C OMET : Un Marco Neural para la Evaluación de MT

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

Resumen

Presentamos C OMET , un marco neuronal para entrenamiento de evaluación de traducción automática multilingüe modelos de situación que obtienen un nuevo estado de niveles de arte de correlación con el juez humano- Nuestro marco de trabajo aprovecha los recientes avances. avances en el lenguaje preentrenado multilingüe modelado que resulta en altamente multilingüe y modelos de evaluación de MT adaptables que explotan información tanto de la entrada fuente como de una referencia de traducción en el idioma objetivo en orden para predecir con mayor precisión la calidad de la MT. A para mostrar nuestro marco de trabajo, entrenamos tres mod- modelos con diferentes tipos de juicios humanos: Evaluaciones Directas , Trans- mediado por humanos Tasa de Edición de Traducción y Calidad Multidimensional Métricas de Ciudad Nuestros modelos logran un nuevo estado de el rendimiento de última generación en el WMT 2019 Met- tarea compartida de rics y demuestran robustez para sistemas de alto rendimiento.

1 Introducción

Históricamente, las métricas para evaluar la calidad de la traducción automática (MT) ha dependido de 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 conteo el número de n-gramas coincidentes entre la MT hipótesis y la traducción de referencia. Métricas como B LEU) Papineni y otros. , 2002) y M ETEOR) Lavie y Denkowski , 2009) siguen siendo populares como un medio para evaluar sistemas de MT debido a sus cálculo ligero y rápido.

Los enfoques neuronales modernos para la traducción automática resultan en mucho mayor calidad de traducción 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 B LEU para proporcionar una estimación precisa de la calidad de MT (Barrault y otros. , 2019).

Mientras que un aumento del interés de investigación en neural métodos para entrenar modelos y sistemas de MT tiene resultó en una reciente y dramática mejora en la MT calidad, la evaluación de MT se ha quedado atrás. El MT la comunidad de investigación aún depende en gran medida de métodos obsoletos métricas y no se ha adoptado ningún nuevo estándar de manera generalizada surgió. En 2019, la Traducción de Noticias WMT La Tarea Compartida recibió un total de 153 sistemas de Traducción Automática envíos Barrault y otros. , 2019 Las Métricas La Tarea Compartida del mismo año solo vio 24 sub- misiones, casi la mitad de las cuales eran participantes en el Tarea Compartida de Estimación de Calidad, adaptada como métricas) Ma y otros. , 2019).

Los resultados de la tarea anteriormente mencionada son altos. ilumina dos desafíos principales para la evaluación de MT que buscamos abordar aquí Ma y otros. , 2019). Es decir, que las métricas actuales lucha por acumular raramente se correlaciona con el juicio humano en seg- nivel mental y no logran diferenciar adecuadamente los sistemas de MT de mayor rendimiento .

En este artículo, presentamos C OMET [1], un PyTorch marco de trabajo basado para la formación altamente multilingüe y modelos de evaluación de MT adaptables que pueden func- como métricas. Nuestro marco de trabajo aprovecha de avances recientes en lenguaje interlingüístico modelado Artetxe y Schwenk , 2019 ; Devlin y otros. , 2019 ; Conneau y Lample , 2019 ; Conneau y otros. , 2019 para generar estimaciones de predicción de hu- juicios de hombre tales como Evaluaciones Directas (DE)) Graham y otros. , 2013), Traducción mediada por humanos Tasa de Edición de Traducción (HTER) (Snover y otros. , 2006) y métricas compatibles con el Calidad Multidimensional Métrica de Ciudad marco de trabajo) Lommel y otros. , 2014).

Inspirado en trabajos recientes sobre Estimación de Calidad (QE) que demostró que es posible lograr altos niveles de correlación con los juicios humanos incluso sin una traducción de referencia (Fonseca y otros. , 2019), proponemos un enfoque novedoso para incorporar

[1] C multilingüe O optimizado M métrica para E valoración de T traducción.

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

traduciendo la entrada del idioma fuente en nuestra evaluación de MT modelos de ation. Tradicionalmente solo los modelos QE han hizo uso de la entrada de la fuente, mientras que la evaluación de MT- las métricas de evaluación dependen en cambio de la traducción de referencia- Como en (Takahashi et al. , 2020), demostramos que usar un espacio de incrustación multilingüe nos permite para aprovechar la información de las tres entradas y demuestra el valor agregado por la fuente como entrada a nuestros modelos de evaluación de MT.

Para ilustrar la eficacia y flexibilidad de la C OMET marco de trabajo, entrenamos tres modelos que estimar diferentes tipos de juicios humanos y muestra un progreso prometedor hacia una mejor correlación a nivel de segmento y robustez ante altas calidad MT.

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

2 Arquitecturas de Modelos

Los juicios humanos sobre la calidad de la traducción automática generalmente provienen en la forma de puntuaciones a nivel de segmento, como DA, MQM y HTER. Para DA, es una práctica común convertir puntuaciones en clasificaciones relativas (DA RR) cuando el número de anotaciones por segmento es limitado) Bojar y compañía. , 2017b ; Ma y otros. , 2018 , 2019). Esto significa que, para dos hipótesis de MT h_i y h_j de la misma fuente , si la puntuación DA asignada a h_i es más alto que la puntuación asignada a h_j, h_i es reconsiderada como una "mejor" hipótesis.[2] Abranger estas diferencias, nuestro marco de trabajo admite dos desarquitecturas distintas: The Modelo estimador y el Modelo de clasificación de traducción . El 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 de sus referencias correspondientes 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 modelo preentrenado, multilingüe modelo como BERT multilingüe (Devlin et al. , 2019), XLM (Conneau y Lample , 2019) o XLM-RoBERTa Conneau y compañía. , 2019). Estos modelos contiene varias capas de codificador transformador que son entrenado para reconstruir tokens enmascarados al descubrir- analizando la relación entre esos tokens y el los que rodean. Cuando se entrena con datos de este objetivo preentrenado tiene múltiples idiomas se ha encontrado que es altamente efectivo en el cruce de idiomas tareas como la clasificación de documentos y el procesamiento natural de lenguaje inferencia de lenguaje) Conneau y compañía. , 2019), gener- adaptándose bien a idiomas y guiones no vistos anteriormente (Pires y otros. , 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 $e_j^{(\ell)}$ para cada uno ficha x_j y cada capa $\ell \in \{0, 1, ..., k\}$. En nuestro marco de trabajo, aplicamos este proceso a la fuente, Hipótesis de MT, y referencia para mapearlas en 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 para el ajuste fino 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 y colaboradores. , 2020) mostró que las diferentes capas pueden alcanzar diferentes niveles de correlación y que utilizando solo la última capa a menudo resulta en un rendimiento inferior. En este trabajo, nosotros utilizó el enfoque descrito en Peters y otros.) 2018) y recopilar información de las fuentes más importantes en- codifica las capas en una única incrustación para cada uno de- ken, e_j , utilizando un mecanismo de atención por capas. Esta incrustación se calcula entonces como:

$$e_{x_j} = \mu E_{x_j}^{T} \alpha \quad (1)$$

donde μ es un coeficiente de peso entrenable, $E_j = [e_j^{(0)}, e_j^{(1)}, \ldots e_j^{(k)}]$ corresponde al vector de incrustaciones de capa para token x_j , y $\alpha = \text{softmax}([\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(k)}])$ es un vector correspondiente correspondiendo a los pesos entrenables por capas. En ordenar para evitar el sobreajuste a la información con- utilizado en cualquier capa individual, utilizamos la eliminación de capas) Kondratyuk y Straka , 2019), en el que con un probabilidad p el peso $\alpha^{(i)}$ está configurado para $-\infty$.

