

The Bible as a Corpus for Language Technologies

GERARDO E. SIERRA MARTÍNEZ 

Instituto de Ingeniería - UNAM

gsierram@iingen.unam.mx

GEMMA BEL ENGUIX 

Instituto de Ingeniería - UNAM

gbele@iingen.unam.mx

Abstract: *This work aims to create an aligned corpus of eleven Spanish translations of the Bible to advance computational linguistics in Spanish. The use of this corpus is essential for applications such as paraphrase detection, lexical grouping identification, and language model evaluation for search systems. In this way, the study covers various aspects of natural language processing, including similarity, lexical extraction, and bias analysis, with the goal of promoting the development of language technologies in Spanish.*

KEYWORDS: LINGUISTIC CORPUS; NATURAL LANGUAGE PROCESSING; PARAPHRASE DETECTION; SEMANTIC CLUSTERING; QUESTION-ANSWERING SYSTEMS

RECEPTION: 16/08/2024

ACCEPTANCE: 22/08/2024

La Biblia como corpus para tecnologías del lenguaje

GERARDO E. SIERRA MARTÍNEZ 

Instituto de Ingeniería - UNAM

gsierram@iingen.unam.mx

GEMMA BEL ENGUIX 

Instituto de Ingeniería - UNAM

gbele@iingen.unam.mx

Resumen: El presente trabajo tiene como objetivo presentar un corpus alineado de once traducciones de la Biblia al español, para avanzar en la lingüística computacional en español. El uso de este corpus resulta fundamental en aplicaciones como la detección de paráfrasis, la identificación de grupos léxicos y la evaluación de modelos de lenguaje para sistemas de búsqueda. De esta manera, este trabajo abarca estudios en procesamiento de lenguaje natural, como la similitud, extracción léxica, y análisis de sesgos, con el propósito de impulsar el desarrollo de tecnologías del lenguaje en español.

PALABRAS CLAVE: CORPUS LINGÜÍSTICO; PROCESAMIENTO DE LENGUAJE NATURAL; DETECCIÓN DE PARÁFRASIS; AGRUPACIÓN SEMÁNTICA; SISTEMAS PREGUNTA RESPUESTA

RECEPCIÓN: 16/08/2024

ACEPTACIÓN: 22/08/2024

INTRODUCCIÓN

La Biblia es el texto más leído en la historia de Occidente y el más traducido del mundo, ya sea en versiones completas o parciales, como el Nuevo Testamento. Aunque es difícil tener un inventario acabado, Wikipedia afirma que hay traducciones completas en, al menos, 700 idiomas (Wikipedia, s.f.).

La Biblia es una compilación de un número variable de libros, que se dividen en dos grupos principales: Antiguo Testamento y Nuevo Testamento. Hay una diferencia en el canon del Antiguo Testamento para las tradiciones protestante (39 libros), católica (46 libros) y ortodoxa (51 libros). El Nuevo Testamento incluye 27 libros para las tres principales tradiciones cristianas.

En general, las versiones de la Biblia en distintas lenguas tienen el objetivo de capturar el mensaje exacto que el texto quiere transmitir. Sin embargo, existen diferentes enfoques de traducción (Klein *et al.*, 2017). Por un lado, la perspectiva *formalmente equivalente* intenta ser fiel al texto, al asignar una traducción literal palabra por palabra del original, prefiriendo la precisión a la claridad o inteligibilidad para los lectores actuales. Por otro lado, las traducciones dinámicamente o *funcionalmente equivalentes* se preocupan más por la claridad de las oraciones, al reformular los textos e interpretar su significado en contexto. Entre estas dos posiciones radicales se encuentra un enfoque óptimamente equivalente, que busca alcanzar la mejor traducción posible para cada pasaje, favoreciendo uno u otro planteamiento principal.

Muchos conjuntos de datos incluyen la Biblia en sus repositorios. Para el caso de las tecnologías del lenguaje, por sus numerosas traducciones y versiones, ésta se utiliza como un recurso valioso para un vasto campo de aplicaciones en áreas como la traducción automática, el procesamiento de lenguaje natural (PLN) y los estudios lingüísticos comparativos.

