UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Estimación automática de la calidad de la traducción automática

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Lena Almor, Iván

Tutor/a: Casacuberta Nolla, Francisco

Cotutor/a externo: GARCIA MARTINEZ, MARIA MERCEDES

CURSO ACADÉMICO: 2021/2022

Resum

En els darrers anys, la intel·ligència artificial (IA) està experimentant un creixement exponencial. L'apogeu del Big Data juntament amb la popularització de l'aprenentatge profund basat en xarxes neuronals han estat els elements detonants d'aquesta revolució tecnològica que cada dia avança a passos més de gegant. Un dels camps principals de recerca dins de la IA és el Processament del Llenguatge Natural (PLN), ja que poder entendre i manipular el llenguatge al nostre gust és un repte diferencial per a la nostra espècie. En específic, la branca del PLN que es tractarà en aquest projecte és la Traducció Automàtica (TA).

Actualment, la traducció automàtica està assolint una qualitat similar a la d'un traductor humà i s'han registrat resultats excel·lents per a molts parells de llengües. Tot i així, en molts casos no sempre és perfecta i cal la revisió de traductors humans professionals. Aquest procés de mesura de qualitat és tediós per als traductors i té un gran cost temporal i econòmic. L'objectiu d'aquest treball és automatitzar aquest procés estalviant molts costos.

Es preten entrenar un model que estime la qualitat d'una traducció sense fer servir una frase de referència. Per fer-ho, caldrà entrenar un model que aprenga a partir d'un conjunt de frases associades a una mètrica de qualitat. En aquest cas, la mètrica que es vol predir representa l'esforç de posedició necessari perquè la frase traduïda siga correcta. Concretament, la mètrica utilitzada és TER (per les sigles en anglès "Translation Error Rate"que significa ràtio d'error en la traducció). Les arquitectures dels models es basaran en models del llenguatge d'aprenentatge profund preentrenats amb moltes dades.

Paraules clau: Traducció Automàtica, Estimació de qualitat de la traducció automàtica, Aprenentatge Automàtic, Aprenentatge Profund, Xarxes Neuronals

Resumen

En los últimos años, la inteligencia artificial (IA) está experimentando un crecimiento exponencial. El auge del Big Data junto a la popularización del Aprendizaje Profundo basado en redes neuronales han sido los elementos detonantes de esta revolución tecnológica que cada día avanza a pasos más agigantados. Uno de los campos principales de investigación dentro de la IA es el Procesado del Lenguaje Natural (PLN), ya que poder entender y manipular el lenguaje a nuestro antojo es un reto diferencial para nuestra especie. En específico, la rama del PLN que se tratará en este proyecto es la Traducción Automática (TA).

Actualmente, la traducción automática está alcanzando una calidad similar a la de un traductor humano y se han registrado excelentes resultados para muchos pares de lenguas. Aún así, en muchos casos no siempre es perfecta y se necesita la revisión de traductores humanos profesionales. Este proceso de medición de calidad es tedioso para los traductores y tiene un gran coste temporal y económico. El objetivo de este trabajo es automatizar este proceso ahorrando muchos de estos costes.

Se pretende entrenar un modelo que estime la calidad de una traducción sin utilizar una frase de referencia. Para ello, habrá que entrenar un modelo que aprenda a partir de un conjunto de frases asociadas a una métrica de calidad. En este caso, la métrica que se quiere predecir representa el esfuerzo de posedición necesario para que la frase traducida sea correcta. Concretamente, la métrica utilizada es TER (por sus siglas en inglés "Translation Error Rate"que significa ratio de error en la traducción). Las arquitecturas de los modelos se basarán en modelos del lenguaje de aprendizaje profundo preentrenados con muchos datos.

Palabras clave: Traducción Automática, Estimación de calidad de la traducción automática, Aprendizaje Automático, Aprendizaje Profundo, Redes Neuronales

Abstract

In recent years, artificial intelligence (AI) is experiencing exponential growth. The rise of Big Data together with the popularisation of Deep Learning based on neural networks have been the triggering elements of this technological revolution that is advancing by leaps and bounds every day. One of the main fields of research within AI is Natural Language Processing (NLP), since being able to understand and manipulate language at our whim is a differential challenge for our species. Specifically, the branch of NLP that will be addressed in this project is Machine Translation (MT).

Currently, machine translation is reaching a quality similar to that of a human translator and excellent results have been recorded for many language pairs. Still, in many cases it is not always perfect and needs to be reviewed by professional human translators. This quality measurement process is tedious for translators and has a high time and financial cost. The aim of this work is to automate this process and save many of these costs.

The aim is to train a model that estimates the quality of a translation without using a reference sentence. To do this, we will have to train a model that learns from a set of sentences associated with a quality metric. In this case, the metric to be predicted represents the post-editing effort required for the translated sentence to be correct. Specifically, the metric used is TER (Translation Error Rate). The model architectures will be based on data-intensive, pre-trained deep learning language models.

Translated with www.DeepL.com/Translator (free version)

Key words: Machine Translation, Machine Translation Quality Estimation, Machine Learning, Deep Learning, Neural Networks

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

En el siguiente capítulo se tratará de explicar las partes introductorias del trabajo. Las cuales están formadas por: la motivación con la que se ha realizado el proyecto, los objetivos que se quieren cumplir y la estructura general del trabajo.

1.1 Motivación

El interés de evaluar traducciones surge a raíz del gran aumento de actividades de traducción que ha ido rodeándonos en nuestro día a día con el paso de los últimos años. Este aumento ha sido motivado por diversas causas, entre las que se encuentran los avances tecnológicos y científicos, el incremento de las relaciones internacionales, el crecimiento del sector turístico, el aumento de los nuevos medios de difusión de la información, la mayor comunicación entre diferentes comunidades lingüísticas, etc. Tradicionalmente, todas estas traducciones habían sido realizadas y evaluadas por los mismos humanos, pero con la llegada de la traducción automática esto cambió. Estas nuevas herramientas han conseguido un gran aumento en velocidad y eficiencia de traducción pero en muchas ocasiones se carece de carácter humano y las traducciones pueden parecer algo forzadas, o simplemente estar mal. El aumento de revisión de la calidad de las traducciones se ha visto multiplicado y con ello surge la necesidad de automatizar también este proceso. Esta evaluación automática pretende mostrar de manera fácil y rápida si una traducción automática es adecuada para un proyecto y la cantidad de edición que se necesita para corregirla.

1.2 Objetivos

El objetivo principal de este trabajo consiste en el estudio y la experimentación de un modelo de estimación de calidad de la traducción automática aplicado a dos idiomas: francés y español. Para ello se ha entrenado con un conjunto de datos formado por tres atributos: una frase de un idioma origen, una traducción de la frase origen generada automáticamente y un valor de una métrica de calidad. El trabajo se divide en los siguientes objetivos:

- Obtener las métricas de calidad estimadas automáticamente por el modelo.
- Observar si existe correlación entre la calidad predicha por el modelo y las frases de test.
- Comparar la precisión entre los modelos monolingües y el multilingüe.

- Estudiar un posible umbral de esfuerzo de posedición para decidir si una frase es buena o mala.

1.3 Estructura de la memoria

Este proyecto está compuesto por seis capítulos. A continuación, se detallan los temas a tratar en los diferentes capítulos:

En el **capítulo 1**, se introduce el tema que se aborda a lo largo del proyecto mediante tres breves partes compuestas por la motivación del proyecto, los objetivos y la presente estructura del trabajo de final de grado.

En el **capítulo 2**, se muestran los avances históricos que se han producido hasta llegar a las tecnologías empleadas en el proyecto.

En el **capítulo 3**, se presenta al lector las distintas técnicas de estimación de calidad automática de la traducción automática que se han empleado en el presente proyecto, así como su arquitectura y funcionamiento.

En el **capítulo 4**, se presenta el marco experimental del trabajo, en este capítulo se presentarán las tecnologías que han sido utilizadas, la preparación y procesado de los diferentes corpus, la partición de los distintos conjuntos de datos generados, los experimentos que se han realizado y las distintas formas de evaluación de los modelos desarrollados.

En el **capítulo 5**, se realizará un análisis aplicado a una serie de experimentos que simulan un entorno de traducción. Luego se complementará con un análisis cualitativo y cuantitativo de los diferentes modelos de evaluación automática

En el **capítulo 6**, se explicarán las conclusiones obtenidas y relacionadas con los objetivos presentados, además, se expondrán diversas líneas de investigación que se podrían tener en cuenta en próximos trabajos relacionados con la evaluación automática de la traducción automática.

CAPÍTULO 2
Estado del arte

En el actual capítulo, se presenta el contexto tecnológico del presente trabajo, exponiendo así las diferentes tecnologías más relevantes que han hecho posible llegar hasta el punto en el que se encuentra actualmente la estimación automática de calidad.

2.1 Traducción automática

Antes de que apareciera la traducción automática neuronal hubo dos grandes familias de sistemas de traducción automática: los basados en reglas y los estadísticos.

Los sistemas basados en reglas [1] son aquellos que siguen una lista de reglas y la aplican de forma secuencial a los textos de entrada. De manera que se va transformando la entrada regla a regla hasta traducir por completo la frase. Estas reglas están descritas por lingüistas expertos que conocen la sintaxis tanto del idioma de entrada cómo el de salida. Por ejemplo, una regla del castellano al inglés podría ser invertir un nombre y un adjetivo cuando aparecen juntos en la frase de origen:

El gato negro -> El negro gato

Luego al traducir palabra por palabra quedaría:

El gato negro -> The black cat

Este método generará frases de gran calidad porque se basa en reglas que capturan un conocimiento a alto nivel y en profundidad de cada idioma. Se ha visto que estos sistemas obtienen buenos resultados para idiomas con muchas casuísticas, por ejemplo el turco o el euskera que tienen un gran abanico morfológico. Otra aplicación de los SBR es en idiomas con pocos recursos donde no se dispone de grandes cantidades de datos. El principal problema de estos sistemas es la poca flexibilidad que tienen, ya que se basan únicamente en un conjunto de reglas y cuando se quiere traducir muchas frases empiezan a salir casos que no se han tenido en cuenta previamente y, por ende, se obtienen malas traducciones.

Para corregir el problema de la flexibilidad surgió un nuevo paradigma basado en los sistemas de traducción automática estadística.[6] Estos sistemas están formados por modelos probabilísticos que se aplican a datos paralelos, es decir, datos que contienen frases y sus correspondientes traducciones. La función de estos modelos es calcular la probabilidad de cada posible traducción para una frase de entrada y elegir aquella que la maximiza. Si se formaliza esta descripción se obtendría la siguiente ecuación [3]:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) \tag{2.1}$$

No obstante, en la práctica estos modelos suelen ser combinados con el modelo log-linear para el término $p(\mathbf{y}|\mathbf{x})$, por lo que el problema se modela de la siguiente manera [16]:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \left\{ \sum_{\mathrm{n}=1}^{N} \lambda_n \log\left(h_n\left(\mathbf{x}, \mathbf{y}\right)\right) \right\} \tag{2.2}$$

en el que $h_n\left(\mathbf{x}, \mathbf{y}\right)$ se corresponde a la n-ésima función perteneciente a la lista de funciones asociadas a las características que valoran si una traducción es buena o no. Luego, lambda, representa el n-ésimo peso asociado a la combinación log-lineal de la función correspondiente y, por último, N es la longitud de la lista de características.

2.2 Traducción Automática Neuronal

Con la llegada del aprendizaje profundo, junto al poder de almacenamiento de datos masivos, se empezaron a desarrollar sistemas de traducción basados en redes neuronales con arquitecturas profundas. Al poco tiempo, se obtuvieron resultados que sobrepasaron a los sistemas estadísticos en varios pares de lenguas, hecho que impulsó a una gran cantidad de investigadores a profundizar en este enfoque y a empezar a destinar gran parte de sus recursos. Esto ha hecho que la traducción automática neuronal sea el actual paradigma de traducción.

La TAN también sigue un enfoque probabilístico cómo su paradigma predecesor, es decir, tiene cómo objetivo estimar una distribución condicional denominada cómo $P(\mathbf{y}|\mathbf{x})$ dado un conjunto de datos D, en el cual x e y son las variables aleatorias que representan la frase de entrada y salida respectivamente. Cabe destacar que existen varios enfoques a la hora de modelar el problema: documento, párrafo y frase.

Empezando de mayor granularidad a menor, observamos la traducción a nivel de frase, que viene descrita por una frase de entrada $\mathbf{x} = (x_1, \ldots, x_S)$, donde S se corresponde a la longitud de la oración, y una frase de salida $\mathbf{y} = (y_1, \ldots, y_T)$ donde T viene dada por la longitud de esta. Formalizando el problema, la distribución condicional se puede describir cómo:

$$\hat{\mathbf{y}}_1^{\hat{T}} = \arg\max_{T, \mathbf{y}_1^T} \prod_{t=1}^{T} Pr_\theta(y_t | y_1^{t-1}, c(x_1^S)) \tag{2.3}$$

donde y_t representa la palabra de salida actual, que viene dada a partir de las palabras anteriores y_1^{t-1} que han sido traducidas previamente con un tipo de representación denotado por la función c de la frase de entrada x_1^S, y empleando los parámetros θ de la función Pr.

Actualmente, existen una gran cantidad de modelos que presentan un codificador y un decodificador, muchos de ellos se combina con redes neuronales profundas[4].

2.2.1. Arquitectura Codificador-Decodificador

La arquitectura codificador-decodificador está dividida en dos partes cómo bien indica su nombre. La primera es el codificador, que tiene cómo objetivo transformar la frase de entrada en una representación vectorial de longitud fija. Luego está el decodificador, que hará el proceso inverso para devolver la frase traducida al idioma de salida. Estas dos componentes están formadas por redes neuronales recurrentes, que son aquellas que se

retroalimentan y tienen "memoria" que tratan de aprender los parámetros que maximizan la probabilidad condicional de cada palabra dada una frase de entrada x y fijándose también en las palabras ya traducidas anteriormente $y_{<t}$. Si formalizamos esta última oración obtendremos lo siguiente:

$$\arg\max \sum_{t=1}^{T} \log p\left(y_t | y_{<t}, \mathbf{x}\right) \tag{2.4}$$

También cabe recalcar que la arquitectura codificador-decodificador está compuesta por cuatro elementos: la capa de *embedding*, la de codificación, la de decodificación y finalmente, la de clasificación.

La capa de *embedding* tiene cómo objetivo fundamental transformar los vectores obtenidos a partir de la frase de entrada, en vectores compuestos por números continuos.

Antes de que apareciera la arquitectura *Transformer*, el tipo de codificador más empleado ha sido la red neuronal recurrente(RNN). En particular, las RNN de tipo LSTM o GRU, ya que permiten modelar una secuencia de caracteres teniendo en cuenta la temporalidad de la frase. Esta red codificadora es la encargada de transformar los embeddings, obtenidos por la anterior capa, en una representación de la frase de entrada. Para poder modelar la frase, la RNN va procesando cada elemento de una secuencia de entrada x donde su representación se define en la ecuación 2.5, vendrá dada por el nombre de estado oculto h_t [29].

$$h_t = COD(x_t, h_{t-1}) \tag{2.5}$$

$$\mathbf{c} = q(h_1, \ldots, h_S) \tag{2.6}$$

donde x_t es una secuencia de entrada al codificador, $h_t \in \mathbb{R}^t$ es el estado oculto en el instante t.

La ecuación 2.6 muestra cómo el vector de contexto c se genera transformando no linealmente los estados ocultos h_t a partir de la función q.

Respecto al decodificador, su objetivo es generar el texto de salida $\mathbf{y} = (y_1, \ldots, y_t)$ a partir de las representaciones obtenidas por el codificador. Dada una representación donde el inicio de palabra esté representado por el token $y_0 = < bos >$ y el estado inicial por $s_0 = h_S$, el decodificador guarda la frase decodificada en un vector de estados $\mathbf{s}_t \in \mathbb{R}^d$ que viene formalizado por la siguiente ecuación:

$$\mathbf{s}_t = DEC(y_{t-1}, \mathbf{s}_{t-1}, \mathbf{c}) \tag{2.7}$$

Finalmente, la capa de clasificación se encarga de predecir la distribución de los tokens de salida. Normalmente, viene dada por una función de activación *softmax*, la cual tiene cómo salida valores reales entre 0 y 1. Teniendo en cuenta que el vocabulario del idioma de salida es V, y |V| es el tamaño del vocabulario, dada la salida del decodificador $\mathbf{s}_t \in \mathbb{R}^d$, la capa de clasificación mapea $\mathbf{h}$ a un vector $\mathbf{z}$ en el espacio del vocabulario $\mathbb{R}^{|V|}$. Luego, se aplica la función *softmax* para asegurar que el vector es una probabilidad válida:

$$softmax\left(z_i\right) = \frac{\exp(z_i)}{\sum_{k=1}^{|V|} \exp(z_k)} \tag{2.8}$$

donde se utiliza z_i para representar el i-ésimo componente del vector $\mathbf{z}$.