Finalmente, como en (Reimers y Gurevych , 2019), aplicamos el promedio de agrupación a la palabra resultante incrustaciones para derivar una incrustación de oración para 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- se excluyen de la DA Datos RR.

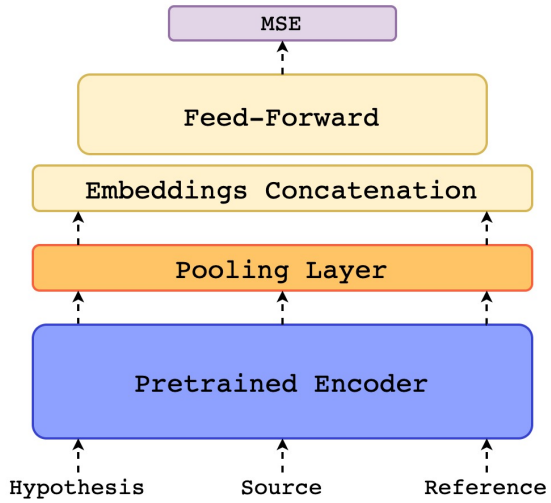

Figura 1: Arquitectura del modelo de estimador. La fuente, la hipótesis y la referencia se codifican de manera independiente en nosotros usando un codificador preentrenado multilingüe. El resultado los incrustaciones de palabras luego pasan a través de un agrupamiento capa para crear una incrustación de oración 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).

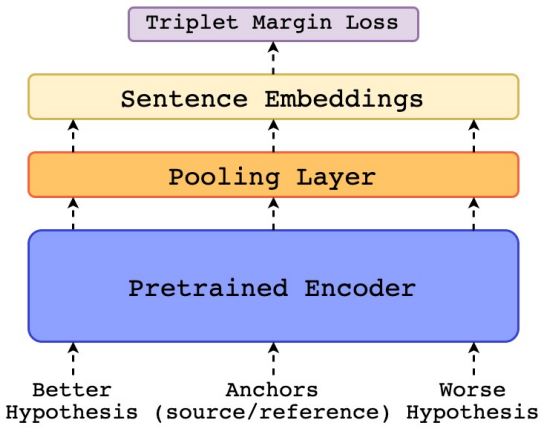

Figura 2: Arquitectura del modelo de clasificación de traducciones. Esta arquitectura recibe 4 segmentos: la fuente, el referencia, una hipótesis "mejor" y una "peor". Estos segmentos se codifican de manera independiente utilizando un precodificador entre lenguas entrenado y una capa de agrupación en finalmente, utilizando la pérdida de margen de tríada (Schroff y otros. , 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 un d incrustación de frases en dimensión para el fuente, la hipótesis y la referencia, adoptamos el enfoque propuesto en RUSE (Shimanaka y otros. , 2018) y extraer las siguientes características combinadas:

- Producto fuente elemento por elemento h ⊙ s
- Producto de referencia elemento por elemento: h ⊙ r
- Diferencia absoluta elemento por elemento de la fuente: |h – s|
- Diferencia absoluta de referencia elemento por elemento. |h – r|

Estas características combinadas se concatenan entonces. a la incrustación de referencia r y hipótesis empropa de cama h en un solo vector x = [h ; r ; h ⊙ s ; h ⊙ r ; | h – s | ; | h – r |] que sirve como entrada para un regresor de avance directo. La fuerza de estos se destaca en resaltar las diferencias entre incrustaciones en el espacio de características semánticas.

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

Tenga en cuenta que elegimos no incluir la fuente original incrustación s) en nuestra entrada concatenada. Temprano la experimentación reveló que el valor agregado por la incrustación de origen como características de entrada adicionales para nuestro el regresor fue insignificante en el mejor de los casos. Una variación en nuestro modelo estimador de HTER entrenado con el vector x = [h ; r ; h ⊙ s ; h ⊙ r ; | h – s | ; | h – r |] como la entrada a la red feed-forward solo logra impulsar rendimiento a nivel de segmento en 8 de los 18 idiomas pares de idiomas descritos en la sección 5 abajo y el mejora promedio en el Tau de Kendall en esos conjuntos las cosas fueron +0.0009. Como se señaló en Zhao et al.) 2020), mientras que los modelos preentrenados multilingües son adaptables a múltiples idiomas, el espacio de características entre los idiomas están mal alineados. Sobre esta base, nosotros desdecidieron 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 reducla ampliación del espacio de características permitiría al modelo enfócate más en la información relevante. Esto no sin embargo, niega el valor general de la fuente a nuestro modelo; donde incluimos características combinadas como h ⊙ s y | h – s | notamos ganancias en correlación como se explora más adelante en la sección 5.5 abajo.

2.4 Modelo de Clasificación de Traducción

Nuestro modelo de clasificación de traducciones (Figura 2) recibe como entrada una tupla $\chi = (s, h^+, h^-, r)$ donde h^+ de anota una hipótesis que fue clasificada más alta que otra hipótesis h^- Luego pasamos χ a través nuestro codificador cruzado de idiomas y capa de agrupación para obobtener una incrustación de oración para cada segmento en el χ . Finalmente, utilizando los incrustaciones $\{s, h^+, h^-, r\}$, calculamos la pérdida de margen de tríada (Schroff y otros., 2015) en relación con la fuente y la referencia:

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

donde

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

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

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

Durante la inferencia, el modelo descrito recibe un trillizo $(s, \hat{h}, r)$ con solo una hipótesis. El puntuación de calidad asignada a $\hat{h}$ es la media armónica entre la distancia a la fuente $d(s, \hat{h})$ y el distancia a la referencia $d(r, \hat{h})$

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

Finalmente, convertimos la distancia resultante en una puntuación de similitud limitada entre 0 y 1 como sigue bajos:

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

3 Corpora

Para demostrar la efectividad de nuestro método descrito arquitecturas de modelos (sección 2), entrenamos tres MT modelos de evaluación donde cada modelo se dirige a un difdiferente tipo de juicio humano. Para entrenar estos modelos, utilizamos datos de tres corpus diferentes: el corpus QT21, el DA RR del WMT Mettarea compartida de rics (2017 a 2019) y una propiedad propietaria 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 un dominio de tecnología de la información o ciencias de la vida) Specia y otros., 2017). Este corpus contiene un total de 173K tuplas con la frase fuente, respectivamente referencia generada por humanos, hipótesis de MT (ya sea de un MT estadístico basado en frases o de un neuMT en bruto) y MT post-editado (PE). El idioma los pares representados en este corpus son: inglés a alemánhombre (en-es), 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 y otros., 2006 ser entre la hipótesis de MT y el PE correspondiente. Finalmente, después de calcular el HTER para cada MT, construimos un conjunto de datos de entrenamiento $D = \{s_n, h_n, r_n, y_n\}_{n=1}^{N}$, donde s_n denota el texto fuente, h_n denota el MT hipótesis, r_n la traducción de referencia, y y_n el Puntuación HTER para la hipótesis h_n De esta manera buscamos aprender una regresión $f(s, h, r) \rightarrow y$ que 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) Barrault y otros., 2019 tengo juicios humanos recopilados en forma de anuncios DAs de equivalencia Graham y otros., 2013 , 2014 , 2017). Estos DAs se asignan entonces en rangos relativosanillos DA RR) (Ma y otros., 2019). El resultado los datos de cada año (2017-19) forman un conjunto de datos $D = \{s_n, h_n^+, h_n^-, r_n\}_{n=1}^{N}$ donde h_n^+ denota un "mejor" hipótesis y h_n^- denota uno "peor". Aquí buscamos aprender una función $r(s, h, r)$ de tal manera que el puntuación asignada a h_n^+ es estrictamente superior a la puntuación asignado a h_n^-) $r(s_n, h_n^+, r_n) > r(s_n, h_n^-, r_n)$). Estos datos [4] Debes proporcionar un texto para traducir. contiene un total de 24 altos y bajos pares de idiomas de recursos como chino a inglés (zh-en) y de inglés a 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/repositorio/xmlui/manejar/11372/LRT-2390

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

mensajes de chat que fueron anotados de acuerdo con el directrices establecidas en Burchardt y Lommel) 2014). Estos datos contienen un total de 12K tuplas, cubren traduciendo 12 pares de idiomas del inglés al alemán (es-de), español (en-es), español latinoamericano español (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). Ten en cuenta que en este corpus inglés siempre se ve como el idioma fuente-idioma, pero nunca como el idioma objetivo. Cada tupla consiste en una oración fuente, generada por un humano referencia, una hipótesis de MT y su puntuación MQM, derivado de las anotaciones de error por uno (o más) anotadores entrenados. La métrica MQM a la que se hace referencia. a lo largo de este documento se define una métrica interna de acuerdo con el marco de MQM (Lommel y otros., 2014) (MQM). Los errores están anotados debajo una tipología interna definida bajo tres errores principales tipos de errores; 'Estilo', 'Fluidez' y 'Precisión'. Nuestro Los puntajes de MQM varían desde $-\infty$ a 100 y son definido como:

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

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

Nuestra métrica MQM toma en cuenta la severidad-cantidad de errores identificados en la hipótesis de MT, conduciendo a una métrica más detallada que HTER o DA. Cuando se utilizaron en nuestros experimentos, estos vallos valores se dividieron por 100 y se truncaron a 0. Como en la sección 3.1 , construimos un conjunto de datos de entrenamiento $D = \{ s_{\text{Por favor, proporcione el segmento que desea traducir.}}, h_{\text{Por favor, proporcione el segmento que desea traducir.}}, r_{\text{Por favor, proporcione el segmento que desea traducir.}}, y_{\text{Por favor, proporcione el segmento que desea traducir.}} \}_{n=1}^{N}$, donde $s_{\text{Por favor, proporcione el segmento que desea traducir.}}$ denota el texto fuente, $h_{\text{Por favor, proporcione el segmento que desea traducir.}}$ denota la hipótesis MT, $r_{\text{Por favor, proporcione el segmento que desea traducir.}}$ el traducción de referencia, y $y_{\text{Por favor, proporcione el segmento que desea traducir.}}$ la puntuación MQM para la hipótesis $h_{\text{Por favor, proporcione el segmento que desea traducir.}}$.

4 Experimentos

Entrenamos dos versiones del modelo Estimador de descrito en la sección 2.3 : uno que retrocede en HTER (C OMET - HTER) entrenado con el corpus QT21, y otro que retrocede en nuestra implementación propietaria tación de MQM (C OMET - MQM 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 - RANGO). En esta sección, introducimos el entrenamiento configuración para estos modelos y evaluación correspondiente configuración.

4.1 Configuración de Entrenamiento

Las dos versiones de los Estimadores (C OMET - (HTER/MQM) comparten la misma configuración de entrenamiento y hiper-parámetros (los detalles están incluidos en el Appendices). Para el entrenamiento, cargamos el preentrenado codificador e inicializar tanto la capa de agrupación como la el regresor de avance directo. Mientras que la capa por capa escalares α desde la capa de agrupación se establecen inicialmente a cero, los pesos del avance directo se inicializan inicializado aleatoriamente. Durante el entrenamiento, dividimos el parámetros del modelo en dos grupos: los parámetros del codificador parámetros, que incluyen el modelo del codificador y el escalares de α ; y los parámetros del regresor, que incluye los parámetros de la alimentación hacia adelante superior red. Aplicamos descongelamiento gradual y discriminación. tasas de aprendizaje innovadoras) Howard y Ruder , 2018), lo que significa que el modelo de codificador está congelado por uno época mientras que el feed-forward se optimiza con un tasa de aprendizaje $3e-5$. Después de la primera época, el todo el modelo está afinado pero la tasa de aprendizaje para los parámetros del codificador están configurados en $1e-5$ con el fin de evita el olvido catastrófico.

A diferencia de los dos estimadores, para el C OMET - RANGO modelo que ajustamos desde el principio. Además, dado que este modelo no agrega ningún nuevos parámetros encima de XLM-RoBERTa (base) además de los escalares de capa α , utilizamos uno solo tasa de aprendizaje $1e-5$ para todo el modelo.

4.2 Configuración de Evaluación

Utilizamos los datos de prueba y la configuración del WMT 2019. Tarea Compartida de Métricas Ma y otros., 2019) con el fin de compara 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 de-Formulación oficial similar a Tau de Kendall. τ , desde el Tarea Compartida de Métricas WMT 2019 (Ma y otros., 2019) definido como:

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

donde Concordante es el número de veces una métrica asigna una puntuación más alta a la "mejor" hipótesis h^{+} y Discordante es el número de veces una métrica asigna una puntuación más alta a la "peor" hipótesis

[5] Para facilitar futuras investigaciones, también proporcionaremos, dentro de nuestro marco de trabajo, instrucciones detalladas y scripts para ejecutar otros metmétricas como CHR 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 WMT19 DA RR corpus. Para B ERTSCORE informamos resultados con el modelo de codificador predeterminado para una comparación completa, pero también con XLM-RoBERTa (base) para la equidad con nuestros modelos. Los valores reportados para YiSi-1 se toman directamente de el documento de tarea compartida (Ma y otros. , 2019).

