidad WMT19. En Actas de la Cuarta Conferencia sobre Traducción Automática (Volumen 3: Documentos de Tarea Compartida, Día 2), páginas 78–84, Florencia, Italia. Asociación para la Lingüística Computacional. Contexto: Página 10 del PDF. Texto a traducir: rence, Italia. Como no se proporcionó ningún texto para traducir, no puedo proporcionar una traducción.

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: Sistema Demostraciones, páginas 117–122, Florencia, Italia. Asociación para la Lingüística Computacional.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, y Evan Herbst. 2007. Moisés: Abierto kit de herramientas de código para la traducción automática est Actas de la 45ª Reunión Anual de la Asociación para la Acompañamiento en Lingüística Computaci Actas del Volumen de la Sesión de Demostración y Póster siones, páginas 177–180, Praga, República Checa. Asociación para la Lingüística Computacional.

Dan Kondratyuk y Milan Straka. 2019. 75 lenguas, 1 modelo: Análisis de dependencias universales universal. En Actas de la Conferencia 2019 Como no se proporcionó ningún texto para traducir, n Conferencia sobre Métodos Empíricos en Lenguaje Natural Procesamiento y la 9na Conferencia Internacional Conjunta Conferencia sobre Procesamiento de Lenguaje Natural (EMNLP-IJCNLP), páginas 2779-2795, Hong Kong, China. Asociación para la Lingüística Computacional.

Alon Lavie y Michael Denkowski. 2009. El meteoro métrica para la evaluación automática de la traducción automá ción. Traducción automática, 23(2-3):105–115. Como no se proporcionó ningún texto para traducir, no pue

Chi-kiu Lo. 2019. YiSi - una calidad de MT semántica unificada evaluación y métrica de estimación para idiomas con diferentes niveles de recursos disponibles. En Actas de la Cuarta Conferencia sobre Traducción Automática (Volumen 2: Documentos de Tarea Compartida, Día 1), páginas 507-513, Florencia, Italia. Asociación para la Lingüística Computacional. Devuelva S Devuelva SOLO el texto traducido, nada más.

Arle Lommel, Aljoscha Burchardt y Hans Uszkoreit. 2014. Métricas de calidad multidimensionales (MQM): A

marco de referencia para declarar y describir la traducción métricas de calidad de texto para traducir, no puedo realizar ninguna traducción. Contexto: Página 11 del PDF Texto a traducir: ducci, 0:455–463.

Qingsong Ma, Ondřej Bojar, y Yvette Graham. 2018. Resultados de la tarea compartida de métricas WMT18: Ambos los caracteres y las incrustaciones logran un buen rendimiento. En Actas de la Tercera Conferencia sobre Traducción Automática: Documentos de Tarea Compartidas, páginas 671-688, Bélgica, Bruselas. Asociación para Computacional.

Qingsong Ma, Johnny Wei, Ondřej Bojar, y Yvette Graham. 2019. Resultados de las métricas WMT19 tarea compartida: Segmento-nivel y fuerte sistema de MT los sistemas 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 Computacional. Contexto: Página 11 del PDF Texto a traducir: rence tics.

Nitika Mathur, Timothy Baldwin y Trevor Cohn. 2019. Ubicando la evaluación en contexto: Contextual los incrustaciones mejoran la evaluación de la traducción automática. En Actas de la 57ª Reunión Anual de la Asociación para la Lingüística Computacional, páginas 2799-2808, Florencia, Italia. Asociación para Computacional.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, y Jeff Dean. 2013. Representación distribuida idad. En Avances en el Procesamiento de Información Neural Sistemas 26, páginas 3111–3119. Curran Associates, Inc. Contexto: Página 11 del PDF Texto a traducir: cion de palabras y frases y

Kishore Papineni, Salim Roukos, Todd Ward, y Wei-Jing Zhu. 2002. Bleu: un método para la evaluación automática evaluación de la traducción automática. En Actas de la 40ª Reunión Anual de la Asociación para Computacional, páginas 311-318, Filadelfia, Pensilvania, EE. UU. Asociación para la Computación Lingüística.