2.2.2. Transformer

En 2017 apareció la arquitectura *Transformer* [30], siendo esta la más utilizada actualmente. Esta arquitectura sustituye las redes recurrentes, anteriormente utilizadas en el codificador y decodificador, por redes feed-forward en conjunto con los mecanismos de atención. Estos mecanismos de atención consiguen aprender las dependencias entre la frase de origen y destino sin tener en cuenta la distancia entre palabras. Esta característica indica una mejora respecto a las redes recurrentes anteriormente empleadas, ya que estas tendían a olvidar las dependencias entre palabras alejadas dentro de una frase. El otro factor diferencial de la arquitectura Transformer fue la codificación posicional, ya que permite ubicar las palabras dentro de la frase sin una recurrencia costosa que, además, impide paralelizar el entrenamiento.

El codificador en la arquitectura *Transformer* viene dado por 6 bloques idénticos formados por dos capas cada uno. La primera capa es un mecanismo multicabezal de autoatención, y la segunda una red neuronal prealimentada o más comúnmente nombrada cómo red feed-forward. Además, se emplea una conexión residual sobre cada una de las dos capas, seguida de una capa de normalización.

El decodificador también viene dado por otros 6 bloques, semejantes a los del codificador pero difiriendo en un par de aspectos. En este caso se añade una primera capa de atención con enmascaramiento, que combinado con el hecho de que las salidas de los *embeddings* están codificadas posicionalmente, fuerza a que las predicciones dependan solamente de las salidas para posiciones menores que i.

Cabe resaltar que en el modelo Transformer se aplica el mecanismo de autoatención sobre tres matrices: Q, matriz que encapsula un conjunto de consultas simultáneamente, K, matriz que encapsula un conjunto de llaves, y finalmente V, matriz que contiene un conjunto de valores asociados a las llaves. Este mecanismo de atención se aplica sobre las matrices en tres partes distintas de la arquitectura. En primer lugar, en el codificador, donde las matrices tienen el mismo valor que se corresponde con la frase de origen. En segundo lugar, se aplica en el decodificador, donde los valores de las matrices se corresponden con la traducción objetivo. Finalmente, en cuanto a los estados de los mecanismos de atención que conectan el codificador y el decodificador, el vector Q representa la traducción final mientras que los vectores K y V representan la frase de salida.

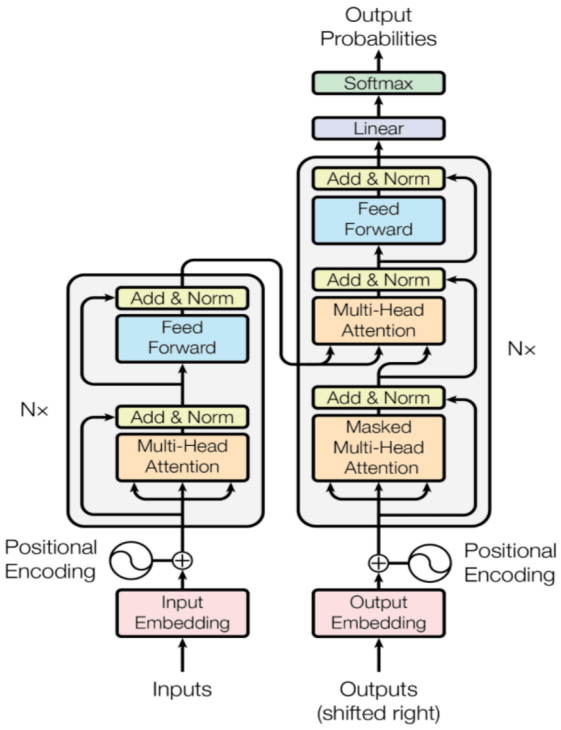

Figura 2.1: Arquitectura Transformer

2.3 Modelos de lenguaje basados en Transformer

Tras la aparición del Transformer, y de las notables mejoras que introdujo respecto a las RNN utilizadas anteriormente, se popularizó su uso y se empezaron a investigar posibles modificaciones para adaptar la arquitectura a las diferentes tareas del PLN.

A continuación se comentarán los modelos más relevantes que están relacionados con el XLM-R, el cual se ha utilizado en el presente trabajo.

2.3.1. BERT

BERT [7], por sus siglas Bidirectional Encoder Representations from Transformers, es un modelo basado en el codificador de la arquitectura Transformer que fue presentado en 2019 para mejorar las representaciones de lenguaje pre-entrenedas de los ya existentes modelos ELMo [18] y GPT [21]. La arquitectura de BERT es la misma que la del Transformer original pero implementando únicamente la parte del codificador y concatenando varios de estos codificadores. BERT ha sido pre-entrenado para dos tareas no supervisadas: Masked Language Model(MLM) y Next Sentence Prediction(NSP).

MLM

Esta tarea tiene cómo fin darle bidireccionalidad al modelo, aspecto que carecía en los ya nombrados ElMo y GPT, para así poder diferenciar palabras similares en diferentes contextos. Un ejemplo serían las frases "Juan es una estrella del rock" y "Han encontrado una nueva estrella a 200.000 km de la Tierra". En este caso BERT obtendría una representación diferente de la palabra estrella para las distintas frases. Para conseguir esta bidireccionalidad, se selecciona un 15 % de los tokens de una frase de manera aleatoria, los cuales se intercambian por la etiqueta [MASK] con el fin de predecir las partes

enmascaradas. Para ello, tendrán que utilizarse los tokens que quedan visibles, es decir, los que no han sido enmascarados y, por tanto, el modelo estará condicionado por el contexto anterior y posterior para lograr el desenmascaramiento de un token concreto. Sin embargo, esto provoca un desajuste entre el pre-entrenamiento y la especialización en una tarea concreta debido a que la etiqueta de enmascaramiento [MASK] no aparece durante el proceso de especialización. Para evitarlo, del 15 % de los tokens que han sido enmascarados, se enmascaran únicamente el 80 %, el otro 10 % son sustituidos por una palabra aleatoria y el 10 % restante no se enmascaran.

NSP

La siguiente tarea tiene cómo objetivo entender las relaciones entre dos frases, aspecto que no se capturaba en el modelo del lenguaje. Para ello, se utiliza un corpus monolingüe donde cada frase de entrenamiento X tiene asociada una frase Y con una etiqueta binaria que indica si Y es la siguiente frase de X. Concretamente, un 50 % de los datos están clasificados cómo que Y es la siguiente frase de X y el otro 50 % son frases asociadas aleatoriamente.

Para representar un par de frases X e Y en el entrenamiento se utilizan dos etiquetas especiales. La primera es *[CLS]* que representa el principio de la frase X y su clasificación, y la segunda, *[SEP]*, la separación entre la frase X e Y, el cual es a su vez el último token de X. De este modo, las dos frases de entrenamiento quedarían representadas como sigue: *[CLS] A [SEP] B*.

2.3.2. RoBERTa

RoBERTa [31] es un modelo que nació tras estudiar con detalle el pre-entrenamiento de BERT, centrándose en cómo afectan el ajuste de hiperparámetros y el tamaño del conjunto de entrenamiento. Para ello se hacen cuatro modificaciones clave al pre-entrenamiento:

- Aumentar el tiempo de entrenamiento, con batches mayores y más datos.
- Eliminar la tarea que consiste en predecir la siguiente frase.
- Entrenar con frases más largas.
- Cambiar dinámicamente el enmascaramiento.

Tras realizar estos cambios, RoBERTa consigue una mejora del 15 % respecto a BERT en diversos benchmarks cómo pueden ser GLUE, RACE o SQuAD.

2.3.3. XLM

Los modelos XLM [17] nacieron con el objetivo de poder codificar frases de cualquier idioma en un mismo espacio vectorial y así también mitigar el sesgo del inglés. Para conseguir estos objetivos se pre-entrenó el modelo transformer en tres tareas distintas: CLM, MLM y TLM.

La tarea CLM, Causal Language Modeling, consiste en obtener la probabilidad de una palabra dadas todas las anteriores palabras de la frase $P(w_t|w_1, \ldots, w_{t-1}, \theta)$.

La tarea MLM es idéntica a la anteriormente mencionada en el modelo BERT.

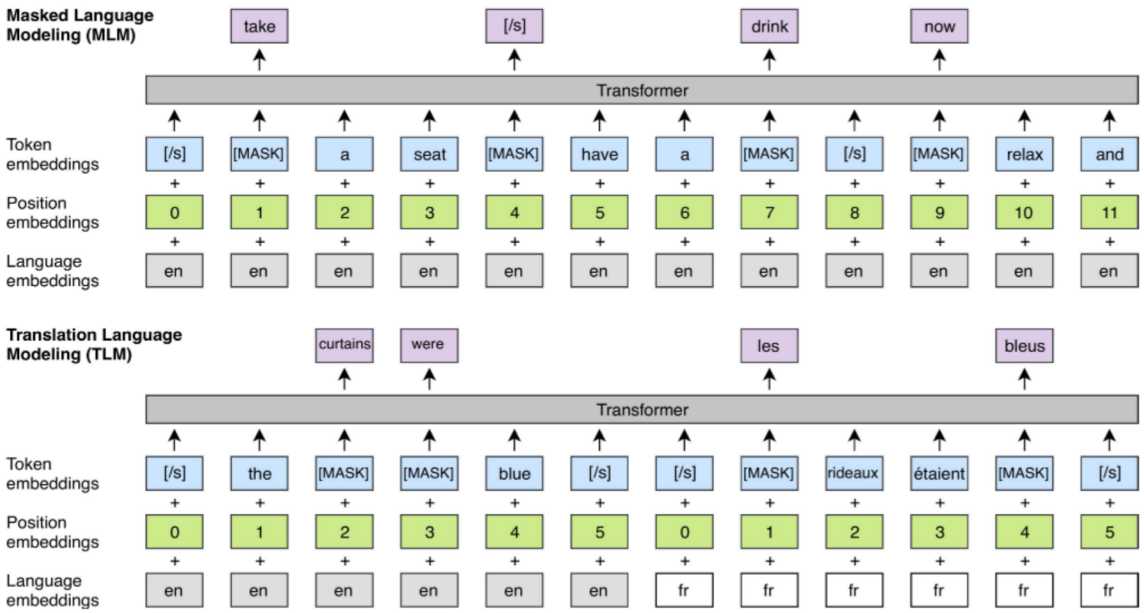

Figura 2.2: Tareas MLM y TLM del modelo XLM

La tarea TLM requiere de datos paralelos a diferencia de las dos anteriores, que son entradas con datos monolingües. Su objetivo es el mismo que MLM pero en este caso se concatenan las frases paralelas y se enmascaran palabras de ambos idiomas.

Estos modelos mostraron mejoras significativas respecto a los mejores resultados obtenidos en diferentes tareas de traducción automáticas vistas en el WMT.

2.3.4. XLM-RoBERTa

XLM-R [5] comparte el mismo objetivo que XLM, la diferencia es que en este caso se ha aumentado la escala del tamaño de entrenamiento de manera significativa. En los anteriores modelos se había entrenado con datos de Wikipedia, en este entrenamiento se han utilizado estos datos junto a una ingente cantidad de datos monolingües sacados de Common Crawl en 100 idiomas distintos. En la figura 1 se muestran los 88 idiomas con más datos, se puede ver el gran aumento que proporciona el corpus de Common Crawl respecto al de Wikipedia.

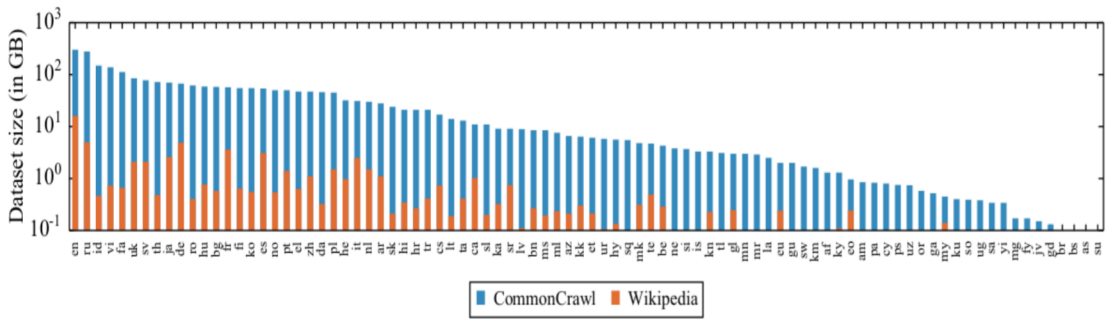

Figura 2.3: Idiomas para los que se ha entrenado XLM-R

XLM-R está pre-entrenado únicamente para la tarea MLM, antes vista en BERT y RoBERTa. Los resultados obtenidos mejoran los de XLM y mBERT, modelo multilingüe basado en BERT, en varias tareas relacionadas con la comprensión de lenguas cruzadas.

2.4 Evaluación automática de la calidad de la traducción automática

Junto a la expansión de la traducción automática también surgió el interés de obtener medidas de confianza para comprobar si estas traducciones eran válidas. Existen tres tipos de medidas de confianza: a nivel de palabra, frase y documento. En este trabajo se abordará la evaluación de las traducciones a nivel de frase.

Al principio se utilizaron sistemas basados en modelos estadísticos, la primera investigación a nivel de frase fue de Blatz et al [2]. En este caso se entrenaron clasificadores y regresores a partir de características extraídas de las traducciones. Luego se usaba la métrica automática NIST[8] para clasificación, donde se marcaba un umbral entre el percentil 5 y 30 para determinar que la traducción era buena. En el caso de la regresión, se mapeaban los scores en dos clases usando los mismos umbrales. Sin embargo, no se había comprobado si verdaderamente estos umbrales determinaban si una traducción era buena o mala.

Quirk et al. [20] utilizó clasificadores y un umbral predefinido para las traducciones "malasz "buenas"teniendo en cuenta un pequeño conjunto de 350 traducciones etiquetadas manualmente en función de su calidad . Los modelos entrenados con este conjunto de datos superaron a los entrenados con un conjunto mayor de datos etiquetados automáticamente.

Gamon et al. [9] entrenó un clasificador usando características lingüísticas extraídas a partir de traducciones humanas y automáticas para distinguir así las traducciones en estos dos tipos. Las predicciones obtenidas tuvieron una muy baja correlación con el juicio humano así que se obtuvo una indicación de que una alta apariencia humana no siempre implica una buena traducción automática y viceversa.

Specia et al. [27] presenta la selección de frases en múltiples sistemas de TA a partir de la estimación de medidas de confianza. Luego, He et al.[10] entrena un clasificador binario para predecir si el resultado de la SMT es más adecuado que la salida de la memoria de traducción utilizando la métrica TER para medir la distancia entre una traducción de referencia.

Con la llegada de la tarea relacionada con estimación de calidad de la WMT, Workshop of Machine Translation por sus siglas, y los datasets anotados que proporcionaron cada año, empezaron a lanzarse y desarrollarse sistemas open-source. Los SMT más reconocidos fueron QuEst [28] y QuEst++ [26]. Luego, junto a la revolución de las redes neuronales profundas llegaron también sistemas que emplean estos modelos. En 2017, el sistema que mejor funcionó en la WMT fue POSTECH [13], que tenía una arquitectura basada en un codificador-decodificador y RNN, que estaba concatenada con otra red neuronal recurrente bidireccional que estimaba los scores de calidad. Está arquitectura necesitaba un pre-entrenamiento muy costoso y una gran cantidad de datos paralelos. Más tarde, fue reimplementada en el sistema deepQuest [11]. Por último, se lanzó el sistema OpenKiwi [12], que mejoró los resultados de deepQuest en varios aspectos, pero aún así seguía siendo una arquitectura computacionalmente costosa.