Métrico	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh
B LEU	0.364	0.248	0.395	0.463	0.363	0.333	0.469	0.235
CHR F	0.444	0.321	0.518	0.548	0.510	0.438	0.548	0.241
Y ʸˢ S ʸˢ -1	0.475	0.351	0.537	0.551	0.546	0.470	0.585	0.355
B ERTSCORE (default)	0.500	0.363	0.527	0.568	0.540	0.464	0.585	0.356
B ERTSCORE (xlmr-base)	0.503	0.369	0.553	0.584	0.536	0.514	0.599	0.317
C OMET - HTER	0.524	0.383	0.560	0.552	0.508	0.577	0.539	0.380
C OMET - MQM	0.537	0.398	0.567	0.564	0.534	0.574	0.615	0.378
C OMET - RANGO	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 igual.

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

5 Resultados

5.1 Del inglés al X

Mesa 1 muestra resultados para los ocho pares de idiomas con inglés como fuente. Contrastamos nuestros tres C OMET modelos contra métricas de referencia como B LEU y CHR F, la métrica ganadora de la tarea 2019 Y ʸˢ S ʸˢ -1, así como el más reciente B ERTSCORE . Observamos que en general nuestros tres modelos entrenado con la C OMET el marco supera, a menudo por márgenes significativos, todas las demás métricas. Nuestro DA El modelo RR Ranker supera a los dos Estimatraductores en siete de ocho pares de idiomas. Además, incluso aunque el Estimador MQM se entrena solo en 12K segmentos anotados, se desempeña aproximadamente a la par con el Estimador HTER para la mayoría de los pares de idiomas, y supera todas las otras métricas en en-ru.

5.2 De X al inglés

Mesa 2 muestra resultados para los siete idiomas al inglés pares de idiomas. Nuevamente, contrastamos nuestros tres C OMET modelos contra métricas de referencia como B LEU y CHR F, la métrica ganadora de la tarea 2019 Y ʸˢ S ʸˢ -1, como así como las métricas B recientemente publicadas ERTSCORE y B LEURT . Como en la Tabla 1 el DA El modelo RR muestra fuertes correlaciones con los juicios humanos superan realizando la recientemente propuesta específica para inglés B LEURT métrica en cinco de siete pares de idiomas. De nuevo, el Estimador MQM muestra una sorprendente fortaleza resultados a pesar de que este modelo fue entrenado con datos que no incluían el inglés como objetivo. Aunque el codificador utilizado en nuestros modelos entrenados es altamente multilingüe, hipotetizamos que este poder- el poderoso resultado de "cero disparos" se debe a la inclusión de la fuente en nuestros modelos.

5.3 Pares de idiomas que no involucran el inglés

Todos nuestros tres C OMET los modelos fueron entrenados en datos que involucran al inglés (ya sea como fuente o como objetivo). Sin embargo, para demostrar que nuestro metrics se generalizan bien, los probamos en las tres WMT 2019 pares de idiomas que no incluyen inglés en ya sea fuente o destino. Como se puede ver en la Tabla 3 , nuestros resultados son consistentes con las observaciones en Mesas 1 y 2 .

5.4 Robustez ante MT de alta calidad

Para el análisis, utilizamos el DA Corpus RR 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 podríamos recuperar datos de al menos diez diferentes sistemas de MT (es decir, todos excepto kk-en y gu-en). Contrastamos contra lo fuerte propuesto recientemente B ERTSCORE y B LEURT , con B LEU como base línea. Los resultados se presentan en la Figura 3 Para lanpares de idiomas donde el inglés es el objetivo, nuestros tres los modelos son mejores o competitivos con todos los demás donde el inglés es la fuente, lo notamos en generalmente, nuestras métricas superan el rendimiento de otros

Tabla 2: Tau de Kendall (τ) correlaciones en pares de idiomas con inglés como objetivo para las Métricas WMT19 DA RR corpus. En cuanto a BERTSCORE , para BLEURT informamos resultados para 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.

Métrico	de-en	Por favor, proporciona el texto que necesitas traducir.	gu-en	kk-es	lt-es	ru-en	zh-en
BLEU	0.053	0.236	0.194	0.276	0.249	0.177	0.321
CHRF	0.123	0.292	0.240	0.323	0.304	0.115	0.371
YiSi-1	0.164	0.347	0.312	0.440	0.376	0.217	0.426
BERTSCORE (default)	0.190	0.354	0.292	0.351	0.381	0.221	0.432
BERTSCORE (xlmr-base)	0.171	0.335	0.295	0.354	0.356	0.202	0.412
BLEURT (base-128)	0.171	0.372	0.302	0.383	0.387	0.218	0.417
BLEURT (grande-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-RANGO	0.202	0.399	0.341	0.358	0.407	0.180	0.445

Tabla 3: Tau de Kendall (τ correlaciones en el lenguaje pares que no involucran inglés para las Métricas WMT19 DA Corpus RR.

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

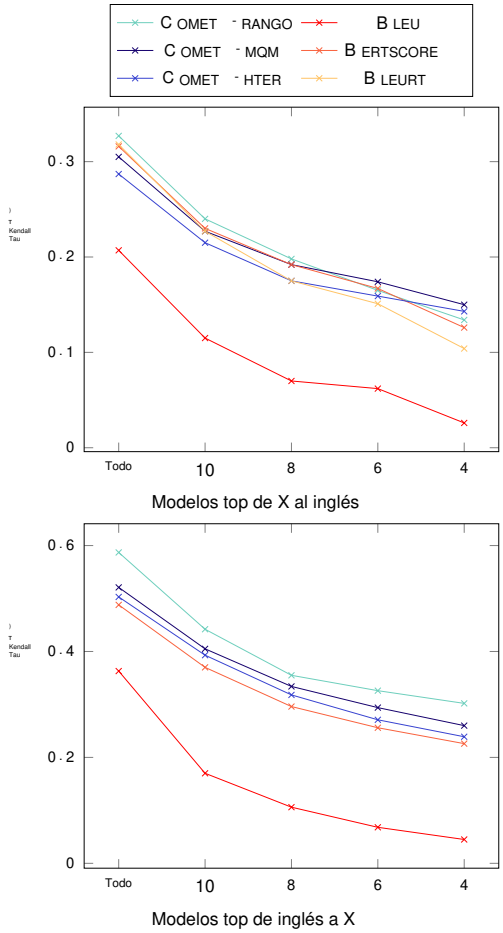

Figura 3: Rendimiento de las métricas en general y en los primeros (10, 8, 6 y 4) Sistemas MT.