Por ejemplo, las múltiples versiones de la Biblia en diferentes idiomas resultan útiles para entrenar modelos de traducción automática, ya que la comparación de textos paralelos facilita la realización de estudios comparativos tendientes a analizar cómo se interpretan y expresan los mismos conceptos en diferentes lenguas y culturas. Este repertorio de multiplicidad de formas para un mismo sentido favorece el estudio de fenómenos como el cambio lingüístico, la variabilidad léxica y las diferencias sintácticas. Asimismo, debido

a que su disposición en capítulos y versículos permite crear bases de datos estructuradas, puede utilizarse para desarrollar y evaluar sistemas de búsqueda y recuperación de información. Adicionalmente, las distintas versiones y traducciones de la Biblia ofrecen un campo fértil para la estilometría, el estudio de estilos literarios mediante métodos cuantitativos, y –en general– para tareas de PLN, como el análisis de sentimiento, el etiquetado de partes del discurso y el reconocimiento de entidades nombradas.

Lo anterior, justifica el desarrollo del corpus que se describe en el presente escrito, mismo que consta de once traducciones de la Biblia a la lengua española; éstas abarcan diferentes siglos y regiones geográficas. Las once versiones se han alineado en función de la distribución de los versículos y la equivalencia de sentido. Se presentan también tres aplicaciones de este recurso: a) *detección de paráfrasis*, b) *identificación de grupos léxicos*, y c) evaluación de los *modelos grandes de lenguaje* (LLM, por sus siglas en inglés), para búsqueda de información y sistemas preguntas-respuesta.

CORPUS PARALELO DE BIBLIAS EN ESPAÑOL

Nuestro corpus contiene once versiones diferentes de la Biblia en español, cuyas ediciones fueron creadas en diferentes períodos, aunque la mayoría se publicaron o editaron durante el siglo xx en distintos países de habla hispana (Sierra *et al.*, 2024). A continuación, se enlistan y describen brevemente tales publicaciones.

La Biblia, que es, los Sacros Libros del Viejo y Nuevo Testamento (OSO). Esta primera Biblia fue traducida por Casiodoro de Reina (OSO) y se publicó por primera vez en Suiza en 1569. La traducción se basó principalmente en el texto hebreo masorético, pero también tomó en consideración otras versiones en latín, griego e, incluso, judeo-español.

La Santa Biblia (REV). Versión de Cipriano de Valera de 1602; consiste en una revisión y actualización de la traducción de Casiodoro de Reina.

La Sagrada Biblia (PET). En 1823, Torres Amat (1772-1847) lanzó su traducción de la Biblia conocida como PET. Esta traducción se basa en gran medida en el trabajo del jesuita Petisco, quien había realizado una versión hacia finales del siglo xviii. A diferencia de las ediciones anteriores, que fueron

traducidas directamente del hebreo, arameo y griego, la de Torres Amat se basa principalmente en la *Vulgata*.¹

La Sagrada Biblia (JNM). Nuestro corpus tiene un vacío de cien años entre el de Torres Amat y el siguiente texto, que es una traducción latinoamericana hecha en Chile por Wilhelm Jünemann. Este autor tradujo tanto el Nuevo Testamento como el Antiguo; no obstante, el primero salió a la luz en 1928, mientras que el Antiguo Testamento se publicó más recientemente, en 1992. Este último volumen fue traducido del griego koiné y se completó antes de 1938, el año de la muerte de Jünemann. A pesar de que no se editó antes, el volumen debe datarse, a todos los efectos, en la década de 1920 o 1930.

La Sagrada Biblia (NAC). La Sagrada Biblia de Nacar y Colunga (NAC) se publicó en España en 1944 y ocupa una posición prominente entre las versiones católicas, tras haberse difundido a través de más de treinta ediciones. Esta traducción se deriva de los textos originales en hebreo, arameo y griego. En el prólogo, los autores expresan su intención de mantener la fidelidad al original, aunque señalan que no necesariamente siguen servilmente sus letras; pues si bien se guían por las palabras del texto, atienden principalmente al sentido de las frases. A pesar de que la versión original incluye 73 libros canónicos, nuestro corpus comprende exclusivamente 66 libros.