Finalmente, a finales de 2020 fue lanzado el framework TransQuest [22], utilizado en el presente trabajo, que utiliza *embeddings* multilingües y hace uso del transformer XLM-R.

CAPÍTULO 3

Estimación automática de la calidad de traducciones en español, francés e inglés

Para abordar el problema de estimación automática de calidad se ha empleado el framework TransQuest [22], el cual tiene implementadas dos arquitecturas para el entrenamiento de la tarea.

La primera tiene cómo nombre MonoTransQuest y se basa en la arquitectura predictor-estimador [14], la cual se ha utilizado en los modelos vistos en el WMT2021 [25] y ha obtenido resultados que marcan el estado del arte en este campo. El predictor está formado por un Transformer, en este caso XLM-R, que tiene cómo entrada las frases de origen y destino separadas por un token de separación. Luego, se han experimentado tres diferentes tipos de pooling para la salida del transformer: estrategia cls, estrategia max y estrategia mean. La primera consiste en quedarse con el token cls, la cual es la que mejores resultados ha obtenido y la que se emplea en el framework, la segunda consiste en calcular la media de todos los vectores de salida y, por último, la estrategia max consiste en quedarse con el máximo de los vectores. Finalmente, el token CLS sirve cómo entrada al estimador, el cual está compuesto por una red neuronal feed-forward y una función softmax, que aprende el valor de calidad de la frase. En la siguiente figura se muestra gráficamente la anterior descripción:

La segunda arquitectura, llamada SiameseTransQuest y basada en los modelos con redes neuronales siamesas [23], está formada por dos Transformer XLM-R, uno para la frase de entrada y otro para la de salida. Los autores del framework han experimentado siguiendo las mismas estrategias anteriormente comentadas. Para ello, se ha calculado la similaridad coseno para los vectores de salida de los dos transformers empleando los tres tipos de pooling mencionados en el anterior apartado: cls, max y mean. En este caso la mejor estrategia ha sido el mean pooling. En la siguiente figura se puede observar arquitectura descrita previamente:

Cómo se podría intuir, el sistema SiameseTransQuest tiene un entrenamiento menos costoso, ya que el MonoTransQuest tiene que aprender a partir del valor de calidad. Sin embargo, se ha visto que la precisión del modelo SiameseTransQuest es menor. Por ello, se ha decidido utilizar el MonoTransQuest para el desarrollo del proyecto.

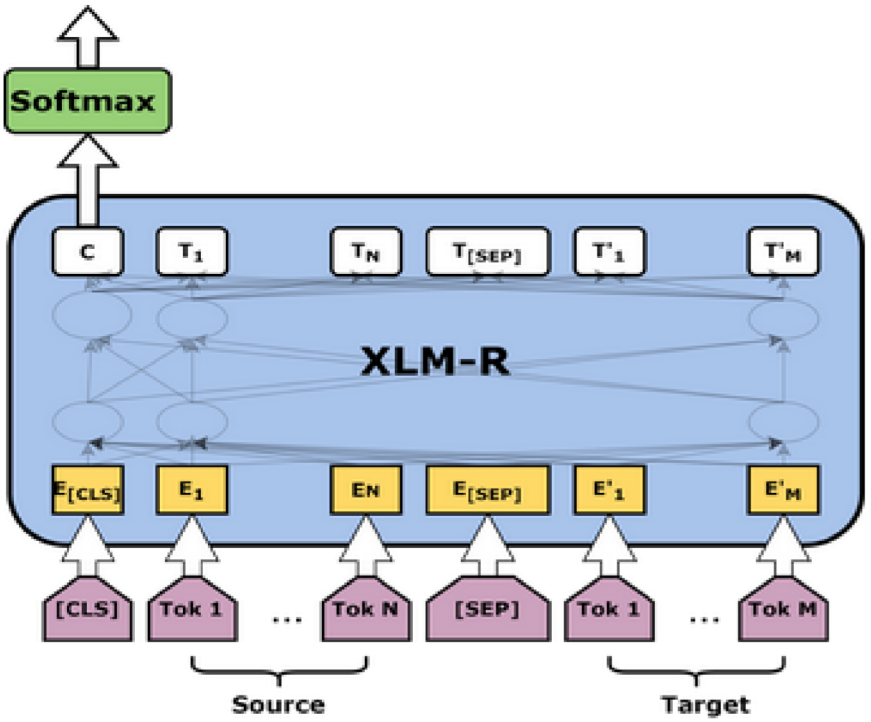

Figura 3.1: Arquitectura MonoTransQuest

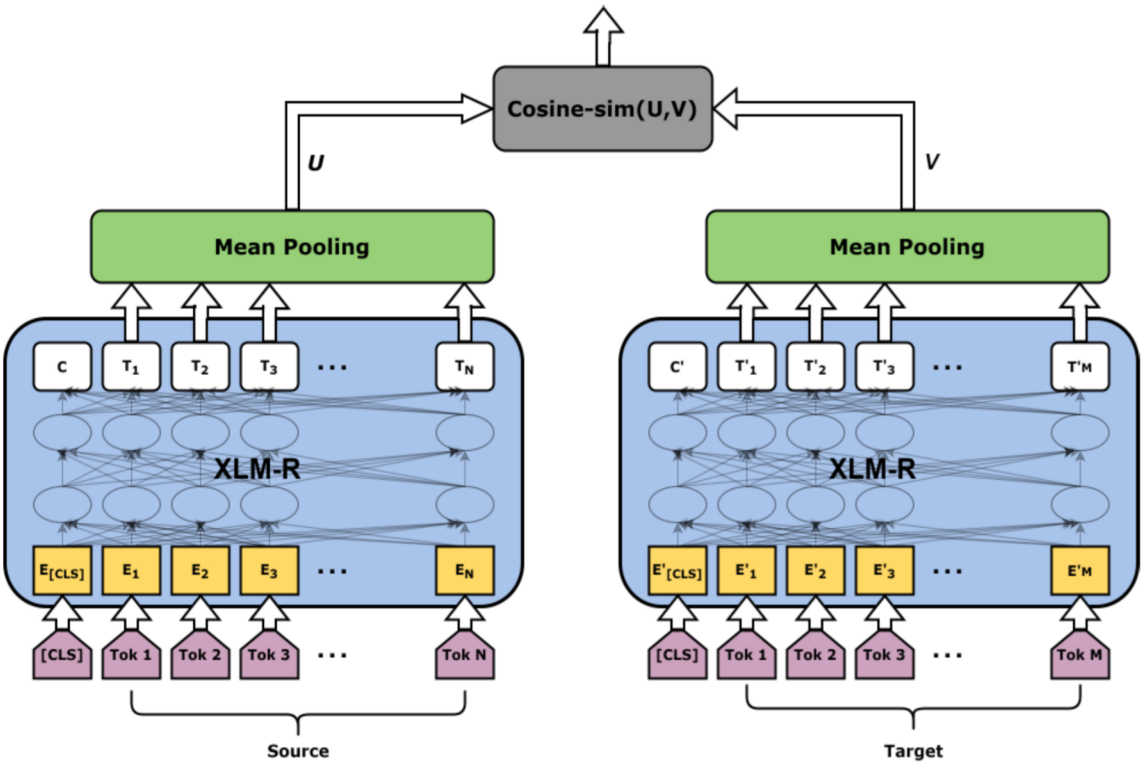

Figura 3.2: Arquitectura SiameseTransQuest

CAPÍTULO 4
Marco experimental

4.1 Tecnologias empleadas

El lenguaje de programación que se ha utilizado para desarrollar los experimentos del presente trabajo ha sido Python, ya que es el más popular dentro de la ciencia de datos por su gran legibilidad y por el gran abanico de librerías relacionadas con el machine learning y el PLN que hay implementadas en este lenguaje.

En cuanto a librerías se refiere, se han utilizado las siguientes:

Pandas - Esta librería está especializada en el manejo y análisis de estructuras de datos. Se utiliza principalmente para leer y aplicar funciones sobre tablas de datos.

Re - Librería que permite el manejo de expresiones regulares. Se utiliza para la limpieza de los corpus.

Sacrebleu - Librería que permite calcular las métricas TER y BLEU.

Transquest - Librería que tiene implementadas las arquitecturas explicadas en el capítulo anterior. Esta libreria está basada en Pytorch y tiene una serie de modelos preentrenados, a parte de las configuraciones de las diferentes arquitecturas.

Matplotlib - Librería para crear visualizaciones estáticas, animadas e interactivas.

Seaborn - Librería complementaria a Matplotlib para crear visualizaciones.

Scipy - Librería cientifíca que tiene implementadas las funciones para calcular las distintas correlaciones empleadas en el presente trabajo.

4.2 Conjunto de datos

El conjunto de datos utilizado para el entrenamiento del modelo proviene del proyecto europeo NTEU, Neural Translation for the EU por sus siglas, el cual tiene como objetivo construir motores de traducción de forma directa entre todas las lenguas europeas. Esto se hace para no tener que utilizar siempre el inglés, que sirve como lengua pivote entre todos los pares europeos.

En este trabajo se han utilizado dos corpus paralelos con las combinaciones español-inglés y francés-inglés. En las tablas 4.1 y 4.2 se muestran las estadísticas para los corpus mencionados anteriormente.

Estadística	Español	Inglés
Número de líneas	1M	1M
Número de palabras	32.6M	27.7M

Tabla 4.1: Estadísticas recogidas del conjunto de datos para la combinación de idiomas español-inglés.

Corpus español-inglés

Si observamos la distribución de la longitud de las frases en español se puede apreciar que la mayoría se sitúan entre 0 y 100 palabras. Luego, la media obtenida es de 32 palabras y hay 1145 frases que superan las 200 palabras.

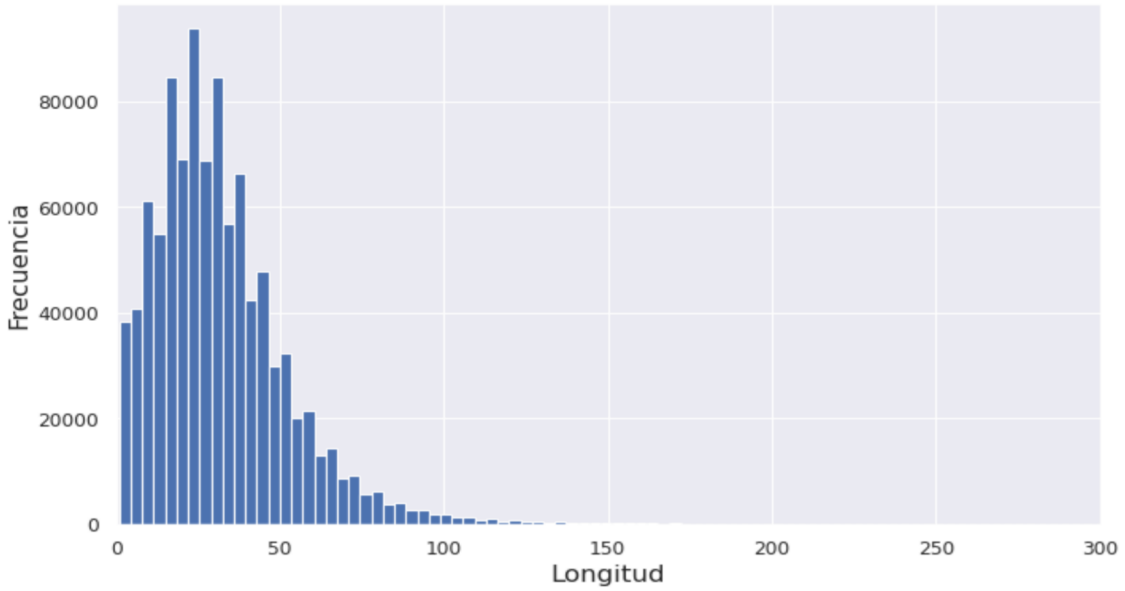

Figura 4.1: Distribución de los datos en español del corpus español-inglés

La distribución para las frases en inglés sigue el mismo patrón que la anterior. En cambio, la media es de 27 palabras y hay 827 frases que superan las 200 palabras. Se puede intuir que en inglés se utilizan menos palabras que en francés.

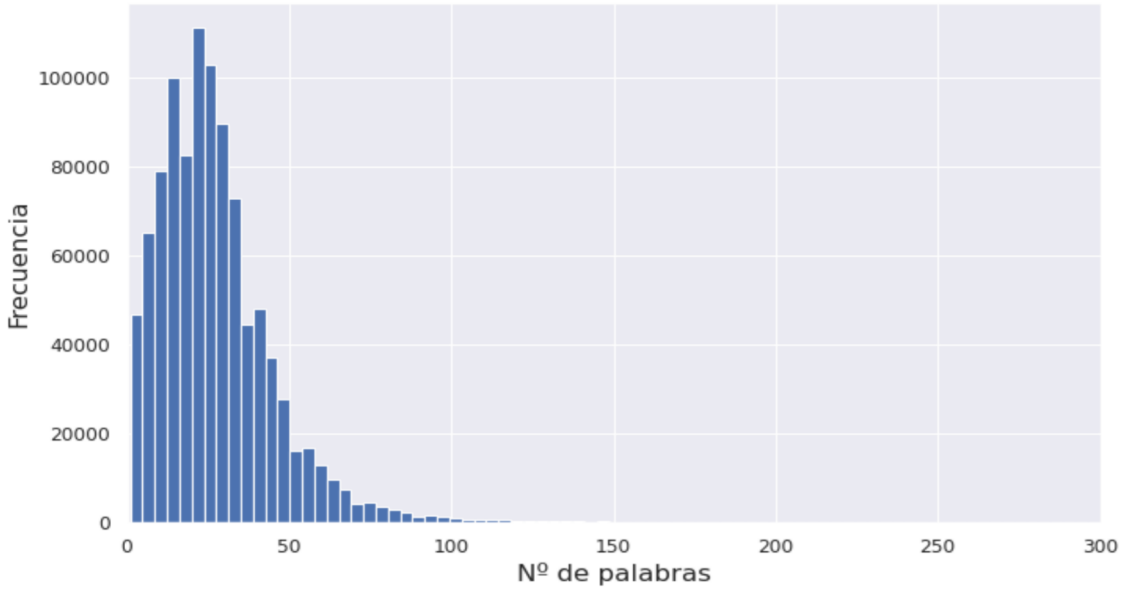

Figura 4.2: Distribución de los datos en inglés del corpus español-inglés

Si representamos las longitudes de la frase origen y destino en un gráfico de dispersión obtendremos lo siguiente:

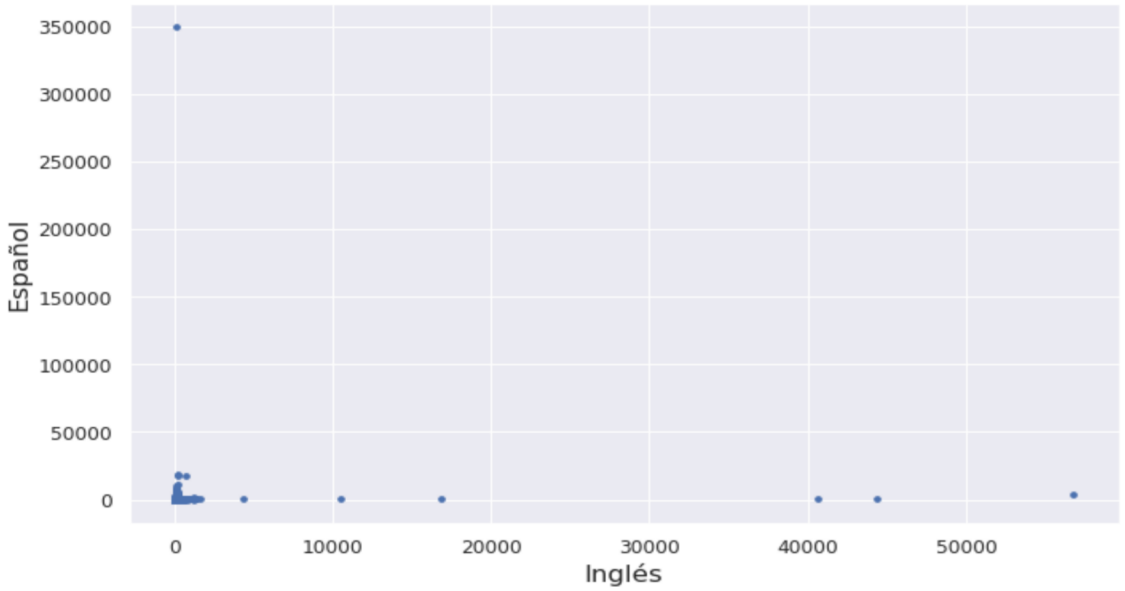

Figura 4.3: Gráfico de dispersión para las frases en español e inglés

Podemos ver que el gráfico sale mal, esto es debido a los puntos que se sitúan en los extremos del gráfico. Se puede considerar que estos datos son anómalos y que se eliminarán en el proceso de limpieza.