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

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, y Luke Zettlemoyer. 2018. Representación de palabras profundamente representaciones contextualizadas. En Actas de la Conferencia encia de la Sección Norteamericana de la Asociación para la Lingüística Computacional: Lenguaje Humano Tecnologías del Lenguaje, Volumen 1 (Documentos Largos), páginas 2227–2237, Nueva Orleans, Luisiana. Asociación para la Lingüística Computacional. Contexto: Página 11 del PDF Texto a traducir: no se proporcionó ningún texto para traducir, no puedo realizar ninguna traducción.

Telmo Pires, Eva Schlinger, y Dan Garrette. 2019. ¿Qué tan multilingüe es el BERT multilingüe? En actas de la 57ª Reunión Anual de la Asociación para la Lingüística Computacional, páginas 4996–5001, Florencia, Italia. Asociación para la Lingüística Computacional. Contexto: Página 11 del PDF Texto a traducir: no se proporcionó ningún texto para traducir, no puedo realizar ninguna traducción.

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

Maja Popović. 2017. chrF++: palabras que ayudan a caracter n-gramas. En Actas de la Segunda Conferencia sobre Traducción Automática, páginas 612–618, Copenhague, Dinamarca. Asociación para la Computacional. Contexto: Página 11 del PDF Texto a traducir: no se proporcionó ningún texto para traducir, no puedo realizar ninguna traducción.

Nils Reimers e Iryna Gurevych. 2019. BERT: Incrustaciones de oraciones usando Siamese BERT redes. En Actas de la Conferencia 2019 sobre Métodos Empíricos en el 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 Lingüística Computacional. Contexto: Página 11 del PDF Texto a traducir: no se proporcionó ningún texto para traducir, no puedo realizar

Victor Sanh, Lysandre Debut, Julien Chaumond, y Thomas Wolf. 2019. Distilbert, una versión destilada de BERT: más pequeño, más rápido, más barato y más ligero preprint arXiv:1910.01108. Como no se proporcionó ningún texto

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

Thibault Sellam, Dipanjan Das y Ankur Parikh. 2020. BLEURT: Aprendiendo métricas robustas para texto generación. En Actas de la 58ª Reunión Anual Reunión de la Asociación para la Lingüística Computacional páginas 7881–7892, En línea. Asociación para la Computacional. Contexto: Página 11 del PDF Texto a traducir: no se proporcionó ningún texto para traducir, no puedo realizar ninguna traducción

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

Hiroki Shimanaka, Tomoyuki Kajiwara, y Mamoru Komachi. 2018. RUSE: Regresor utilizando oraciones incrustaciones para la evaluación automática de la traducción ción. En Actas de la Tercera Conferencia sobre Traducción Automática: Documentos de Tarea Compartidas, páginas 751-758, Bélgica, Bruselas. Asociación para la Computacional. Contexto: Página 11 del PDF Texto a traducir: no se proporcionó ningún texto para traducir, no puedo realizar ninguna traducción.

Hiroki Shimanaka, Tomoyuki Kajiwara, y Mamoru Komachi. 2019. Evaluación de la Traducción Automática ación con BERT Regresor. Comprimir preprint arXiv arXiv:1907.12679. Como no se proporcionó ningún texto para tra

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, y John Makhoul. 2006. Un estudio de la tasa de edición de traducción con anotaciones humanas dirigidas-ción En En las Actas de la Asociación para la Máquina Traducción en las Américas, páginas 223–231.

Lucia Specia, Fr´ed´eric Blain, Varvara Logacheva, Ramón Astudillo, y André F. T. Martins. 2018. Hallazgos de la tarea compartida WMT 2018 sobre calidad estimación En Actas de la Tercera Conferencia sobre Traducción Automática: Documento página Tareas Compartidas 689-709, Bélgica, Bruselas. Asociación para Com- Lingüística Computacional.