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- contribución de la entrada del idioma fuente en nuestros modelos capacidad para aprender predicciones precisas, entrenamos dos versiones de nuestro DA Modelo RR Ranker: uno que utiliza solo la referencia, y otro que utiliza ambas referencias. ence y source. Ambos modelos fueron entrenados usando el corpus WMT 2017 que solo incluye idioma pares desde inglés (en-de, en-cs, en-fi, en-tr). En en otras palabras, mientras que nunca se observó el inglés como un idioma objetivo durante la formación para ambas variantes del modelo, el entrenamiento de la segunda variante en- incluye incrustaciones de origen en inglés. Luego hicimos pruebas estas dos variantes de modelo en el corpus WMT 2018 para estos pares de idiomas y para el di-reverso direcciones (con la excepción de en-cs porque cs-en) no existe para WMT 2018). Los resultados en la Tabla 4 muestra claramente que para la clasificación de la traducción archi- la arquitectura, incluyendo la fuente, mejora el conjunto general correlación con los juicios humanos. Además, la inclusión de la fuente expuso la segunda vari- hormiga del modelo a incrustaciones en inglés que es

Tabla 4: Comparación entre C OMET - RANGO (sección 2.4) y una versión de solo referencia de la misma en datos de WMT18. Ambos modelos fueron entrenados con WMT17, lo que significa que el modelo solo de referencia nunca está expuesto al inglés. durante el entrenamiento.

Métrico	en-cs	en-de	en-fi	en-tr	cs-en	de-en	Por favor, proporciona el texto que necesitas traducir.	tr-en
C OMET - RANGO (ref. solo)	0.660	0.764	0.630	0.539	0.249	0.390	0.159	0.128
C OMET - RANGO	0.711	0.799	0.671	0.563	0.356	0.542	0.278	0.260
Δ τ	0.051	0.035	0.041	0.024	0.107	0.155	0.119	0.132

reflejado en un nivel superior Δ τ 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 requerido para ejecutar todas las líneas de base informadas.[6] Todo los modelos reportados en este documento fueron entrenados en un única GPU Tesla T4 (16GB). Además, nuestro marco-el trabajo se basa en PyTorch Lightning (Halcón , 2019), un envoltorio ligero de PyTorch, que fue creado para máxima flexibilidad y reproducibilidad.

7 Trabajo Relacionado

Las métricas clásicas de evaluación de MT suelen ser de carácter caracterizado como n -métricas de coincidencia de gramas porque, utilizando características hechas a mano, estiman la calidad de MT al contabilizar el número y la fracción de n -gramos que aparecen simultáneamente en un candidato hipótesis de traducción y una o más humanas referencias. Métricas como B LEU) Papineni y otros. , 2002), M ETEOR) Lavie y Denkowski , 2009), y CHR F (Popovi ć , 2015) han sido ampliamente estudiados-murió y mejoró Koehn y otros. , 2007 ; Popovi ć , 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 y otros. , 2014 ; Peters y otros. , 2018 ; Devlin y otros. , 2019 han surgido como un com-alternativa comúnmente utilizada a n -coincidencia de gramas capturando la similitud semántica de las palabras. Incrustación métricas basadas como M ETEOR -V ECTOR) Sirvan y otros. , 2016), B LEU 2 VEC) T Attar y Fishel , 2017), Y Ye S Ye -1 Lo , 2019), M SOBRE S NÚCLEO) Zhao y colaboradores. , 2019), y B ERTSCORE) Zhang y colaboradores. , 2020 crear 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 se-similitud semántica, resultando en un límite superior de correlación vinculado entre los juicios humanos y las puntuaciones producido por tales métricas.

Métricas aprendibles) Shimanaka y otros. , 2018 ; Mathur et al. , 2019 ; Shimanaka y colaboradores. , 2019) en intentar optimizar directamente la correlación con hu-juicios de hombre, y recientemente han mostrado promesa-resultados de la búsqueda. B LEURT) Sellam et al. , 2020), aprender métrica capaz basada en BERT (Devlin et al. , 2019), afirma un rendimiento de última generación durante los últimos 3 años de la tarea compartida de métricas WMT. Porque B LEURT se basa en English-BERT (Devlin y otros. , 2019), solo puede usarse cuando el inglés es el idioma objetivo que limita su aplicabilidad. Además, hasta donde sabemos, todos los previamente las métricas aprendibles propuestas se han centrado en la opti-minimizando DA que, debido a una escasez de anotadores, puede demostrar ser inherentemente ruidoso (Ma y otros. , 2019).

Evaluación MT sin referencias , también conocido como La Estimación de Calidad (EC), históricamente a menudo ha re-progresado en HTER para la evaluación a nivel de segmento Bo-jar y otros. , 2013 , 2014 , 2015 , 2016 , 2017a Más recientemente, MQM ha sido utilizado para el nivel de documento evaluación Specia y otros. , 2018 ; Fonseca et al. , 2019). Al aprovechar pre-entrenamientos altamente multilingües codificadores entrenados como BERT multilingüe (De Vlin y otros. , 2019) y XLM (Conneau y Lam por favor , 2019), los sistemas QE han estado mostrando aus-correlaciones sospechosas con juicios humanos. Ke-Pler et al. , 2019a). Simultáneamente, el OpenKiwi marco) Kepler y otros. , 2019b) ha facilitado para que los investigadores avancen en el campo y construyan modelos de QE más fuertes.

8 Conclusiones y Trabajo Futuro

En este documento presentamos C OMET , una nueva novela neu-Marco conceptual para la formación de modelos de evaluación de MT eso puede servir como métricas automáticas y ser fácilmente

[6] Estos se llevarán a cabo en: https://github.com/Unbabel/COMET

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

Para demostrar la efectividad de nuestro marco de trabajo, buscamos abordar los desafíos reportados en el Tarea Compartida de Métricas WMT 2019 Ma y otros. , 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 son el peso abrumador de paparámetros y tiempo de inferencia. Una vía principal para trabajo futuro en C OMET examinará el impacto de soluciones más compactas como DistilBERT (Sanh y otros. , 2019).

Además, mientras esbozamos el potencial imImportancia del texto fuente anterior, notamos que nuestro C OMET - RANGO el modelo pondera la fuente y la referencia diferentemente durante la inferencia pero igualmente en su entrenamiento función de pérdida de ing. El trabajo futuro investigará el optimalidad de esta formulación y examinar más a fondo la interdependencia de las diferentes entradas.

Reconocimientos

Estamos agradecidos con Andr é Martins, Austin Matthews, Fabio Kepler, Daan Van Stigt, Miguel Vera, y los revisores, por sus valiosos comentarios y desdiscusiones. 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 oraciones multilingües para cero transferencia cruzada de idiomas de un solo disparo y más allá . Transac- Asociación para la Lingüística Computacional tics , 7:597–610.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Grajamón, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post y Marcos Zampieri. 2019. Hallazgos de la conferencia 2019 sobre traducción automática (WMT19) . En Actas de la Cuarta Con- Conferencia sobre Traducción Automática (Volumen 2: Compartido Papeles de Tarea, Día 1) , páginas 1–61, Florencia, Italia. As- Asociación para la Lingüística Computacional.