Si limitamos los ejes a un valor de 500 ya se puede observar con mejor detalle la dispersión de los datos. Se puede ver que los datos siguen en mayor medida una línea diagonal, este hecho nos indica que las frases en ambos idiomas tienen un número de palabras parecidas.

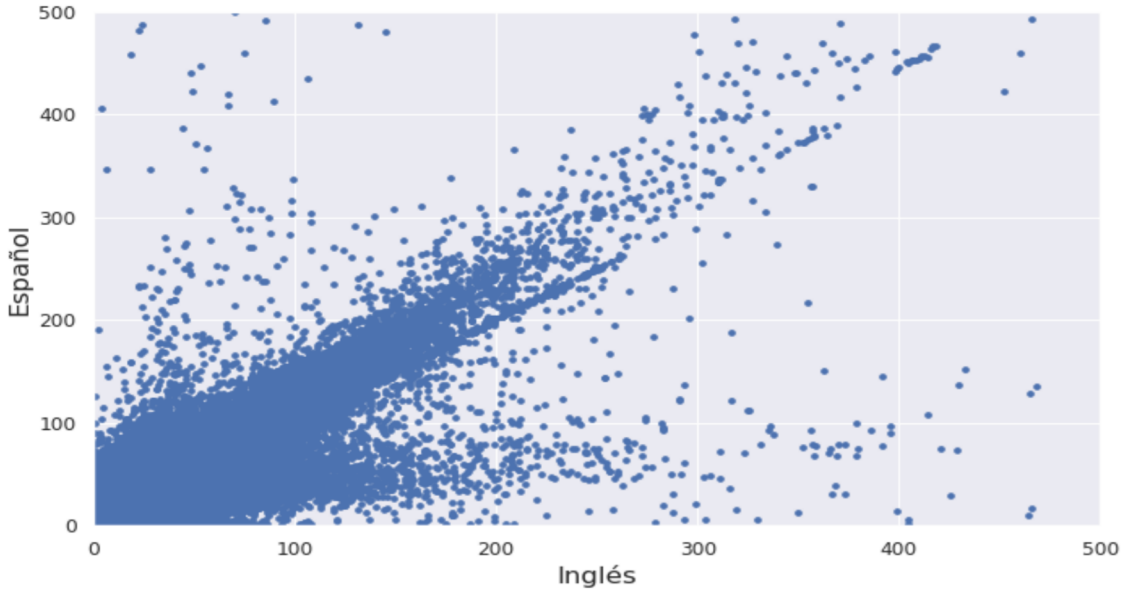

Figura 4.4: Gráfico de dispersión para las frases en inglés y español tras limitar los ejes

Corpus francés-inglés

Estadística	Francés	Inglés
Número de líneas	1M	1M
Número de palabras	30.8M	27.1M

Tabla 4.2: Estadísticas recogidas del conjunto de datos para la combinación de idiomas francés-inglés.

Si observamos la distribución de la longitud de las frases en francés se puede ver que sigue el mismo patrón que las anteriores distribuciones, donde los valores se sitúan entre 0 y 100 palabras. En este caso la media obtenida es de 30 palabras y hay 812 frases que superan las 200 palabras.

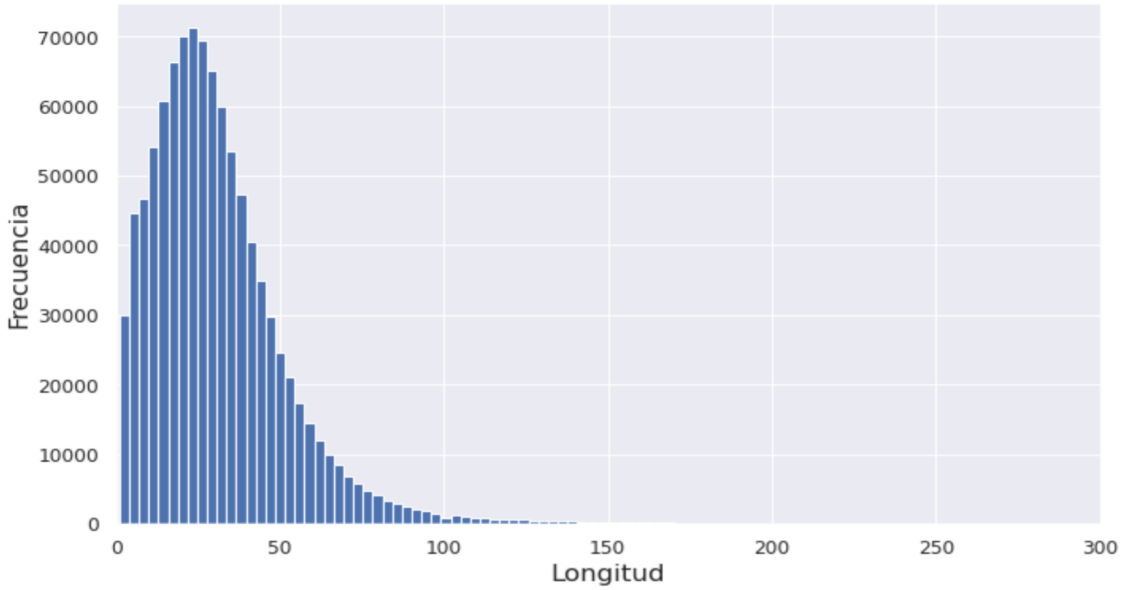

Figura 4.5: Distribución de los datos en francés del corpus francés-inglés

En el caso de las frases en inglés del corpus francés-inglés se obtiene una media de longitud de 27 palabras. Se observan 590 frases con más de 200 palabras. En este corpus se aprecian menos valores atípicos, además, la media de palabras entre lenguas difiere en menor medida respecto al par español-inglés.

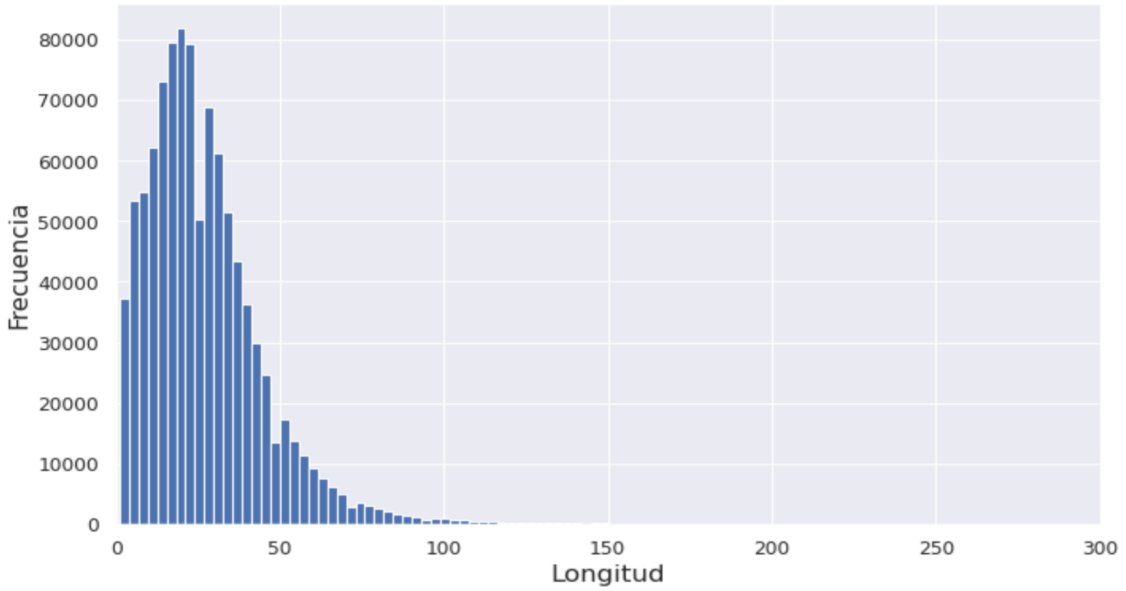

Figura 4.6: Distribución de los datos en inglés del corpus francés-inglés

Si volvemos a representar las longitudes de la frase origen y destino para este corpus en un gráfico de dispersión obtenemos lo siguiente:

Se puede apreciar que la escala de los ejes difiere en varias magnitudes. Esto es porque las frases de origen oscilan entre valores normales y no hay datos con más de 300 palabras, en cambio para las frases de destino hay un grupo de frases anómalas situadas a la derecha.

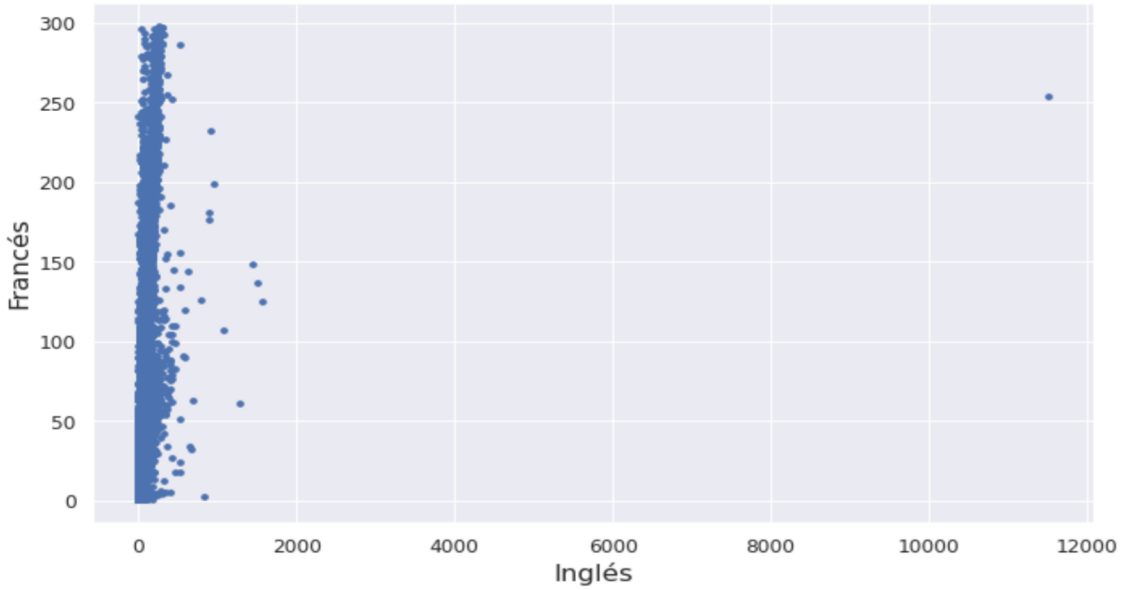

Figura 4.7: Gráfico de dispersión para las frases en francés e inglés

Si volvemos a limitar los ejes a un valor de 500 ya se puede observar con mejor detalle la dispersión de los datos. Se puede ver que los datos siguen en mayor medida una línea diagonal, este hecho nos indica que las frases en ambos idiomas tienen un número de palabras parecidas.

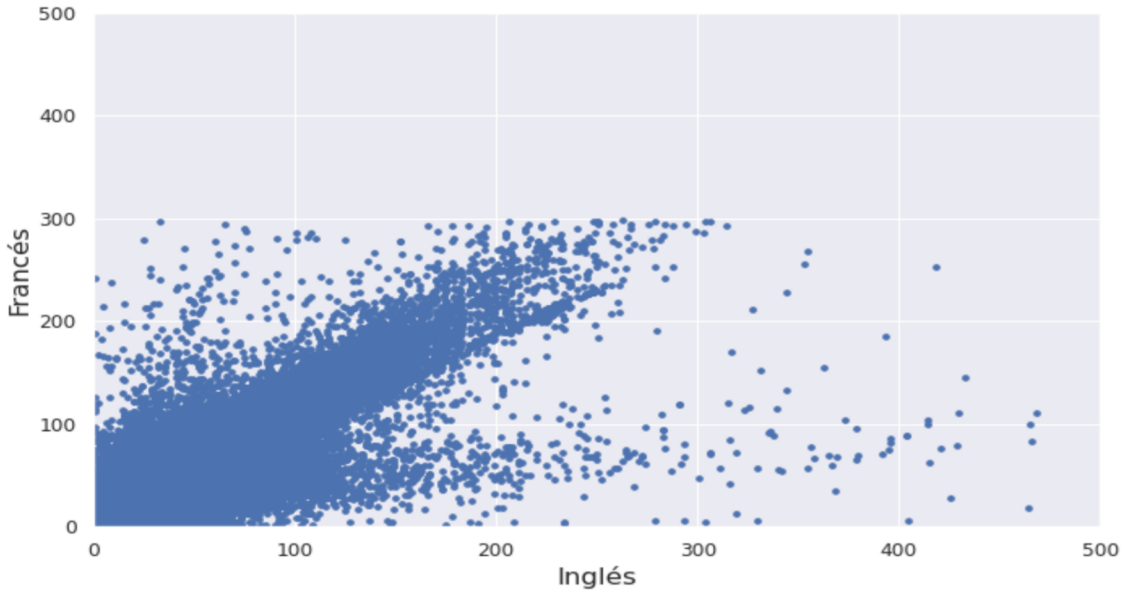

Figura 4.8: Gráfico de dispersión para las frases en francés e inglés tras limitar los ejes

4.3 Preparación de los datos

Como se ha podido ver en el capítulo anterior, existen datos anómalos los cuales tienen errores ya sean de longitud o por diferentes motivos. Para solucionar estos problemas se aplican preprocesos a los diferentes corpus con el objetivo de limpiar los datos y que sean de la máxima calidad posible para así entrenar un buen modelo para los diferentes idiomas. Por un lado, tenemos el conjunto de preprocesos que conforman la normalización de un segmento, al cual llamaremos **normalizadores**. Por otra parte, se define el conjunto de métodos que determinan si un texto es válido, al cual llamaremos **validadores**. A continuación, se explicarán cada uno de los métodos pertenecientes a cada conjunto.

4.3.1. Normalizadores

El objetivo de estás técnicas de normalización es conseguir que las oraciones que contienen cierto ruido corregible, se puedan seguir empleando aplicándoles dichas correcciones. A continuación, se muestra la lista de normalizadores que se han utilizado para la limpieza:

- Normalización de espacios en blanco repetidos a un único espacio en blanco, se aplica tanto en la frase de origen como en la frase de destino.
- Normalización de las puntuaciones repetidas a una puntuación, por ejemplo, se sustituyen puntuaciones como “????” por “?”, se aplica tanto en la frase de origen como en la frase de destino.
- Normalización de tildes en la correcta codificación unicode.
- Para la preparación de datos en francés se ha hecho un normalizador especial que añade el espacio especial (non-breaking) que tiene esta lengua con los signos de puntuación (“!”, “?”, “;”, “:”).
- Para la preparación de datos en español se ha eliminado la tilde en las siguientes palabras: solo, esto, esta, estos, estas, este, aquel, aquello, aquella, aquellos, aquellas, ese, eso, esa, esas, esos. Ya que antes contenían una tilde y en la nueva norma de español ya no la contienen.

4.3.2. Validadores

El objetivo de estos procesos de validación es verificar si una frase cumple una serie de condiciones para así poder pertenecer al conjunto de datos finales. A continuación, se muestra la lista de validadores que se han implementado en este trabajo:

- Eliminación de segmentos vacíos.
- Eliminación de segmentos que contengan en su mayoría números, puesto que dichos segmentos no contienen información relevante para la generación de nuevos segmentos a partir de dichos segmentos.
- Eliminación de segmentos que no contengan en el lado del idioma origen y en el lado del idioma destino la misma cantidad de números y elementos de puntuación.
- Eliminación de segmentos que sólo contengan direcciones de correos electrónicos o de Internet.