La Biblia de Jerusalén (JER). Como un proyecto de traducción colaborativa, llevada a cabo por la *École biblique et archéologique française* de Jérusalem, el objetivo principal de este esfuerzo de traducción es facilitar la exégesis bíblica. Con principios de crítica histórica, se incluyen en el texto comentarios de esta índole, así como lingüísticos, los cuales proporcionan valiosos análisis e información. Desde su publicación inicial en 1966, la Biblia de Jerusalén ha experimentado varias revisiones. Sin embargo, la versión incorporada a nuestro corpus es la edición original de 1966. La traducción se basa en los textos originales en hebreo, arameo y griego, y sigue las directrices empleadas en la traducción francesa.

La Santa Biblia (EMN). Evaristo Martín Nieto dirigió y lanzó en 1961 esta versión en español realizada por un equipo de traductores que buscó

¹ Traducción de la Biblia al latín, según la tradición realizada por San Jerónimo y Santa Paula en el siglo IV.

proporcionar una representación fiel de los textos originales, adhiriéndose estrechamente a la redacción y estructura de los idiomas fuente.

La Nueva Biblia - Edición pastoral para Latinoamérica (LAT). Versión orientada específicamente a los lectores latinoamericanos. Se imprimió por primera vez en 1972, después del Concilio Vaticano II, tiene una ideología progresista en el marco de la Iglesia Católica. Fue considerada una herramienta de la teología de la liberación. Es muy popular entre la gente de América Latina. Está traducida del hebreo, arameo y griego con un estilo de equivalencia funcional.

Nueva Biblia Española (ESP). Esta versión también es una traducción con un enfoque de equivalencia funcional a partir de los textos originales; realizada por Alonso Shöckel y Juan Mateos, vio la luz en 1975. Con el propósito de hacer el texto más accesible para las personas y ayudarlas a conectar con el mensaje de la Biblia, acorde con su propio contexto cultural y lingüístico, los traductores emplearon un estilo libre que utilizaba expresiones y lenguaje que contribuyera a cerrar la brecha cultural entre el mundo moderno y el contexto histórico en el que se escribieron los textos bíblicos.

La Biblia (SER). Bajo la coordinación de Serafín de Ausejo, en 1975 se publicó esta versión regida por los principios de equivalencia formal. Realizada por un equipo de traductores, tuvo como objetivo mantener una representación fiel de los textos originales, enfatizando la precisión en la traducción de los idiomas fuente. A la postre, es considerada la versión en lengua española más literal.

La Biblia de las Américas (AME).² El título *Biblia de las Américas* refleja adecuadamente el público y el propósito de esta edición, ya que está dirigida, principalmente, a individuos de habla hispana que residen en el continente americano. La traducción en sí es resultado de un esfuerzo colaborativo que involucró a individuos de diversas corrientes protestantes, quienes trabajaron a partir de los textos originales. Esta versión se publicó por primera vez en 1986, lo que la convierte en una adición relativamente reciente dentro del conjunto de traducciones bíblicas disponibles en español.

² En adelante, se identificará cada versión de la Biblia con el código alfabético de tres caracteres que se muestra, entre paréntesis y tras su nombre, en este listado: OSO, REV, PET, JNM, NAC, JER, EMN, LAT, ESP, SER Y AME (*N. del Ed.*)

El principal criterio para la selección de los libros fue que las ediciones estuvieran disponibles íntegramente en formato digital, con posibilidad de ser descargadas y convertidas en documentos de texto. Al conformar el corpus, se excluyeron subtítulos y notas, con la finalidad de conservar únicamente el contenido de los versículos.

Dado que el número de libros en las diversas ediciones varía de 64 a 73, se suscitaron discrepancias sobre la inclusión o exclusión de algunos. Otro aspecto problemático para establecer paralelismos lo constituyen las diferencias en el número de capítulos dentro de un mismo libro, así como en el número de versículos dentro de cada capítulo, diferencias que se deben a factores como enfoques de traducción diferentes, desafíos en la interpretación del texto, empleo de textos fuente diferentes o divergencias ideológicas entre los traductores. Por ejemplo, mientras los libros de Judas o Lamentaciones muestran coincidencias en la cantidad de versículos y contenido en todas las ediciones, el libro de Salmos presenta diferencias significativas, ya que en éste se observaron hasta diez variantes.