Ond˘rej 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 de 2013 tienda sobre Traducción Automática Estadística . En Proceder Actas del Octavo Taller sobre Máquina Estadística Traducción , páginas 1-44, Sofía, Bulgaria. Asocia- 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 de la Noveno Taller de Traducción Automática Estadística , páginas 12-58, Baltimore, Maryland, EE. UU. Asocia- 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 máquinas traducción (WMT17) . En Actas de la Sec- Segunda Conferencia sobre Traducción Automática , páginas 169–214, Copenhague, Dinamarca. Asociación para Com- Lingüística Computacional.

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 2016 sobre la traducción automática . En Actas de la Primera Conferencia sobre Traducción Automática: Volumen 2, Documentos de Tareas Compartidas , páginas 131-198, Berlín, Alemuchos. Asociación para la Lingüística Computacional.

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

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

Aljoscha Burchardt y Arle Lommel. 2014. Practi Directrices de Uso de MQM en Investigaciones Científicas 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 Zettle Moyer y Veselin Stoyanov. 2019. Sin supervisión aprendizaje de representación interlingüística a gran escala · arXiv preprint arXiv:1911.02116 ·

Alexis Conneau y Guillaume Lample. 2019. Cruzar preentrenamiento del modelo de lenguaje lingual · En H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch´e Buc, E. Fox y R. Garnett, editores, Avances en Neu- Sistemas de Procesamiento de Información General 32 , páginas 7059–7069. Curran Associates, Inc.

Michael Denkowski y Alon Lavie. 2011. Meteor 1.3: Métrica automática para optimización confiable y evaluación Evaluación de sistemas de traducción automática . En Proceder Actas del Sexto Taller sobre Máquina Estadística Traducción , páginas 85-91, Edimburgo, Escocia. As- 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 debajo de de pie · En Actas de la Conferencia 2019 del Capítulo Norteamericano de la Asociación para Lingüística Computacional: Lenguaje Humano Tecnologías, Volumen 1 (Documentos Largos y Cortos) , páginas 4171-4186, Minneapolis, Minnesota. Asoci- Asociación para la Lingüística Computacional.

WA Falcon. 2019. PyTorch Lightning: El liviano Envoltorio de PyTorch para investigación de IA de alto rendimiento · GitHub ·

Erick Fonseca, Lisa Yankovskaya, Andr´e F. T. Martins, Mark Fishel y Christian Federmann. 2019. Encontrar resultados de las tareas compartidas de WMT 2019 sobre la estimación de calidad información . En Actas de la Cuarta Conferencia sobre Traducción Automática (Volumen 3: Documentos de Tareas Compartidas, 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 Pro Actas del 7mo Taller de Anotación Lingüística e Interoperabilidad con Discourse , 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? ¿mejor con el tiempo? En Actas de la 14ª Conferencia 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. ¿Pueden los sistemas de traducción automática- los elementos deben ser evaluados solo por la multitud · Lan Natural Ingeniería de Idiomas , 23(1):330.

Yinuo Guo y Junfeng Hu. 2019. Meteor++ 2.0: Adopta el conocimiento de paráfrasis a nivel sintáctico en ma- evaluación de traducción china . En Actas de la Cuarta Conferencia sobre Traducción Automática (Volumen 2: Documentos de Tarea Compartida, Día 1) , páginas 501-506, Flo- Florencia, Italia. Asociación para la Lingüística Computacional. tics.

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

Fabio Kepler, Jonay Tr´enous, Marcos Treviso, Miguel Vera, Antônio Góis, M. Amin Farajian, Antônio V. Lopes, y Andr´e F. T. Martins. 2019a. Unba- participación de Bel en la calificación de traducción WMT19 tarea compartida de estimación de entidad . En Actas de la Cuarta Conferencia sobre Traducción Automática (Volumen 3: Documentos de Tarea Compartida, Día 2) , páginas 78-84, Flo- Florencia, Italia. Asociación para la Lingüística Computacional. tics.

Fabio Kepler, Jonay Tr´enous, 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 Manifestaciones , páginas 117-122, Florencia, Italia. As- 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, Ond˘rej Bojar, Alexandra Constantin y Evan Herbst. 2007. Moisés: Abre kit de herramientas de origen para la traducción automática estadística . En Actas de la 45ª Reunión Anual de la As- Compañero de la Asociación para la Lingüística Computacional Actas del Volumen de la Sesión de Demostraciones y Pósters siones , páginas 177-180, Praga, República Checa. Como- Asociación para la Lingüística Computacional.

Dan Kondratyuk y Milan Straka. 2019. 75 lan- Idiomas, 1 modelo: Analizando dependencias universales universalmente · En Actas de la Conferencia 2019 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. Como Asociación para la Lingüística Computacional.

Alon Lavie y Michael Denkowski. 2009. El meteoro métrica para la evaluación automática de la traducción de máquinas traducción · Traducción Automática , 23:105–115.

Chi-kiu Lo. 2019. YiSi - una calidad unificada de traducción semántica MT métrica de evaluación y estimación para idiomas con diferentes niveles de recursos disponibles . En Proceder Actas de la Cuarta Conferencia sobre Traducción Automática (Volumen 2: Documentos de Tareas Compartidas, Día 1) , páginas 507-513, Florencia, Italia. Asociación para la Computación. Lingüística Computacional.

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

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

Qingsong Ma, Ond ˇrej Bojar, y Yvette Graham. 2018. Resultados de la tarea compartida de métricas WMT18: Ambos personajes y incrustaciones logran un buen rendimiento romance . En Actas de la Tercera Conferencia sobre Traducción Automática: Documentos de Tareas Compartidas , páginas 671-688, Bélgica, Bruselas. Asociación para Com- Lingüística Computacional.

Qingsong Ma, Johnny Wei, Ond ˇrej Bojar, y Yvette Graham. 2019. Resultados de las métricas WMT19 tarea compartida: Sistemas de MT fuertes y a nivel de segmento los temas representan grandes desafíos · En Actas de la Cuarta Conferencia sobre Traducción Automática (Volumen 2: Documentos de Tarea Compartida, Día 1) , páginas 62-90, Flo- Florencia, Italia. Asociación para la Lingüística Computacional. tics.

Nitika Mathur, Timothy Baldwin y Trevor Cohn. 2019. Poniendo la evaluación en contexto: Contextual las 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 Compu- Lingüística Computacional.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Cor- Rado y Jeff Dean. 2013. Representación distribuida traducciones de palabras y frases y su composición- ciudad . En Avances en el Procesamiento de Información Neural Sistemas 26 , páginas 3111-3119. Asociados Curran, S.A. (Sociedad Anónima)

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 la Comunicación Lingüística 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 presentación . En Actas de la Conferencia 2014 sobre Métodos Empíricos en el Procesamiento del Lenguaje Natural (EMNLP) , páginas 1532-1543, Doha, Qatar. Asso- Asociación para la Lingüística Computacional.