- Eliminación de segmentos en los que la frase fuente sea idéntica a la frase destino ya que esto no sería útil porque no está traducida o no aporta información.
- Eliminación de segmentos donde el origen o destino sea tres veces más larga que la contraria.
- Eliminación de frases donde las comillas dobles, los paréntesis o corchetes no se abran y cierren correctamente.

Tras aplicar los anteriores métodos a los dos corpus se han obtenido las métricas de la siguiente tabla:

	Frases Iniciales	**Frases actuales**
Corpus Español-Inglés	1M	872.9K
Corpus Francés-Inglés	1M	864.4K

Tabla 4.3: Estadísticas recogidas de los diferentes conjuntos de datos tras el proceso de limpieza

Si representamos los datos procesados en un gráfico de dispersión como en el anterior capítulo obtenemos los siguientes resultados:

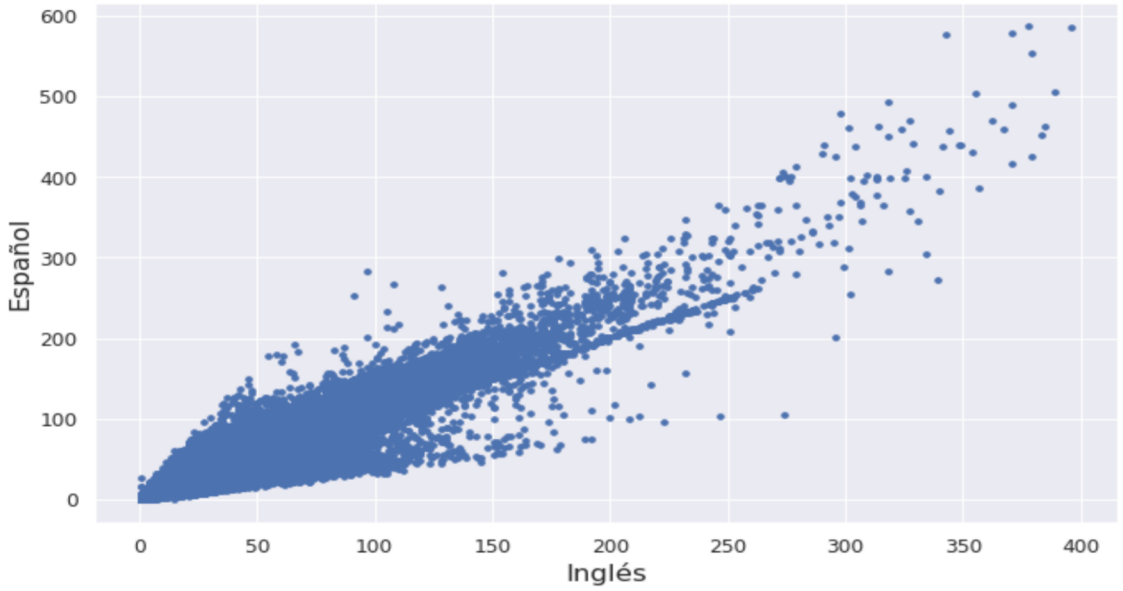

Figura 4.9: Gráfico de dispersión para las frases en español e inglés tras limpiar los datos

Si comparamos este gráfico con la figura 4.3 se puede apreciar que ya no existen los valores anómalos que provocaban que la gráfica saliera mal. Además, la linea diagonal se aprecia mejor que en la figura 4.4, hecho que nos hace pensar que la limpieza ha mejorado la calidad de los datos.

En el caso del corpus francés-inglés se puede apreciar el mismo efecto respecto a las figuras 4.7 y 4.8.

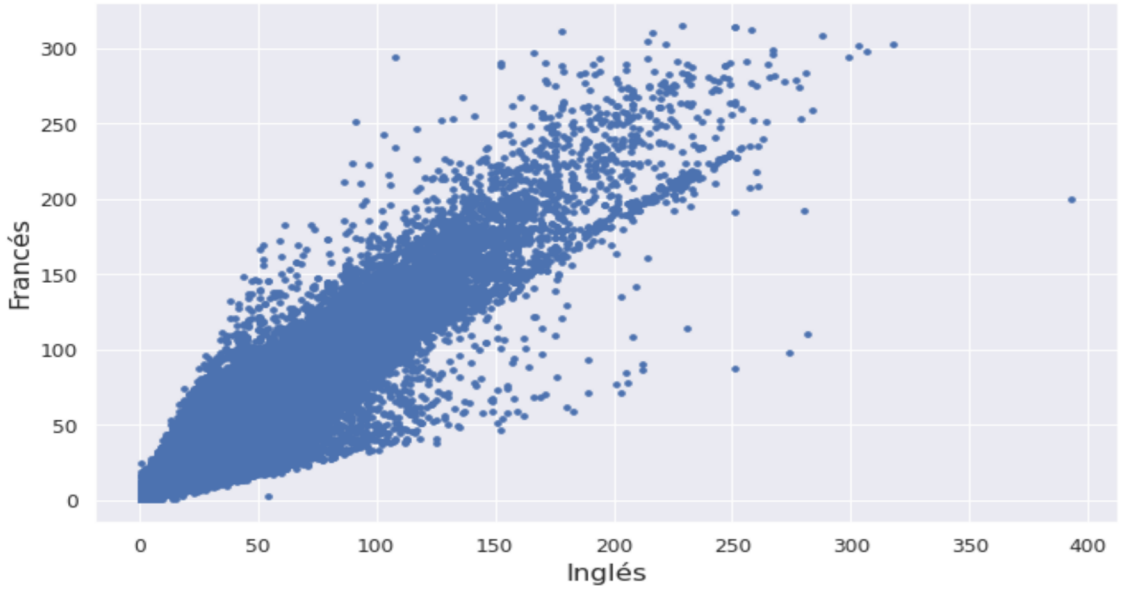

Figura 4.10: Gráfico de dispersión para las frases en francés e inglés tras limpiar los datos

Tras limpiar los datos, se tradujeron los datos de entrada con un motor de traducción proporcionado por Pangeanic. Al obtener las traducciones automáticas, se calculó el TER entre estas y las frases de destino, que son las utilizadas como referencia.

Finalmente, el aspecto que tiene el conjunto de datos con el que se va a entrenar y evaluar el modelo es el siguiente:

Origen	Traducción	TER
FECHA DE NACIMIENTO	DATE OF BIRTH	0.0
Detalles del tutor	Details of the guardian	25.0

Tabla 4.4: Aspecto del conjunto de datos final.

4.3.3. Partición del conjunto de datos

En relación a la partición de los conjuntos de datos en entrenamiento, validación y test se ha utilizado la siguiente metodología. Se ha seleccionado para ambos corpus un 5 % del conjunto total para validar el modelo y otro 5 % para testearlo. Respecto al modelo multilingüe, se han juntado los datos de entrenamiento pertenecientes a los dos pares de idiomas y se han mezclado aleatoriamente. En la tabla 4.5 se pueden ver las estadísticas de los diferentes conjuntos.

Conjunto	# Muestras
Entrenamiento	
Francés	778K
Español	786K
Multilingüe	1.5M
Validación	
Francés	43K
Español	44K
Test	
Francés	43K
Español	44K

Tabla 4.5: Estadísticas de los conjuntos de entrenamiento, validación y test para los modelos español-inglés, francés-inglés y español/francés-inglés

En las siguientes figuras se puede ver como se ha respetado la distribución de TER para escoger el conjunto de test. Se puede apreciar tanto en la figura 4.11 como en la figura 4.12 que los datos oscilan en mayor medida entre el 0 y 100, habiendo picos para valores como 0, 50 y 100. En el caso del corpus francés-inglés se obtienen resultados parejos, como se puede observar en las figuras 4.13 y 4.14

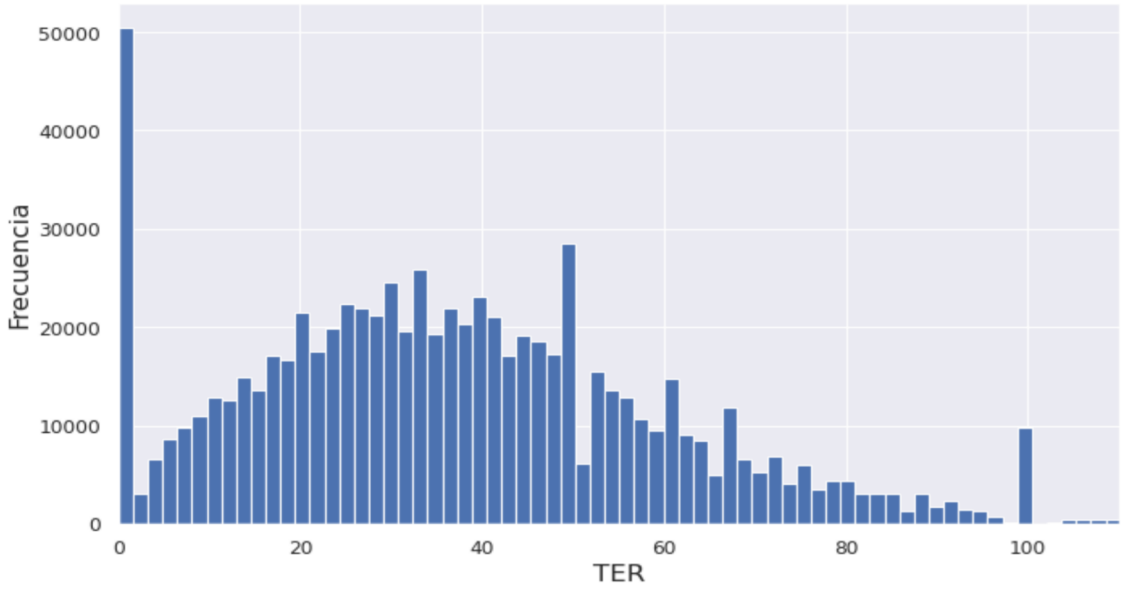

Figura 4.11: Distribución de la variable TER en los datos de entrenamiento pertenecientes al par de lenguas español-inglés

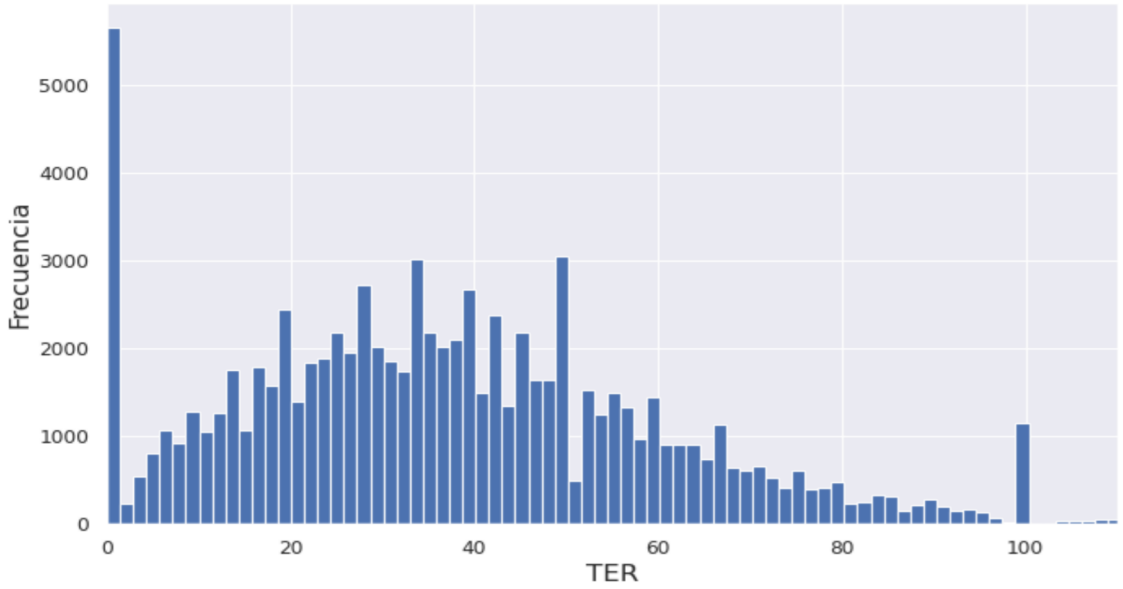

Figura 4.12: Distribución de la variable TER en los datos de test pertenecientes al par de lenguas español-inglés

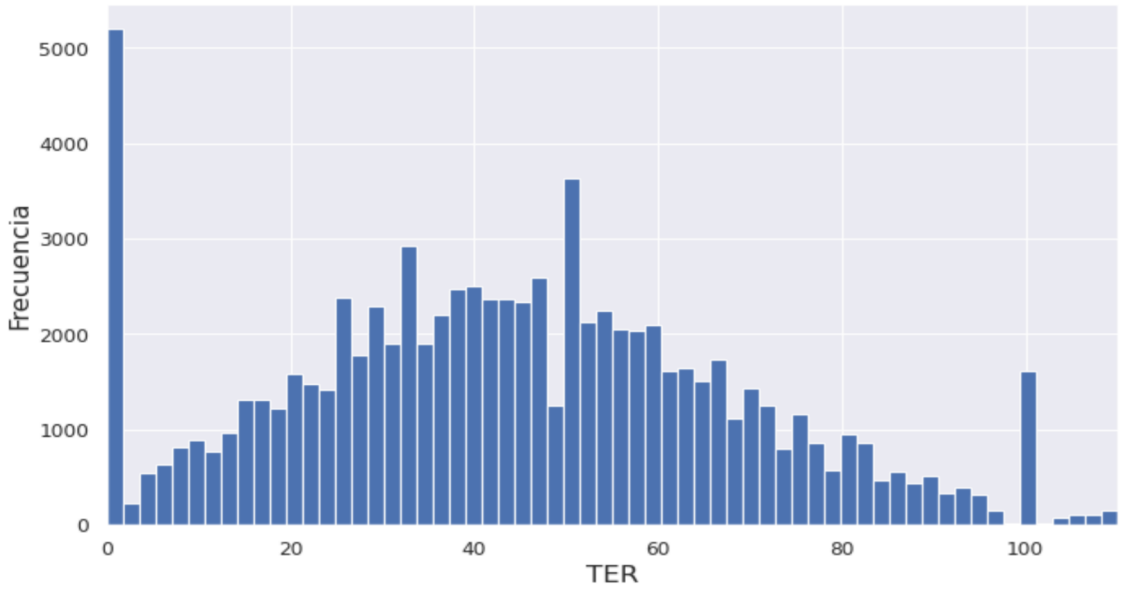

Figura 4.13: Distribución de la variable TER en los datos de entrenamiento pertenecientes al par de lenguas francés-inglés

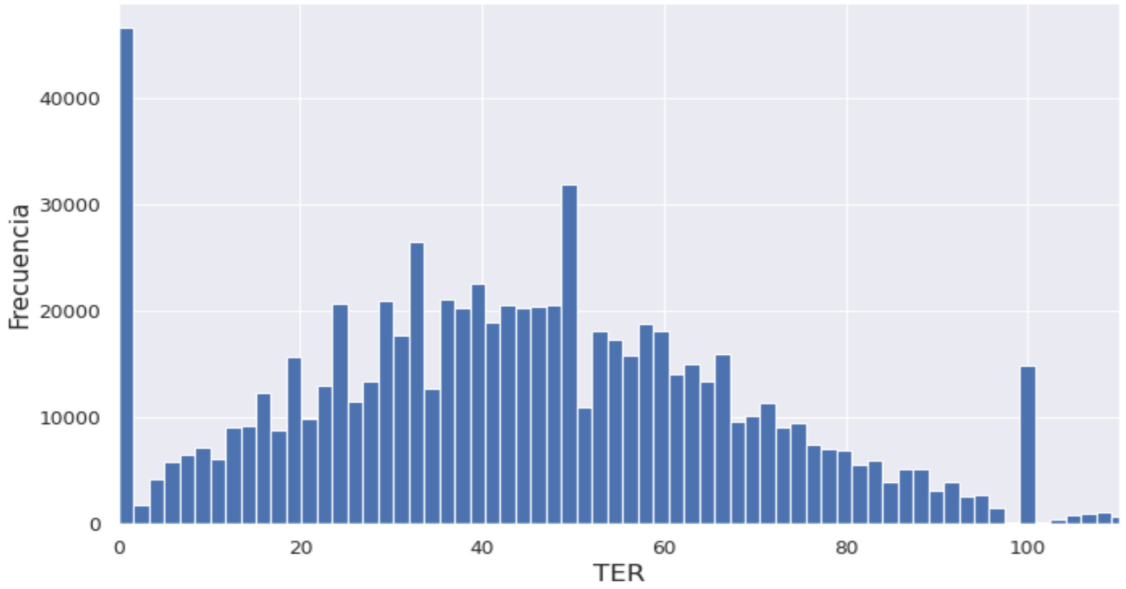

Figura 4.14: Distribución de la variable TER en los datos de test pertenecientes al par de lenguas francés-inglés

4.4 Experimentación

Con el objetivo de obtener un sistema de evaluación automática de la traducción automática, se va a experimentar con los diferentes corpus presentados anteriormente. A parte, se va a comparar el rendimiento de los modelos monolingües con el del multilingüe.