Lucia Specia, Kim Harris, Fr´ed´eric Blain, Aljoscha Burchardt, Viviven Macketanz, Inguna Skadina, Matteo Negri, , y Marco Turchi. 2017. Transla- Calidad de la traducción y productividad: Un estudio sobre mor- idiomas de 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 de máquinas-ción utilizando entradas del idioma fuente y translingüístico modelo de lengua En Actas de la 58ª Reunión Anual Reunión de la Asociación para la Lingüística Computacional lingüística, páginas 3553-3558, En línea. Asociación para Lingüística Computacional.

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

Ian Tenney, Dipanjan Das y Ellie Pavlick. 2019. BERT redescubre la clásica cadena de procesamiento NLP Cerramiento se proporcionó ningún texto para traducir, no puedo realizar ning Actas de la 57ª Reunión Anual de la Asso- Asociación para la Lingüística Computacional, páginas 4593–4601, Florencia, Italia. Asociación para Computacional Lingüística.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, y Yoav Artzi. 2020. Bertscore: Eval- Evaluando la generación de texto Conferencia Internacional proporcionó ningún texto para traducir, no puedo realizar ninguna traducc Conferencia sobre Representaciones de Aprendizaje Cerramiento se proporcionó ningún texto para traducir, no puedo realizar ninguna t

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, y Steffen Eger. 2020. En el límlimitaciones de los codificadores multilingües según lo expuesto por evaluación de traducción automática sin referencia Conferencia se proporcionó ningún texto para traducir, no puedo realizar ning Actas de la 58ª Reunión Anual de la Asso- Asociación para la Lingüística Computacional, páginas 1656–1671, En línea. Asociación para la Lingüística Computacional. lingüística.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, y Steffen Eger. 2019. MoverScore: Evaluación de la generación de texto con em- contextualizado ropa de cama y distancia del movimiento Encuentro de tierra de la Conferencia 2019 sobre Métodos Empíricos en Procesamiento de Lenguaje Natural y la 9na Interna- Conferencia Conjunta Internacional sobre Procesamiento del Lenguaje Natural procesamiento (EMNLP-IJCNLP), páginas 563–578, Hong Kong, China. Asociación para la Computación Lingüística.

A Apéndices

En la Tabla 5 enumeramos los hiperparámetros utilizados para entrenar nuestros modelos. Antes de inicializar estos modelos, se ejecutó un La semilla dom se estableció en 3 en todas las bibliotecas que realizan operaciones "aleatorias" (, , , y).

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

Hiper-parámetro	C_{OMET} (Est-HTER/MQM)	C_{OMET} RANKER
Modelo de Codificador	XLM-RoBERTa (base)	XLM-RoBERTa (base)
Optimizador	Adam (parámetros predeterminados)	Adam (parámetros predeterminados)
en épocas congeladas	1	0
Tasa de aprendizaje	3e-05 y 1e-05	1e-05
Tamaño del lote	16	16
Función de pérdida	EMC	Margen de Triplete (1
Abandono por capas	0.1	0.1
Precisión FP	32	32
Unidades ocultas de retroalimentación anticipada	2304,1152	Traduce el siguiente texto al español
Activaciones Feed-Forward	Tanh	Traduce el siguiente texto al español
Abandono anticipado	0.1	Como asistente, necesito más info

Tabla 6: Estadísticas para el corpus QT21.

	en-de	en-cs	en-lv	de-en
Tuplas totales	54000	42000	35474	41998
Promedio de tokens	17.30	17.37	18.39	17.71
Promedio de tokens	16.70	17.37	18.39	17.18
Promedio de tokens	17.65	15.64	16.42	17.78

Tabla 7: Estadísticas para el WMT 2017 DA corpus RR.

	en-cs	en-de	en-lv	Texto a traducir: en-tr
Tuplas totales	32810	6454	3270	3456
Promedio de tokens	19.70	22.15	21.42	17.57
Promedio de tokens	22.87	23.41	21.73	26.08
Promedio de tokens	19.45	22.58	16.06	17.25

Pares de idiomas RR de inglés y no inglés.