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

Telmo Pires, Eva Schlinger y Dan Garrette. 2019. ¿Cuán multilingüe es BERT multilingüe? En Pro Actas de la 57ª Reunión Anual de la Asociación Asociación para la Lingüística Computacional , páginas 4996–5001, Florencia, Italia. Asociación para la Computación. Lingüística Computacional.

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

Maja Popovi´c. 2017. chrF++: palabras que ayudan a charac- tener n-gramas · En Actas del Segundo Congreso Conferencia sobre Traducción Automática , páginas 612-618, Copenhague, Dinamarca. Asociación para la Computación. Lingüística Computacional.

Nils Reimers e Iryna Gurevych. 2019. Oración 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 Natu- Procesamiento del Lenguaje Natural (EMNLP-IJCNLP) , páginas 3982-3992, Hong Kong, China. Asociación para Lingüística Computacional.

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 · arXiv preprint arXiv:1910.01108 .

F. Schroff, D. Kalenichenko, y J. Philbin. 2015. Facenet: Una incrustación unificada para el reconocimiento facial y agrupación . En Conferencia IEEE 2015 sobre Com- Visión por Computadora 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 Computa- Lingüística Computacional.

Christophe Servan, Alexandre B´erard, Zied Elloumi, Hervé Blanchon y Laurent Besacier. 2016. Word2Vec vs DBnary: Mejorando METEOR us- ¿Representaciones vectoriales o recursos léxicos? En Actas de COLING 2016, la 26ª Interna- 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 de máquinas situación . En Actas de la Tercera Conferencia sobre Traducción Automática: Documentos de Tareas Compartidas , páginas 751-758, Bélgica, Bruselas. Asociación para Com- Lingüística Computacional.

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

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

Lucia Specia, Frédéric 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: Documentos de Tareas Compartidas , páginas 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. Traduce Calidad de la traducción y productividad: Un estudio sobre rich mor- idiomas de morfología . En Cumbre de Traducción Automática XVI , páginas 55-71, Nagoya, Japón.

Kosuke Takahashi, Katsuhito Sudoh y Satoshi Naka mura. 2020. Evaluación automática de la traducción de máquinas utilizando entradas en el idioma de origen y translingüístico modelo de lenguaje . 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 Tättar y Mark Fishel. 2017. bleu2vec: el métrica dolorosamente familiar en espacio vectorial continuo esteroides . En Actas de la Segunda Conferencia sobre Traducción Automática , páginas 619-622, Copen- Hagen, Dinamarca. Asociación para la Computación Lingüística.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, y Yoav Artzi. 2020. Bertscore: Eval- Evaluando la generación de texto con Bert . En Internacional Conferencia sobre Representaciones de Aprendizaje .

Wei Zhao, Goran Glava˘s, Maxime Peyrard, Yang Gao, Robert West y Steffen Eger. 2020. En el límite limitaciones de los codificadores multilingües según lo expuesto por evaluación de traducción automática sin referencia . En Actas de la 58ª Reunión Anual de la Asociación Asociación para la Lingüística Computacional , páginas 1656–1671, En línea. Asociación para la Lingüística Computacional. lingüística.

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

A Apéndices

En la Mesa 5 enumeramos los hiperparámetros utilizados para entrenar nuestros modelos. Antes de inicializar estos modelos, se realizó una la semilla del DOM se estableció en 3 en todas las bibliotecas que realizan operaciones "aleatorias" (antorcha , numpy , aleatorio y Cuda).

Tabla 5: Hiperparámetros utilizados en nuestro C OMET marco de trabajo para entrenar los modelos presentados.

Hiper-parámetro	C OMET (Est-HTER/MQM)	C OMET - RANGO
Modelo de Codificación	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 de lote	dieciséis	dieciséis
Función de pérdida	MSE	Margen de Trío $\epsilon = 1.0$
Abandono por capas	0.1	0,1
Precisión FP	32	32
Unidades ocultas de avance directo	2304,1152	–
Activaciones Feed-Forward	Tanh	–
Abandono hacia adelante	0,1	–

Tabla 6: Estadísticas para el corpus QT21.

	en-de	en-cs	en-lv	de-en
Tuplas totales	54000	cuarenta y dos mil	35474	41998
Avg. tokens (referencia)	17.80	15.56	16.42	17.71
Avg. tokens (fuente)	16.70	17.37	18.39	17.18
Avg. tokens (MT)	17.65	15.64	16.42	17.78

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

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

Pares de idiomas RR al inglés.

DA

Tabla 9: Estadísticas para el WMT 2019

Por favor, proporciona el texto que necesitas traducir.

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

RR de pares de idiomas inglés y no inglés.

DA

Tabla 9: Estadísticas para el WMT 2019

	Tuplas totales	Avg. tokens (referencia)	Avg. tokens (fuente)	Avg. tokens (MT)
de-fr	4862	27.32	21.36	25.68
de-cs	23194	22.27	25.22	21.89
fr-de	1369	22.68	28.60	23.36
en-zh	18658	9.25	24.39	6.83
en-ru	24334	24.79	24.45	23.37
en-lt	17401	21.00	24.46	20.97
en-kk	18172	18.89	23.78	19.92
en-gu	11355	33.32	24.32	32.97
en-fi	31820	20.12	25.23	19.69
en-de	99840	25.65	24.97	24.98
en-cs	27178	22.92	24.98	22.60

estadísticas.

3.3

Tabla 10: Corpus MQM (sección

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

Pares de idiomas RR.