Para el entrenamiento, se utiliza un tamaño de batch de 8, así como un total de 3 épocas. Además se emplea el optimizador Adam [15] con un ratio de aprendizaje de 2e5 y un aumento del ratio de aprendizaje lineal sobre el 10 % de los datos de entrenamiento. Durante el proceso de entrenamiento los parámetros del modelo XML-R, así como los de las siguientes capas de clasificación, van siendo actualizados. Finalmente, se ha aplicado *early stopping* con una paciencia de 10 pasos.

4.5 Evaluación

En el presente trabajo se han utilizado dos tipos de métricas de evaluación, unas relacionadas con la calidad de la traducciones y otras relacionadas con la correlación de los datos.

4.5.1. Métricas relacionadas con la calidad de la traducciones

En el presente trabajo se han utilizado dos métricas relacionadas con la evaluación de la calidad de la traducción: BLEU [19] y TER [24].

BLEU

BLEU, BiLingual Evaluation Understudy por sus siglas, es una métrica para evaluar la calidad de un texto traducido automáticamente de una lengua a otra. BLEU no es una tasa de error, es una puntuación, es decir, cuanto más alta sea la puntuación BLEU es mejor.

Para obtener esta métrica se calcula la media geométrica de la precisión modificada por n-gramas, p_n, la cual consiste en el conjunto de todos los recuentos de n-gramas candidatos y sus correspondientes recuentos máximos de referencia, dichos recuentos de candidatos se recortan por su valor máximo de referencia correspondiente, sumado, y dividido por el número total de n-gramas candidatos, y multiplicado por el factor BP que penaliza las frases cortas. De esta manera, BLEU se puede formalizar de la siguiente forma:

$$BLEU = BP * exp\left(\sum_{n=1}^{N} \frac{log\ p_n}{N}\right) \tag{4.1}$$

La definición más común de BLEU se calcula sobre la concatenación de todas las frases de test, y normalmente emplea n-gramas de orden 4. Por otro lado, el resultado final de la métrica es un valor entre 0 y 1, dicho valor suele multiplicarse por 100 para obtener una mayor capacidad de interpretación representándose así como porcentaje.

TER

TER, Translation Error Rate por sus siglas, mide el esfuerzo de posedición que se necesita para corregir una traducción respecto a una referencia. Esta medida se define como el número mínimo número de ediciones necesarias para cambiar el resultado de la traducción automática de forma que coincida exactamente con la referencia, normalizada por la longitud de la referencia. Las ediciones incluyen la inserción, la supresión y la sustitución de palabras individuales, como cualquier métrica de distancia de edición estándar, así como desplazamientos de secuencias de palabras. TER se formaliza de la siguiente manera:

$$TER = \frac{\#\ ediciones}{\#\ palabras\ dereferencia} \tag{4.2}$$

4.5.2. Métricas relacionadas con la correlación de los datos

Después del proceso de entrenamiento de los modelos se tiene que realizar una evaluación de estos, con el fin de valorar la certeza de los valores de calidad obtenidos y si tienen correlación con los valores verdaderos. En este trabajo se van a utilizar dos coeficientes de correlación: Pearson y Spearman.

Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson es una medida de dependencia lineal entre dos variables aleatorias cuantitativas, además es independiente de la escala de la medida de las variables. Este coeficiente se modeliza de la siguiente forma:

$$\rho X, Y = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} \tag{4.3}$$

Donde la X y la Y hacen referencia a las predicciones y a los datos de test.

Coeficiente de correlación de Spearman

El coeficiente de correlación de Spearman tiene la misma interpretación que el de Pearson, la principal diferencia entre ellos es que el de Spearman es más robusto frente a

valores extremos. Para calcular este coeficiente se ordenan los datos y se hace un ranking de estos. Luego se aplica la siguiente fórmula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{4.4}$$

Donde d es la diferencia entre los rankings de la predicción y el valor real.

CAPÍTULO 5
Resultados

En este capítulo se van a exponer los resultados obtenidos tras experimentar con los diferentes conjuntos de datos y los diferentes modelos expuestos anteriormente.

Para ello se realizará un análisis aplicado a un problema de traducción junto a un análisis cualitativo y otro cuantitativo.

5.1 Análisis aplicado a un problema de traducción

En la siguiente sección se ha planteado un escenario donde se han establecido diferentes umbrales basados en la métrica TER para así simular el esfuerzo de posedición que se necesitaría para corregir las traducciones. Esta herramienta serviría para elegir el umbral que abarque el mayor número de frases con el menor esfuerzo y error posible.

5.1.1. Resultados Francés-Inglés

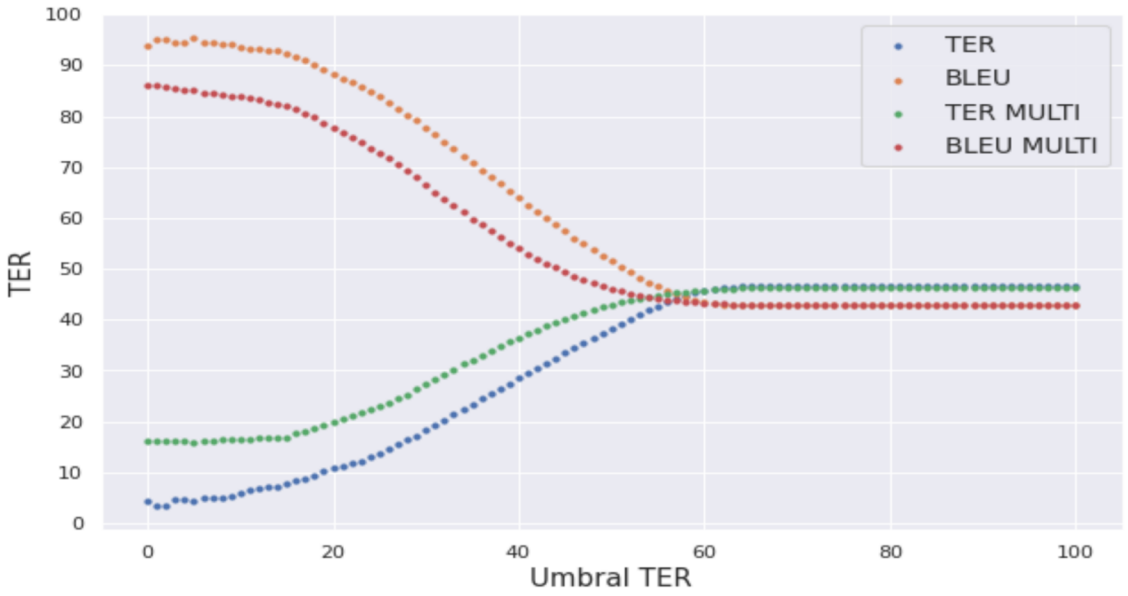

Figura 5.1: TER y BLEU medio respecto al umbral de TER escogido para determinar si una frase es buena o mala. Las líneas se corresponden a los modelos monolingüe y multiligüe para el par francés-inglés

El primer experimento se basa en ir computando el esfuerzo de posedición según un umbral que indica si las frases predichas son buenas o no. Es decir, si se escoge un umbral de 10, todas las frases que el modelo prediga con un valor menor que 10 serán consideradas como buenas y el resto serán sustituidas por la frase de referencia para simular que han sido corregidas por un lingüista, luego se calculará el TER real a nivel de corpus. De esta manera se puede observar a partir de que umbral el esfuerzo aumenta o disminuye en mayor o menor medida.

De la figura 5.1 se puede deducir que el modelo monolingüe funciona mejor que el multilingüe, ya que las líneas de TER y BLEU están por debajo y por arriba respectivamente.

Luego, se ha calculado el mismo experimento simulando que las predicciones han resultado iguales a los datos reales, es decir, ha sido una predicción perfecta. Este modelo perfecto ha sido nombrado como oráculo y sirve para comparar cuánto se acerca nuestro modelo a la perfección.

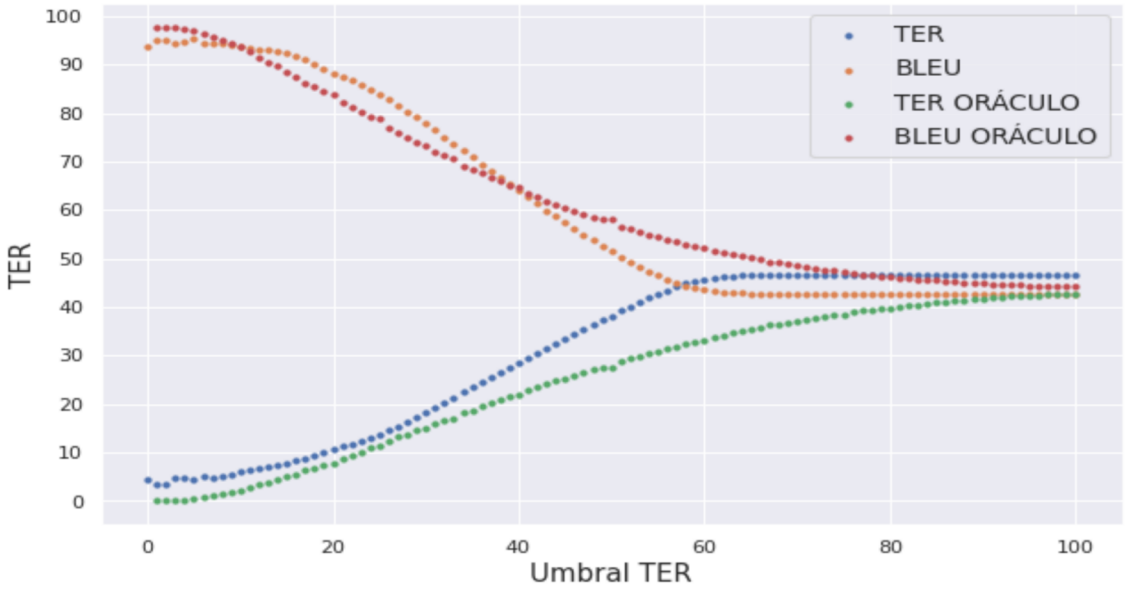

Figura 5.2: TER y BLEU medio respecto al umbral de TER escogido para determinar si una frase es buena o mala. Las líneas se corresponden al modelo monolingüe y al oráculo para el par francés-inglés

Tras observar la figura 5.2 se puede ver como en el umbral 30 el modelo se acerca mucho al oráculo. De hecho, para la métrica BLEU mejora el resultado del oráculo, esto suceso puede ocurrir cuando el BLEU y el TER no están muy correlacionados. De este modo se podría decir que el 30 es un buen umbral para fijar si una frase es buena o mala.

Para complementar este experimento se han calculado el número de frases exactas que serían poseditadas según el umbral escogido.

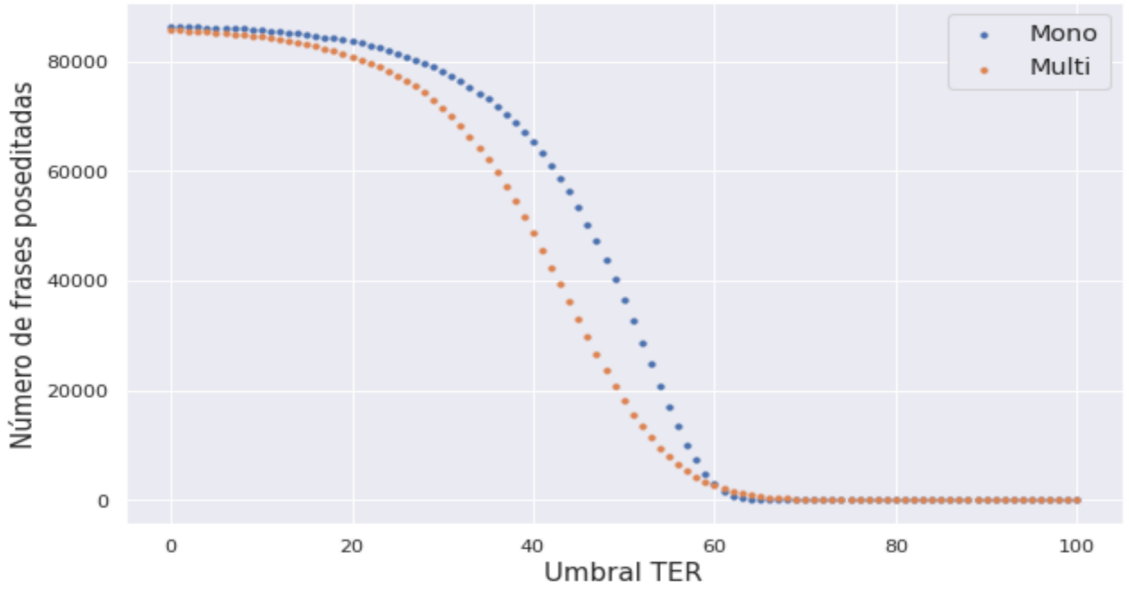

Figura 5.3: Nº de frases poseditadas respecto al umbral de TER escogido para determinar si una frase es buena o mala. Las líneas se corresponden al modelo monolingüe y al multilingüe para el par francés-inglés

Si se observa la figura 5.3 se puede ver como para el umbral 30 se poseditan muchas frases, por lo tanto aunque el error fuese muy bajo, no se asegura una cantidad muy alta de frases predichas. Este umbral seria bueno para un cliente que quiera frases fiables, aunque sean pocas, y con un esfuerzo de posedición relativamente bajo, 30 o menos.

Finalmente, se ha realizado un experimento donde se han ordenado las filas de menor a mayor predicción, luego se han ido quitando las peores frases, desde el 0 % de peores frases hasta el 100 %. Todo esto para simular que se han corregido una cantidad determinada de las peores frases. Finalmente, se ha calculado el TER global tras reemplazar las frases malas por la referencia.

Para este experimento se ha vuelto a calcular el modelo oráculo y ahora también se ha calculado un modelo aleatorio, el cual simula que se han hecho predicciones aleatorias y luego se ha hecho el experimento. También se ha añadido la figura 5.5 donde se muestra el valor globlal del BLEU, métrica más representativa a la hora de mostrar la calidad de las frases y que complementa el experimento para dar una mayor certeza.