Las disparidades representaron un reto importante al intentar alinear de manera paralela el contenido de las ediciones que conforman el corpus, ya que –como se ha dicho– éste no siempre coincide y el número de versículos es variable. Para facilitar la comparación de los textos de las diferentes ediciones, se utilizó un código creado en el lenguaje de programación Python, que ha permitido etiquetar los versículos en cada libro y categorizar los capítulos en función de las diferencias en la numeración de los versos. Esto facilitó, finalmente, obtener una alineación automática, tras la cual se llevó a cabo una revisión manual.

Como resultado del proceso anterior, se pudo determinar cuál fue la versión con el mayor número de versículos para cada segmento de la obra, misma que sirvió como guía para el libro o capítulo en cuestión, de manera que ninguna edición en particular se empleó como referencia de alineación para toda la obra. Los capítulos o versículos omitidos en una versión particular fueron marcados y etiquetados como **texto omitido**.³

³ Para facilitar la lectura de los diferentes elementos de esta nota, se establecieron los siguientes usos tipográficos: Negritas (para resaltar las etiquetas empleadas en el proceso de marcación del texto). Subrayados (para destacar un elemento dentro del ejemplo (versículo), marca que corresponde al analista y no a la fuente de origen (*N. del Ed.*))

También hubo casos en los que faltaba el último versículo de un capítulo en una versión, pero el contenido omitido, en realidad, existía al principio del capítulo siguiente. Ante tales situaciones, se reorganizó la información para presentarla en la posición que ocupaba en otras versiones.

En (1), ejemplo procedente de Lucas 17: 36, la Biblia OSO (a) contiene una información adicional que no está presente en otras versiones, como en ESP (b) o en LAT (c), lo que da como resultado un hueco en el contenido:

- (1) a. Dos mujeres estarán moliendo juntas; la una será tomada, y la otra será dejada. Dos estarán en el campo; el uno será tomado, y el otro será dejado (OSOLUC017036).⁴
- b. Estarán dos moliendo juntas, a una se la llevarán y a la otra la dejarán (ESPLUC017036).
- c. Dos mujeres estarán moliendo juntas, pero una será llevada y la otra dejada (LATLUC017036).

En total, nuestro corpus comprende 754 libros, los cuales abarcan una colección voluminosa de 7 970 269 palabras. Para facilitar el acceso al mismo y la recuperación de sus materiales con fines académicos, la colección está disponible en la plataforma en línea GitHub,⁵ así como en nuestro gestor de corpus GECO.⁶

El corpus alineado de traducciones de la Biblia en español tiene múltiples aplicaciones en tareas de procesamiento de lenguaje natural (PLN). A continuación, se detallan algunas líneas de investigación que se han llevado a cabo usando este recurso como base textual. En concreto, se explica cómo se

⁴ Los caracteres alfanuméricos que se muestran entre paréntesis tras los ejemplos corresponden a los datos de identificación del versículo y la fuente exacta de dónde fueron extraídos. El código se forma de la siguiente manera: las tres primeras letras son de la versión de la Biblia, las tres siguientes letras corresponden al libro (LUC para Lucas, por ejemplo), los siguientes tres números refieren al capítulo y los últimos tres al versículo.

⁵ <https://github.com/GIL-UNAM/SpanishParaphraseCorpora/tree/main/Biblias>

⁶ <http://www.geco.unam.mx/geco3/proyecto/CPBE>

implementan las aplicaciones en las áreas de detección de paráfrasis, agrupación semántica y evaluación de modelos grandes de lenguaje (LLM).

APLICACIÓN PARA DETECCIÓN DE PARÁFRASIS

Se considera que dos textos, o dos oraciones, son paráfrasis cuando son semánticamente equivalentes, independientemente de la causa de dicha equivalencia (Das y Smith, 2009). La detección automática de paráfrasis es una tarea de PLN que sirve de base para múltiples aplicaciones, tales como detección de plagio, sistemas de pregunta-respuesta o traducción automática (Kong *et al.*, 2020).