Pares de idiomas RR a inglés.

DA

Tabla 8: Estadísticas para el WMT 2019

DA

Tabla 9: Estadísticas para el WMT 2019

Texto a traducir: (referencia)■■Devuelve SOLO el texto traducido, nada más.
You didn't provide any text to translate. Could you please provide the text you wan
Devuelve SOLO el texto traducido, nada más.

Texto a traducir: (referencia)■■Dev
You didn't provide any text to tran
Devuelve SOLO el texto traduc

Your request is unclear. Please provide the text you want to be translated into Spanish.

Traducir el siguiente texto al español. ■Contexto: página 16 del PDF■■Texto a traducir: en

Texto a traducir: en-tr■■Devuelve SOLO el texto traducido, nada más.

en-es

Pares de idiomas RR.

) estadísticas.

3.3■■Devuelve SOLO el texto traducido, nada más. DA

Tabla 10: Corpus MQM (sección

Tabla 11: Estadísticas para el WMT 2018

You didn't provide any text to translate. Please provide the text you want translated into Spa

Contexto: página 16 del PDF■■Texto para traducir: en-ru■■Devuelve SOLO el texto tradu

Traducir el siguiente texto al español.■Con

Contexto: Página PDF 16■■Texto para traducir: en-ja■■Devuelve SOLO el texto traducido

Contexto: página 16 del PDF■■Texto a tra

en-zh■■Devuelve SOLO el texto traducido

(referencia)

You didn't provide any text to translate. Please provide the text you want translated

Traduce el siguiente texto al español.■Contexto: Página 16 del PDF■■Texto p

(referencia)