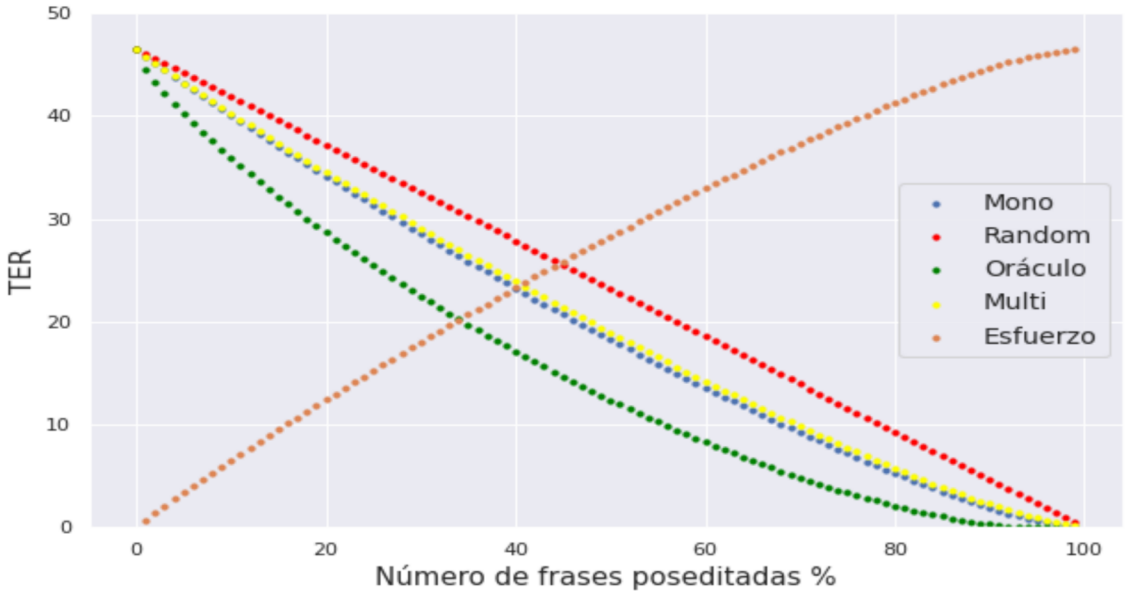

Figura 5.4: TER global según el porcentaje de frases poseditadas para el par de lenguas francés-inglés

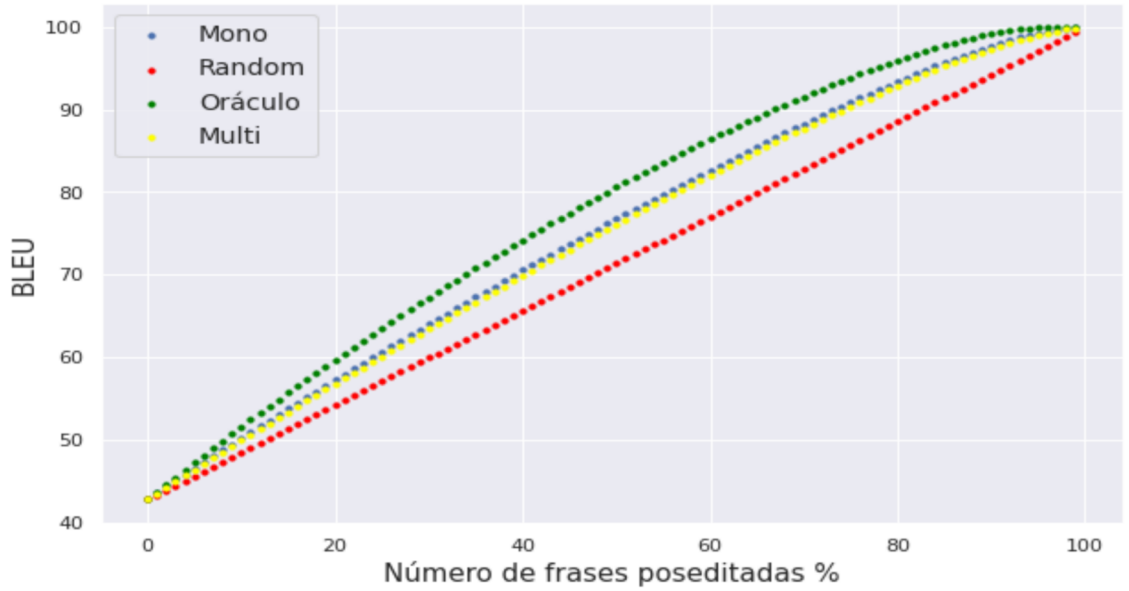

Figura 5.5: BLEU global según el porcentaje de frases poseditadas para el par de lenguas francés-inglés

Lo primero que se puede apreciar en la figura 5.4 es que las líneas de los modelos se sitúan entre el oráculo y el modelo aleatorio, cosa que nos hace pensar que nuestro modelo funciona correctamente en cierta medida. Luego, se puede ver como el modelo monolingüe se acerca más al oráculo que el multilingüe, hecho que se ha repetido en todos los experimentos. Finalmente, se ha calculado el esfuerzo, medido en unidades de TER, que se necesitaria para corregir las frases que han sido poseditadas. Así pues, en un escenario donde un cliente quisiera unos datos con una calidad de 10 puntos de TER o menos, nos ahorraríamos un 25 % de frases a poseditar aproximadamente y el resto habría que poseditarlas sabiendo que el esfuerzo medio es de casi 40 puntos de TER.

5.1.2. Resultados Español-Inglés

Para el par de lenguas Español-Inglés se han realizado los mismos experimentos.

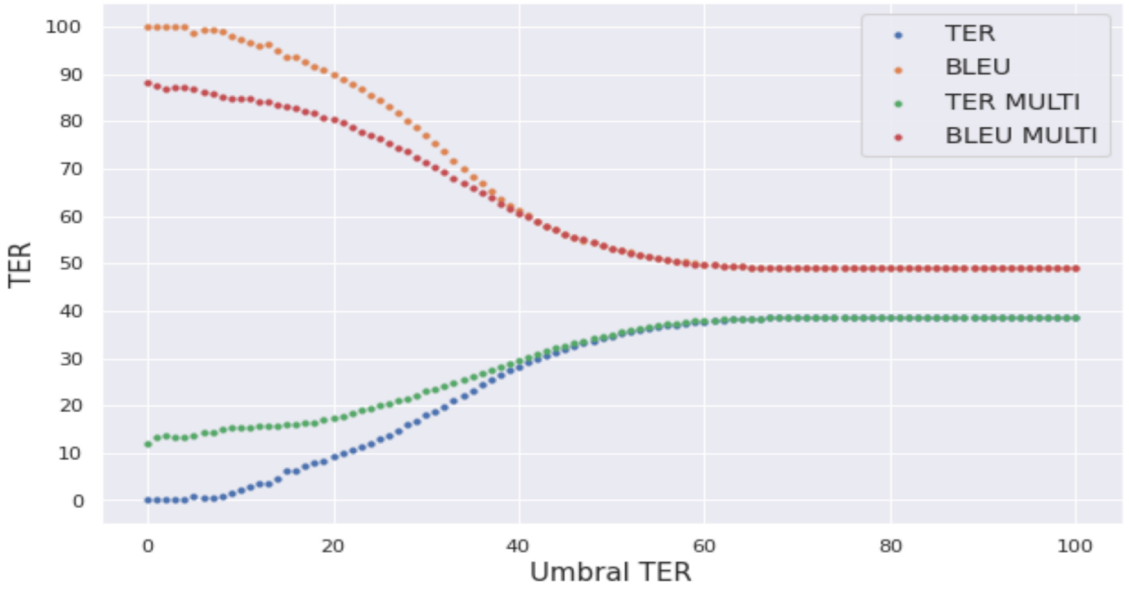

Figura 5.6: TER y BLEU medio respecto al umbral de TER escogido para determinar si una frase es buena o mala. Las líneas se corresponden a los modelos monolingüe y multiligüe para el par español-inglés

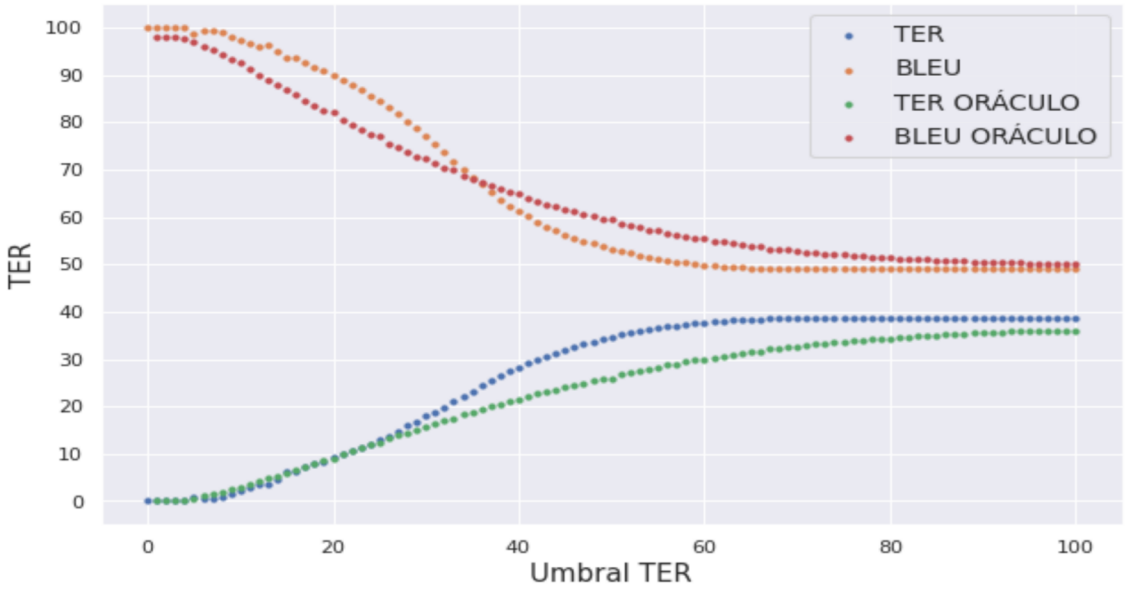

Figura 5.7: TER y BLEU medio respecto al umbral de TER escogido para determinar si una frase es buena o mala. Las líneas se corresponden al modelo monolingüe y al oráculo para el par español-inglés

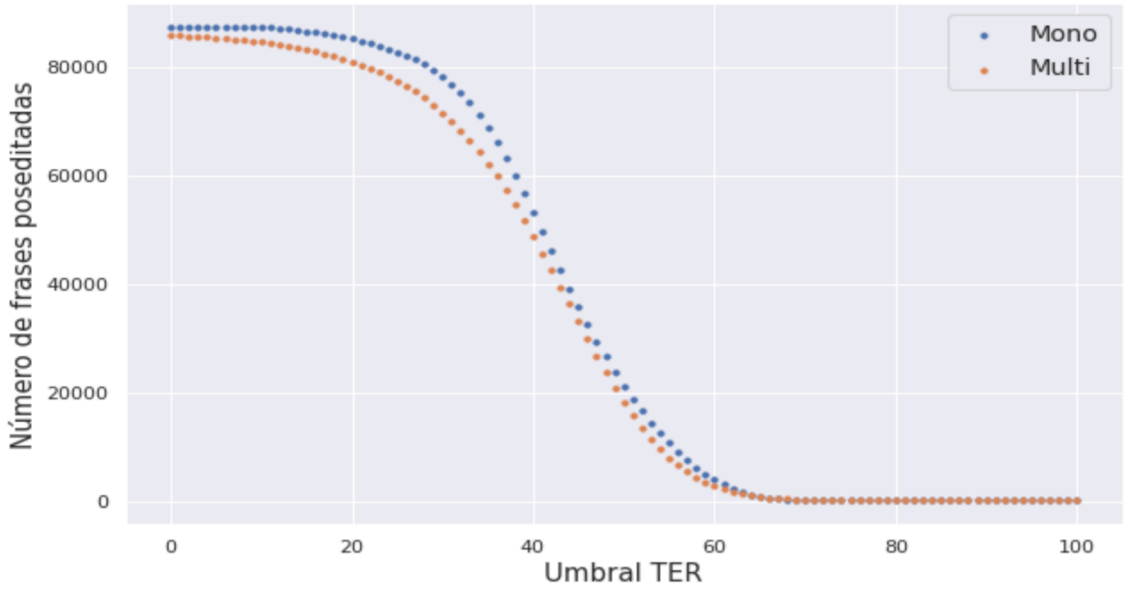

Figura 5.8: Nº de frases poseditadas respecto al umbral de TER escogido para determinar si una frase es buena o mala. Las líneas se corresponden al modelo monolingüe y al multilingüe para el par español-inglés

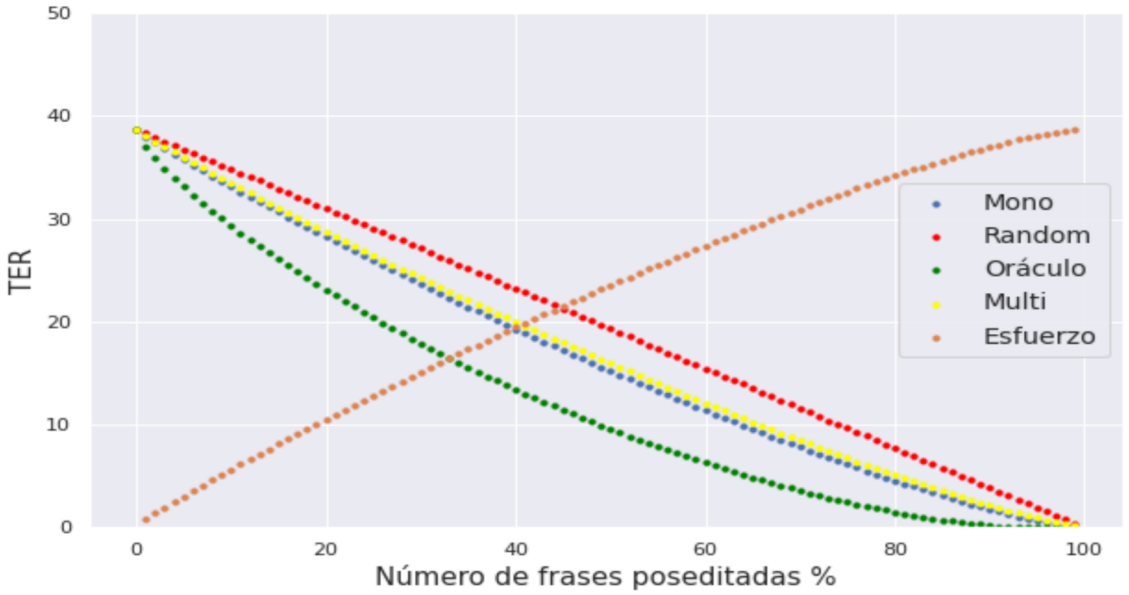

Figura 5.9: TER global según el porcentaje de frases poseditadas para el par de lenguas español-inglés

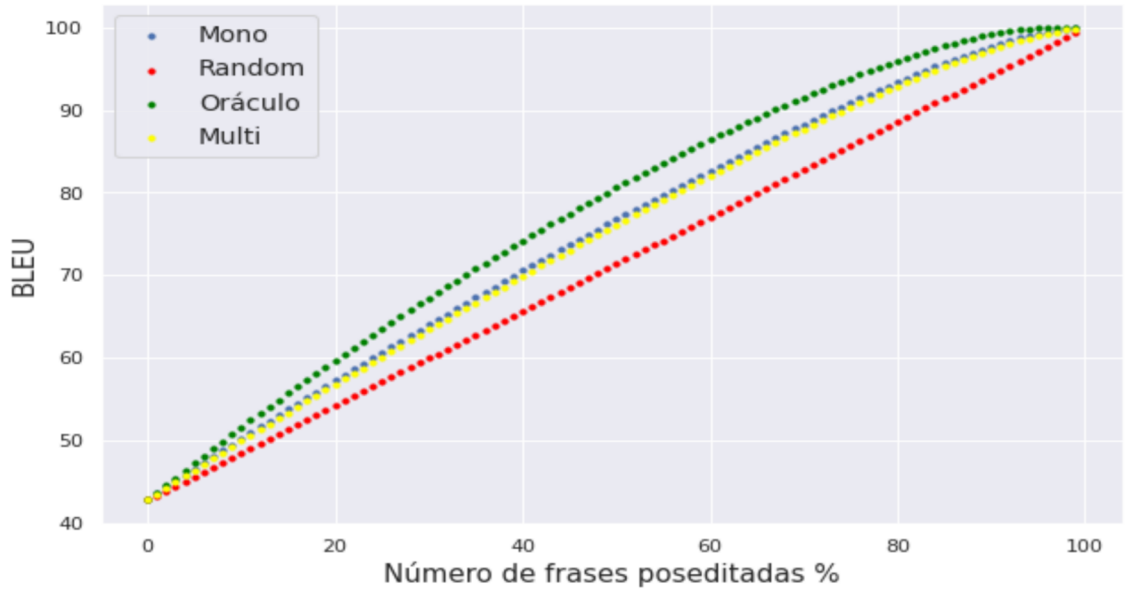

Figura 5.10: BLEU global según el porcentaje de frases poseditadas para el par de lenguas español-inglés

Las figuras anteriormente expuestas tienen la misma interpretación que las del modelo francés-inglés. La única diferencia que se puede observar respecto a los anteriores resultados es que en este caso el modelo monolingüe y el multilingüe difieren en menor medida, hecho que favorece el posible uso de un modelo multilingüe.