Se pueden categorizar las paráfrasis teniendo en cuenta el tipo y la profundidad de variantes entre dos textos que comparten un significado (Mota-Montoya *et al.*, 2016), lo que da lugar a dos clases principales. En la denominada *paráfrasis baja* se dan preferentemente cambios de tipo léxico, como la sustitución de una palabra por un sinónimo o hiperónimo, entre otros. La *paráfrasis alta* emplea, además, otro tipo de recursos, como cambios en el orden, estrategias de reescritura sintáctica y modificaciones de tipo discursivo, como el reordenamiento lógico del texto.

Para modelar la detección de paráfrasis de manera automática se necesitan corpus etiquetados sobre los que poder entrenar algoritmos de aprendizaje automático. Entre los más usados por los investigadores se encuentran el Microsoft Research Paraphrase Corpus (MRPC) (Dolan y Brockett, 2005), el Quora Question Pairs (QQP) y el PAWS (Zhang *et al.*, 2019).

El MRPC consta de 5801 pares de oraciones escogidos de sitios web dedicados a las noticias. Los pares se muestran etiquetados con 0 (que se aplica a las no paráfrasis) y 1 (para las paráfrasis). Por otra parte, el QQP contiene casi 800 mil pares de oraciones etiquetados también de manera binaria; éstos fueron extraídos de la conocida red social Quora, la cual tiene por objetivo compartir conocimiento en formato pregunta/respuesta.⁷ Por último, PAWS es una colección de 65 000 pares de oraciones etiquetadas de Wikipedia.⁸

⁷ <https://quoradata.quora.com>

⁸ <https://dumps.wikimedia.org>

Es importante hacer notar que todos los corpus de este tipo se encuentran en inglés. En realidad, con anterioridad al nuestro, no existía un recurso en español que pudiera servir de referencia para el entrenamiento de modelos de aprendizaje automático para la detección de paráfrasis.

Entre las posibles estrategias para elaborar corpus de paráfrasis, una muy utilizada es tomar diferentes traducciones de un texto fuente hacia una misma lengua. Ésta es una práctica aconsejable específicamente para libros que hayan sido bastante versionados, en diferentes épocas y geografías, a la lengua destino. A pesar de sus evidentes ventajas, esta forma de recopilación tiene el problema de que algunas versiones no ofrecen una traducción suficientemente adherida a la literalidad del texto, sino que reformulan muy libremente sus ideas generales.

Como puede advertirse, la Biblia es un recurso muy valioso para llevar a cabo todo tipo de estudios lingüísticos: comparativos, cuantitativos, diacrónicos, etc. Entre los corpus bíblicos existentes en diferentes lenguas, destacan el conformado por la colección anotada de los 66 libros del canon protestante de la Biblia en ocho lenguas, y por los del Nuevo Testamento en doce lenguas, elaborado por Resnik *et al.* (1999); el corpus de traducciones del Nuevo Testamento en lenguas Indoeuropeas, construido por el proyecto PROIEL (Pragmatic Resources in Old Indo-European Languages), con información morfológica, sintáctica y discursiva (Eckhoff *et al.*, 2018); el corpus multilingüe de Mayer and Cysouw (2014), compuesto por 994 traducciones alineadas, las cuales abarcan 76 familias lingüísticas; el corpus alineado de 4272 traducciones de la Biblia en 1611 lenguas (incluidas 27 versiones en inglés), y el corpus multilingüe de Christodouloupoulos and Steedman (2015), elaborado por McCarthy *et al.* (2020), el cual incluye traducciones totales o parciales de la Biblia en cien lenguas, por lo que se trata de una de las fuentes textuales más extensamente usadas en PLN.

No son tan frecuentes los corpus de traducciones de la Biblia a una sola lengua y, sin embargo, hay algunos casos reseñables. En español, se cuenta con el trabajo de Enrique-Arias y Pueyo Mena (2008), quienes prepararon un corpus con 14 traducciones alineadas de la Biblia en español medievales; mientras que en inglés, es posible acceder a las 27 versiones del corpus de McCarthy (2020).