DA

Tabla 11: Estadísticas para el WMT 2018

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

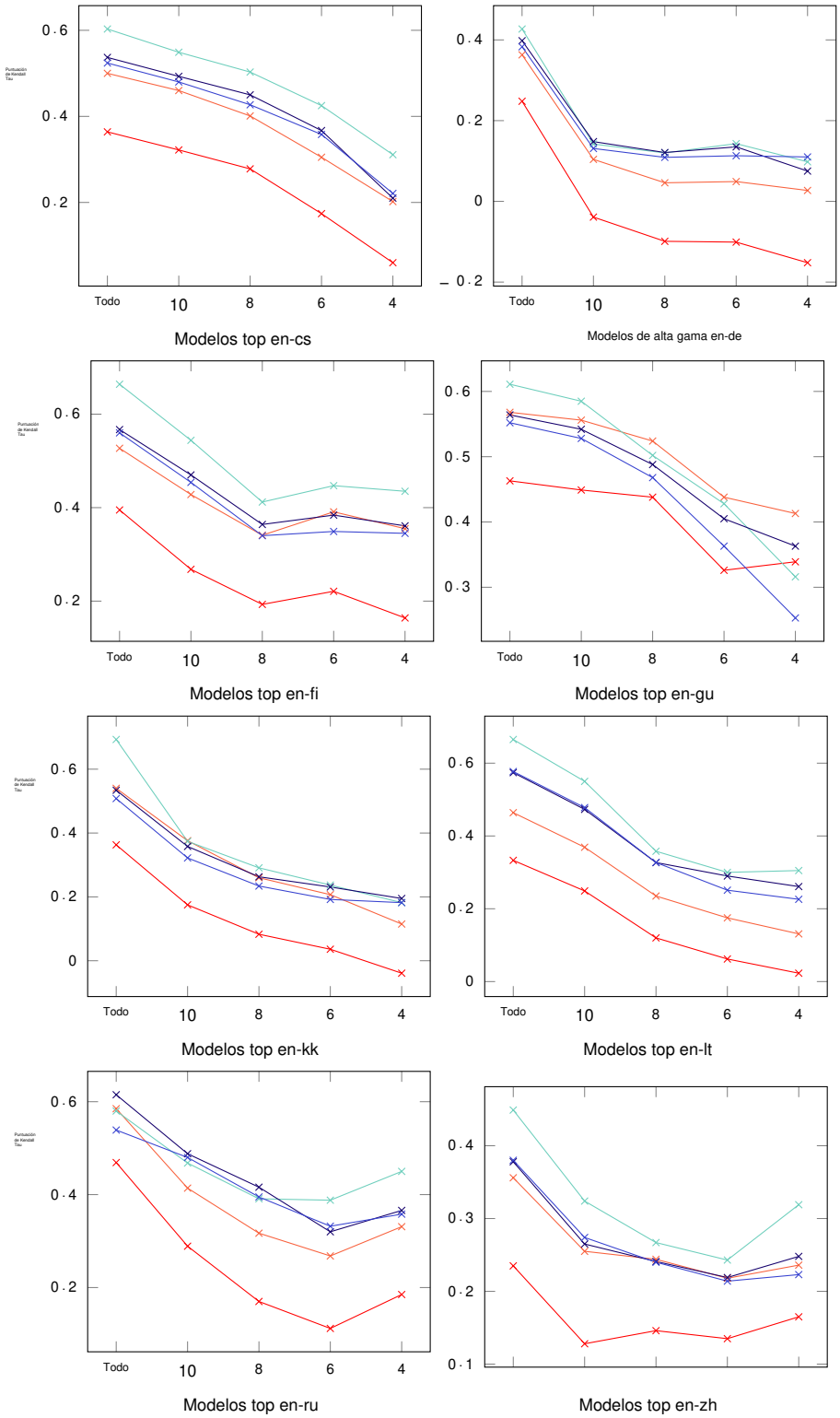

Tabla 12: Rendimiento de las métricas en todos y los mejores (10, 8, 6 y 4) sistemas de MT para todos los idiomas desde el inglés. pares. El esquema de colores es el siguiente: — C OMET - RANGO, — C OMET - HTER, — C OMET - MQM, — B LEU, — B ERTSCORE

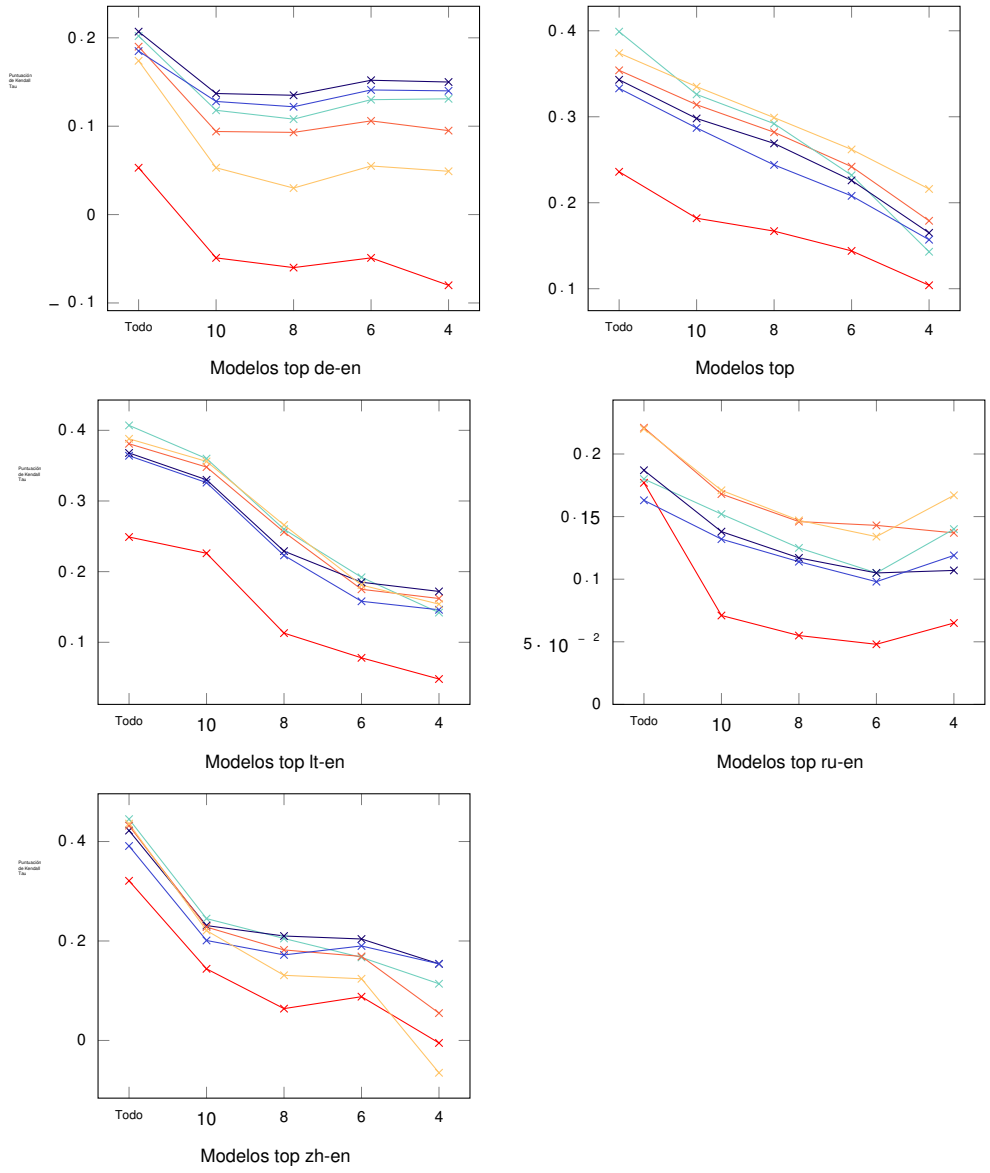

Tabla 13: Rendimiento de métricas en todos y los mejores (10, 8, 6 y 4) sistemas de MT para todos los idiomas hacia el inglés. pares. El esquema de colores es el siguiente: C OMET - RANGO , C OMET - HTER , C OMET - MQM , B LEU , B ERTSCORE , B LEURT