You didn't provide any text to tran

Traduce el siguiente texto al es

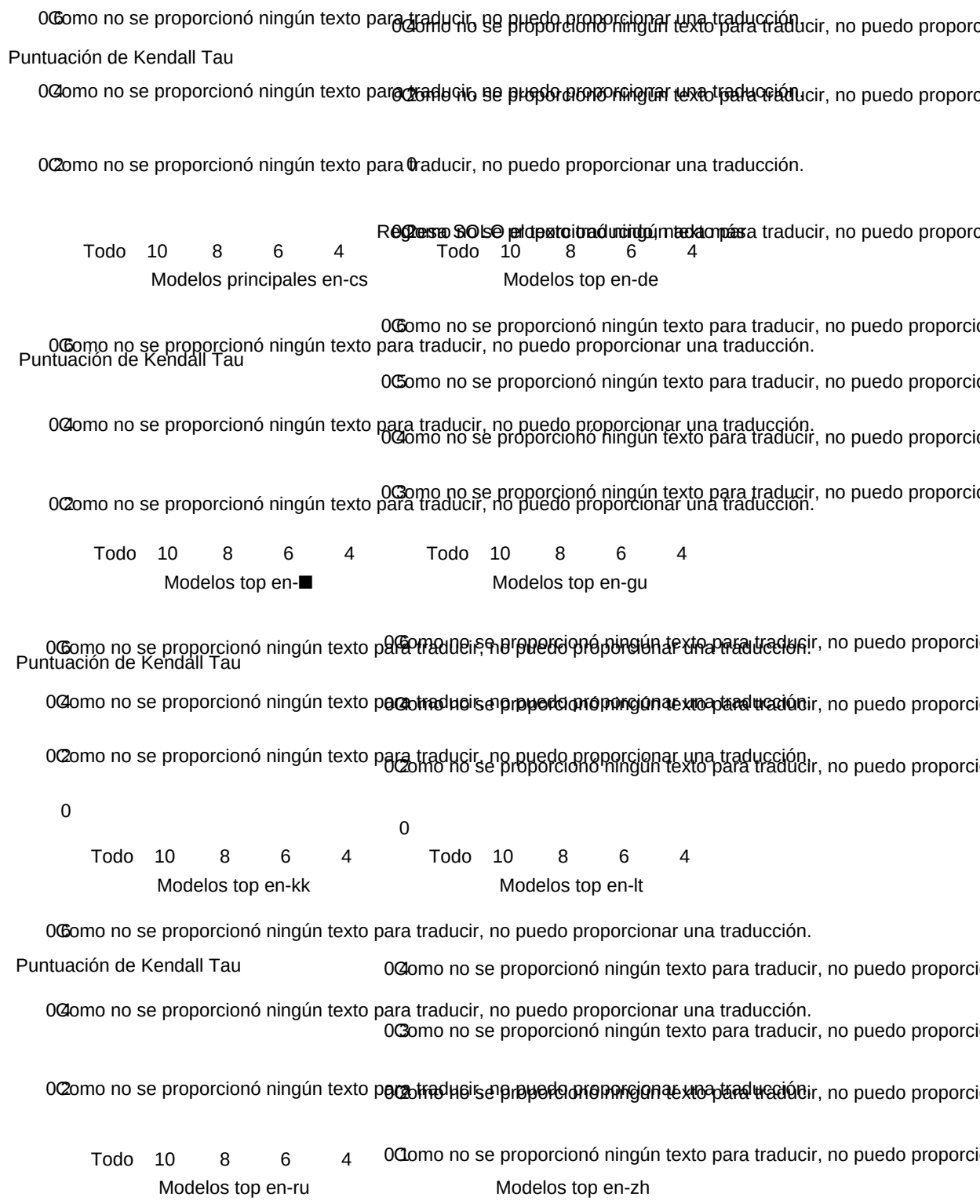

Tabla 12: Rendimiento de métricas en todos y los mejores (10, 8, 6 y 4) sistemas MT para todos los idiomas desde ing
pares. El esquema de colores es el sig
BERTSCORE

0.2 Como no se proporcionó ningún texto para traducir, no puedo proporcionar una traducción.

Puntuación de Kendall Tau

0.1 Como no se proporcionó ningún texto para traducir, no puedo realizar la traducción. Por favor, proporcione el texto

0

Regresa SOLO el texto traducido, nada más

Todo 10 8 6 4

Modelos top de-en

0.4 Como no se proporcionó ningún texto para traducir, no puedo proporcionar una traducción.

0.3 Como no se proporcionó ningún texto para traducir, no puedo propor

0.2 Como no se proporcionó ningún texto para traducir, no puedo propor

0.1 Como no se proporcionó ningún texto para traducir, no puedo propor

Todo 10 8 6 4

Modelos top ■-en

0.4 Como no se proporcionó ningún texto para traducir, no puedo proporcionar una traducción.

Puntuación de Kendall Tau

0.3 Como no se proporcionó ningún texto para traducir, no puedo proporcionar una traducción.

0.2 Como no se proporcionó ningún texto para traducir, no puedo proporcionar una traducción.

0.1 Como no se proporcionó ningún texto para traducir, no puedo proporcionar una traducción.

Todo 10 8 6 4

Modelos superiores lt-en

0.2 Como no se proporcionó ningún texto para traducir, no puedo real

0.15 Como no se proporcionó ningún texto para traducir, no puedo realiz

0.1 Como no se proporcionó ningún texto para traducir, no puedo real

5 · 10 Regresa SOLO el texto traducido, nada más

0

Todo 10 8 6 4

Modelos top ru-en

0.4 Como no se proporcionó ningún texto para traducir, no puedo realizar la traducción. Por favor, proporcione el texto

Puntuación de Kendall Tau

0.2 Como no se proporcionó ningún texto para traducir, no puedo realizar la traducción. Por favor, proporcione el texto

0

Todo 10 8 6 4

Modelos top zh-en

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