Dicho lo anterior, el corpus alineado de traducciones de la Biblia en español resulta un recurso fundamental para el tratamiento de la paráfrasis. Diversos

aspectos confluyen para ello. En primer lugar, es sabido que los traductores de la Biblia adoptan los principios de la traducción formalmente equivalente, o bien funcionalmente equivalente. A pesar de encontrarse en extremos opuestos de los enfoques de traducción, en ambas perspectivas el traductor intenta respetar el texto original de la manera más estricta, ya que se considera un texto sagrado. Por lo tanto, la libertad del traductor no pone en peligro, en principio, la equivalencia del texto.

En segundo lugar, la división de los textos bíblicos en versículos es una característica que facilita enormemente el trabajo en un gran número de tareas de PLN. A pesar de la gran ventaja que supone esta numeración, la alineación paralela automática que podría derivarse de ella no es perfecta. Por distintas razones, ya sean de carácter pragmático, ideológico o filológico, los versículos no siempre se alinean de forma perfecta en todas las versiones. Pero, en general, éstos suponen una gran ayuda para el investigador. Como ya se ha explicado, en el corpus que presentamos, se ha llevado a cabo un trabajo pormenorizado de revisión de la alineación, lo que permite experimentar a partir de pares o grupos de versículos con identidad semántica.

Computacionalmente, el problema de la detección de paráfrasis no es otro sino el consistente en evaluar la similitud semántica entre dos entradas. Para atacar el problema desde el aprendizaje automático, se necesita un conjunto de datos etiquetado, el preprocesamiento y vectorización de los textos y, finalmente, unos algoritmos de clasificación, en este caso, la arquitectura de una red neuronal siamesa (Reimers y Gurevych, 2019). A continuación, se reporta el proceso de elaboración de un *dataset*⁹ para la detección automática de paráfrasis a partir del corpus alineado de biblias y los parámetros generales de configuración de los experimentos realizados con este recurso.

Como se ha mencionado, para determinar si hay o no paráfrasis se necesita una doble entrada de dos textos que puedan ser comparados. No es necesario que el algoritmo defina si uno de ellos es la fuente y el otro la copia, sino, únicamente, si ambas entradas son equivalentes o no. Como el corpus con el que trabajamos está completamente alineado, la elaboración del *dataset* es sencilla. En este artículo, se toman como textos de inicio dos libros, 1 Corintios y Génesis, y dos traducciones, la Biblia Latinoamericana (LAT) y la Biblia de

⁹ Conjunto de datos específicamente preparado para una tarea concreta de PLN.

las Américas (AME). El *dataset* consta de pares de versículos de 1 Corintios y pares de versículos de Génesis. Primero, se empareja cada uno de los versículos de un libro (por ejemplo, 1 Corintios) en una traducción (AME) con todos los versículos de la otra traducción (LAT).

En el conjunto de datos se marcan sólo como paráfrasis aquellos versículos con la misma etiqueta (excepto el código que identifica cada biblia). El resto, no son considerados paráfrasis. Así, en el ejemplo (2), se obtiene el par LAT1CO001001-AME1CO001001, que se etiqueta automáticamente con el valor positivo para paráfrasis (es decir, con un 1) debido a la coincidencia en el código del versículo:

- (2) a. De Pablo, apóstol Cristo Jesús por decisión Dios que lo ha llamado, y Sóstenes nuestro hermano (LAT1CO001001).
- b. Pablo, llamado a ser apóstol de Jesucristo por la voluntad de Dios, y Sóstenes, nuestro hermano (AME1CO001001).

También se forman pares con etiquetas no coincidentes, que se marcan como no-paráfrasis (con valor 0), como sucede en (3) con (a) y (b) (véase el último dígito de las etiquetas LAT1CO001001 y AME1CO001002).

- (3) a. De Pablo, apóstol Cristo Jesús por decisión Dios que lo ha llamado, y Sóstenes nuestro hermano (LAT1CO001001).
- b. A la iglesia de Dios que está en Corinto, a los que han sido santificados en Cristo Jesús, llamados a ser santos, con todos los que en cualquier parte invocan el nombre de nuestro Señor Jesucristo, Señor de ellos y nuestro (AME1CO001002).