5.2 Análisis cuantitativo

En esta sección, se va a presentar el análisis cuantitativo, cuyo objetivo es obtener la correlación entre los valores generados por el modelo de evaluación automática y el conjunto de test anteriormente comentado.

La tabla que se va a mostrar a continuación refleja el coeficiente de correlación de Pearson, situado a la izquierda de cada celda, y el coeficiente de correlación de Spearman, situado a la derecha de cada celda. Cada fila representa un modelo distinto y cada columna representa el test del respectivo idioma.

	Francés	Español
Es-En	*	**0.364 - 0.447**
Fr-En	**0.404 - 0.477**	*
Es/Fr - En	0.365 - 0.414	0.327 - 0.381

Tabla 5.1: Resultados del análisis cuantitativo

A la vista de los resultados obtenidos en la tabla 5.1 se puede decir que el modelo multilingüe da peores resultados que los modelos monolingües. Además, se observa que el coeficiente de Spearman es más alto en todos los casos, cosa que nos puede indicar que hay algunos datos extremos que penalizan al coeficiente de Pearson.

5.3 Análisis cualitativo

En la siguiente sección, se va a presentar el análisis cualitativo que se ha realizado de los modelos de evaluación automática de la traducción automática. Para exponer el análisis cualitativo de los distintos modelos, se van a presentar ejemplos de predicciones para cada idioma.

5.3.1. Resultados Español-Inglés

En la figura 5.2 se pueden apreciar una serie de frases bien predichas por ambos modelos. En muchos casos se observa que las frases con aspecto legal, como es la primera de esta tabla, están bien predichas, esto puede estar relacionado con que hay bastantes frases que siguen el mismo patrón. En cuanto a las frases mal predichas vistas en la figura 5.3, se puede observar que para frases con un TER alto el modelo tiende a fallar, puede que en el entrenamiento se hayan visto pocas frases que tengan un TER alto.

Origen	Referencia	Traducción	Métricas
Recordando la resolución 56/6 de la Asamblea General, de 9 de noviembre de 2001, relativa al Programa Mundial para el Diálogo entre Civilizaciones, en la que la Asamblea reconocía la valiosa contribución que el diálogo entre civilizaciones podía aportar para que se conocieran y comprendieran mejor los valores comunes compartidos por toda la humanidad,	Recalling General Assembly resolution 56/6 of 9 November 2001 on the Global Agenda for Dialogue among Civilizations, in which the Assembly recognized the valuable contribution that dialog among civilizations can make to an improved awareness and understanding of the common values shared by all humankind,	Recalling General Assembly resolution 56/6 of 9 November 2001 on the Global Agenda for Dialogue amongst Civilizations, in which the Assembly recognised the valuable contribution that dialogue amongst civilisations can make to enhancing awareness and understanding of common values shared by all humankind,	TER = 17.77 Monolingüe = 23.22 Multilingüe = 24.40
Por su naturaleza misma, esos ENM podrían disuadir al país sede asociado de una .^{ev}asión de sus responsabilidades".	By their very nature, such MNAs have the potential to deter a "breakout"by the host partner.	By their very nature, such MNAs could deter the associated host country from ‘evading its responsibilities.’	TER = 58.823529 Monolingüe = 55.28 Multilingüe = 53.67

Tabla 5.2: Ejemplos de frases bien predichas para el par de idiomas español-inglés.

Origen	Referencia	Traducción	Métricas
La descentralización y autonomía efectivas de las iniciativas relativas a la gestión de los recursos naturales, al mismo tiempo que respetan los rasgos culturales específicos, deben ir acompañadas de un aumento de la capacidad de las partes de la sociedad civil interesadas, las ONG y las organizaciones comunitarias para formular, ejecutar y evaluar programas de desarrollo local integrado en el marco de los PAN, y también del fomento de las relaciones de asociación entre la comunidad, el sector privado y el sector público.	Effective decentralization and ownership of initiatives for natural resources management, while respecting cultural specificities, should go hand in hand with strengthening the capacities of civil-society stakeholders, NGOs and CBOs for the formulation, execution and evaluation of integrated local area development programs under the NAP, while expanding community/-private/public partnerships.	Effective decentralisation and autonomy of natural resource management initiatives, while respecting specific cultural features, should be accompanied by increased capacity of interested civil society actors, NGOs and community-based organisations to formulate, implement and evaluate integrated local development programmes within the framework of the NAPs, as well as fostering partnerships between the community, the private sector and the public sector.	TER = 97.91 Monolingüe = 42.92 Multilingüe = 42.238
a) Trabajos nocturnos industriales.	(a) Night work in industry.	(a) Industrial nite work.	TER = 80 Monolingüe = 38.86 Multilingüe = 37.43

Tabla 5.3: Ejemplos de frases mal predichas para el par de idiomas español-inglés.

5.3.2. Resultados Francés-Inglés

En la tabla 5.4 se pueden apreciar una serie de frases bien predichas por ambos modelos. En muchos casos se observa que las frases con un TER cercano a 50 tienden a ser mejor predichas que las que tienen valores más extremos, esto puede pasar porque había bastantes frases con un TER cercano a 50 en el entrenamiento. En cambio, también se pueden apreciar frases que están perfectas y que el modelo acierta. Aunque esto suele pasar para frases de poca longitud. En la figura 5.5, se puede observar que para frases con un TER de 0 y que tienen una longitud mayor el modelo empieza a fallar. Además pasa lo mismo que en el anterior modelo, las frases con un TER cercano a 100 tienden a estar mal predichas.

Origen	Referencia	Traducción	Métricas
En ce qui concerne le document de travail proposé par Mme Chung, M. Salama fait observer qu'il avait présenté récemment un document de travail sur un thème analogue et avait offert de collaborer avec Mme Chung à l'élaboration du document qu'elle proposait, éventuellement en actualisant son propre document de travail.	In relation to Ms. Chung's suggested working paper, Mr.Salama noted that he had recently submitted a working paper on a similar topic and offered to work with Ms. Chung in the preparation of her paper, possibly by updating his first working paper.	With regard to the working paper proposed by Ms. Chung, he noted that he had recently presented a working paper on a similar theme and offered to collaborate with Ms. Chung in the preparation of the paper she proposed, possibly updating her own working paper.	TER = 45.23 Monolingüe = 42.96 Multilingüe = 42.43
Lundi 31 octobre 2005	Monday, 31 October 2005	Monday, 31 October 2005	TER = 0 Monolingüe = 0 Multilingüe = 0

Tabla 5.4: Ejemplos de frase bien predichas para el par de idiomas francés-inglés.

Origen	Referencia	Traducción	Métricas
État au 28 février 2005 des contributions à l'Opération des Nations Unies au Burundi (ONUB) (En dollars des États-Unis)	Status of contributions to the United Nations Operation in Burundi (ONUB) as at 28 February 2005	Status of contributions to the United Nations Operation in Burundi (ONUB) as at 28 February 2005	TER = 0 Monolingüe = 27.46 Multilingüe = 27.11
Organigramme et répartition des postes pour l'exercice biennal 2006-2007	Organizational structure and post distribution for the biennium 2006-2007	Organizational structure and post distribution for the biennium 2006-2007	TER = 0 Monolingüe = 38.96 Multilingüe = 37.37
Deux groupes d'experts ont été créés, l'un - placé sous les auspices de l'OMS - consacré aux conséquences de l'accident pour la santé des populations, et l'autre - placé sous les auspices de l'AIEA - consacré aux conséquences environnementales.	Two expert groups were created, one addressing the consequences of the accident for human health, operating under WHO auspices, and another addressing the environmental impact, operating under IAEA auspices.	Two expert groups were created, one - under the auspices of the WHO - dedicated to the consequences of the accident for the health of the people, and the other - under the auspices of the IAEA - dedicated to the environmental consequences.	TER = 103.44 Monolingüe = 56.88 Multilingüe = 58.34

Tabla 5.5: Ejemplos de frases mal predichas para el par de idiomas francés-inglés.

CAPÍTULO 6
Conclusiones

En este último apartado del proyecto, se va a exponer la evaluación de los objetivos que se nombraron al comienzo del trabajo. Además, se va a definir las diferentes líneas que se podrían seguir para el desarrollo futuro del presente trabajo.

6.1 Evaluación de los objetivos

La principal contribución de este trabajo ha sido aplicar un modelo de estimación de calidad a dos idiomas poco explorados en la literatura científica. Ya que la mayoría de artículos científicos se centran en los idiomas seleccionados para la WMT, donde se encuentra el alemán o el ruso entre ellos.

Para ello, se ha presentado una herramienta para estimar el esfuerzo del traductor en la posedición de una frase traducida. Para obtener este esfuerzo automáticamente se han entrenado una serie de modelos para dos pares de lenguas: español-inglés y francés-inglés. Tras observar los resultados se puede concluir que los modelos monolingües tienen mayor precisión que el multilingüe, aunque habrá que tener en cuenta si se prefiere almacenar dos modelos monolingües a cambio de ese aumento de precisión.

Además, se han realizado una serie de experimentos donde se ha visto como se comporta el modelo. Un posible escenario que se ha planteado es aquel donde los usuarios pueden indicar un umbral para seleccionar las frases como buenas o malas traducciones, de esta manera, pueden tomar los siguientes caminos:

- Mantener las frases traducidas sin post-edición ahorrando al traductor tiempo y después poseditar o eliminar las frases consideradas como mal traducidas.

- Eliminar las malas traducciones ahorrando tiempo de post-edición de las malas traducciones (tan malas que es mejor traducir sin la traducción automática) y poseditar sólo las buenas traducciones.

Otra enfoque distinto sería aquel en el que nosotros tenemos una empresa de traducción y queremos añadir este servicio a nuestra plataforma de posedición. Se podría mostrar el valor predicho a cada frase traducida y así el traductor contratado podría ver más fácilmente el error previsto de la frase. Por ejemplo, en una plataforma de posedición es posible añadir este método dando el valor predicho a cada frase traducida. Este valor predicho es similar a otra medida común utilizada en el marco de la traducción conocida como umbral de coincidencia difusa que da un grado de coincidencia entre un segmento del documento de origen y un segmento de la memoria de traducción.

6.2 Trabajo futuro

Una primera línea futura podría ser aumentar la cantidad y la calidad de los datos de entrenamiento. Ya que tras la limpieza se ha entrenado con menos de un millón de datos y estos modelos tan pesados suelen requerir una gran cantidad de datos para generalizar.

Otra línea sería entrenar modelos para más idiomas, ya que sobre este tema hay pocos experimentos con idiomas que no son los vistos en la WMT. También se podría probar como funciona el modelo multilingüe para una mayor cantidad de idiomas, ya que en ese caso puede que la pérdida de precisión este justificada con el ahorro de espacio en memoria de los modelos.

En relación a la anterior línea comentada, se podría hacer un estudio comparando los gastos causados por el coste espacial de los modelos monolingües con los gastos ocasionados por el coste de precisión de los modelos multilingües. Y así estudiar como varían los beneficios a medida que aumentamos el número de modelos e idiomas.

Por último, una mejora a este trabajo sería utilizar métricas más confiables para calcular la calidad de los datos. Ya que el TER es una métrica la cual tiene correlación con el esfuerzo de posedición pero a veces falla por falta de flexibilidad. Por ejemplo, el hter o el direct assessement son medidas que necesitan de un componente humano pero luego son más confiables a la hora de evaluar si una traducción está bien o está mal.

Bibliografía

[1] Winfield S Bennett and Jonathan Slocum. The lrc machine translation system. *Computational linguistics*, 11(2-3):111–121, 1985.

[2] John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. Confidence estimation for machine translation. In *Coling 2004: Proceedings of the 20th international conference on computational linguistics*, pages 315–321, 2004.

[3] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311, 1993.

[4] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, 2014.

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[6] Vincent J Della Pietra. The mathematics of statistical machine translation: Parameter estimation. *Using Large Corpora*, page 223, 1994.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[8] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, 2002.

[9] Michael Gamon, Anthony Aue, and Martine Smets. Sentence-level mt evaluation without reference translations: Beyond language modeling. In *Proceedings of the 10th EAMT Conference: Practical applications of machine translation*, 2005.

[10] Yifan He, Yanjun Ma, Josef van Genabith, and Andy Way. Bridging smt and tm with translation recommendation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 622–630, 2010.

[11] Julia Ive, Frédéric Blain, and Lucia Specia. Deepquest: a framework for neural-based quality estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3146–3157, 2018.

[12] Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André FT Martins. Openkiwi: An open source framework for quality estimation. *arXiv preprint arXiv:1902.08646*, 2019.

[13] Hyun Kim and Jong-Hyeok Lee. Recurrent neural network based translation quality estimation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 787–792, 2016.

[14] Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, 2017.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

[17] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

[18] E Matthew. Peters, mark neumann, mohit iyyer, matt gardner, christopher clark, kenton lee, luke zettlemoyer. deep contextualized word representations. In *Proc. of NAACL*, 2018.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002.

[20] Christopher Quirk. Training a sentence-level machine translation confidence measure. In *LREC*. Citeseer, 2004.

[21] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. 2018.

[22] Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. Transquest: Translation quality estimation with cross-lingual transformers. *arXiv preprint arXiv:2011.01536*, 2020.

[23] Tharindu Ranasinghe, Constantin Orăsan, and Ruslan Mitkov. Semantic textual similarity with siamese neural networks. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1004–1011, 2019.

[24] M. Snover, Bonnie J Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, 2006.

[25] Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André FT Martins. Findings of the wmt 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, 2021.

[26] Lucia Specia, Gustavo Paetzold, and Carolina Scarton. Multi-level translation quality prediction with quest++. In *Proceedings of ACL-IJCNLP 2015 system demonstrations*, pages 115–120, 2015.

[27] Lucia Specia, Dhwaj Raj, and Marco Turchi. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50, 2010.

[28] Lucia Specia, Kashif Shah, José GC De Souza, and Trevor Cohn. Quest-a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, 2013.

[29] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, volume 27, page 3104–3112, Cambridge, MA, USA, 2014.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010, Red Hook, NY, USA, 2017.

[31] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China, 2021.

APÉNDICE A
OBJETIVOS DE DESARROLLO SOSTENIBLE

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenible	**Alto**	**Medio**	**Bajo**	**No procede**
ODS 1. **Fin de la pobreza.**				X
ODS 2. **Hambre cero.**				X
ODS 3. **Salud y bienestar.**				X
ODS 4. **Educación de calidad.**	X			
ODS 5. **Igualdad de género.**			X	
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**	X			
ODS 9. **Industria, innovación e infraestructuras.**	X			
ODS 10. **Reducción de las desigualdades.**		X		
ODS 11. **Ciudades y comunidades sostenibles.**				X
ODS 12. **Producción y consumo responsables.**				X
ODS 13. **Acción por el clima.**				X
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**		X		
ODS 17. **Alianzas para lograr objetivos.**		X		

Reflexión sobre la relación del TFG con los ODS y con el/los ODS más relacionados.

El presente trabajo fin de grado está relacionado con varios de los objetivos de desarrollo sostenible (ODS). Desde el inicio de los tiempos, la comunicación ha supuesto un intercambio constante de conocimiento y recursos entre distintas sociedades, con el objetivo de socializar y realizar negociaciones, lo que ha supuesto el nacimiento de las sociedades primitivas y la estabilidad y avance de las sociedades de las posteriores. Sin embargo, hoy en día el objetivo principal es tener la capacidad de establecer una comunicación efectiva y sencilla teniendo en cuenta la enorme cantidad de diferentes idiomas y sus características propias. Por ello, la traducción cada vez juega un papel más y más en el desarrollo social, cultural y económico. Debido a esto se hace obvio la creciente necesidad de tener un acceso rápido, sencillo y globalizado a traducciones de calidad. Además, ello supone un impulso en la exportación de productos y conocimiento, algo esencial en el desarrollo de las relaciones entre los países. La traducción automática es una solución barata y rápida que permite el acceso a traducciones a todo el mundo, creando puentes de comunicación entre personas de todo el mundo.

Por último, el presente trabajo supone un gran impacto en la industria, en concreto en del lenguaje, e innovación, debido a que consigue mejorar las traducciones, tanto del francés al inglés como del español al inglés, con el empleo de técnicas de evaluación automática, lo que posibilita un mayor entendimiento y una mejor comunicación. Además, el crecimiento económico también se ve influenciado, ya que al conseguir mejores traducciones, se consigue ahorrar en posibles post-procesos de corrección humana de dichas traducciones.