Al comparar el contenido de (3a) y (3b), resulta evidente que dichos textos no son paráfrasis entre ellos. Para evitar la profusión de pares anotados como no paráfrasis que, en realidad, están muy alejados entre sí, se implementó la siguiente estrategia: Se decidió que el 25 % de los pares del *dataset* serían paráfrasis, y se considerarían todos los posibles pares. Así, a partir de los datos etiquetados con el valor 1, se calculó el número de pares etiquetados con 0 para cada libro. Para elegir cuáles de los pares 0 entrarían a formar parte del

conjunto de datos se realizó la prueba de la distancia coseno. Los pares con el coeficiente más alto se incluyeron en el *dataset*, cuya configuración se puede observar en la tabla 1.

TABLA 1. NÚMERO DE PARES DE CADA SUBCONJUNTO EN RELACIÓN CON EL CONJUNTO DE DATOS PARA EL EXPERIMENTO

Libro	Entrenamiento (64 %)	Validación (16 %)	Prueba (20 %)
Génesis	3921	981	1266
1 Corintios	1118	280	350

Una vez elaborado el conjunto de datos con el que se realizarían los experimentos, éste se dividió en tres partes: entrenamiento (64 %), validación (16 %) y prueba (20 %). Con los pares de entrenamiento, se alimenta el sistema de aprendizaje automático. Con los pares de validación, se comprueba si el algoritmo ha generalizado. Finalmente, el conjunto de prueba sirve para evaluar la eficiencia del proceso.

El siguiente paso consistió en la vectorización de las entradas textuales y en la extracción de características. Se usaron dos tipos de métricas para evaluar la distancia entre las dos oraciones de entrada, las *conjuntistas* (Dice, Jaccard, Jaro-Winkler, Levenshtein y coseno), y las *vectoriales* (diferencia de longitud, n-gramas, tf-idf y WMD). Los experimentos usaron diferentes combinaciones de estas métricas con el fin de buscar la mejor configuración.

Por lo que respecta a la implementación neuronal, se optó por una red siamesa (Reimers y Gurevych, 2019), con dos entradas y dos ramas iniciales de procesamiento independientes. La red siamesa se adecua muy bien a la tarea que consiste en procesar dos entradas de manera independiente para después compararlas. En este caso, la red consta de una capa de entrada, una capa de vectorización y una capa LSTM (Long Short Term Memory), esto es, un tipo de redes neuronales recurrentes. Una vez la entrada es procesada por la red, se comparan los vectores resultantes mediante la distancia euclidiana.

Para la representación matemática de los textos en la capa de vectorización se usó un modelo BERT (Bidirectional Encoder Representations from Transformers) (Devlin *et al.*, 2018).

Los resultados obtenidos están alrededor del 0.95 de F1-score, tanto si se usan modelos de aprendizaje automático clásicos como si se emplean modelos

basados en *transformers*. Como la medida del F1-score está entre 0 y 1, los resultados son altamente satisfactorios.

APLICACIÓN PARA AGRUPACIÓN SEMÁNTICA

El objetivo principal del agrupamiento semántico es identificar pares de palabras que puedan ser utilizadas de manera intercambiable en un cierto contexto, de manera que las palabras de un grupo están fuertemente asociadas, ya que comparten características semánticas, mientras que palabras de otro grupo mostrarán características diferentes.

Contar con grupos de palabras semánticamente similares resulta de gran utilidad para las tecnologías del lenguaje debido a su relevancia en varias aplicaciones, como la detección de plagio, la respuesta a preguntas, el entendimiento textual, el resumen y la evaluación de los sistemas de traducción automática. Un método para encontrar grupos semánticos a partir del alineamiento de definiciones fue desarrollado por Sierra y McNaught (2000), específicamente diseñado para identificar pares de palabras donde uno de sus miembros puede ser reemplazado por el otro en el contexto de una definición sin alterar el concepto subyacente.

Este método, mediante la alineación de definiciones que expresan el mismo concepto, pero a través de diferentes formulaciones, logra determinar las correspondencias más probables entre las palabras en la oración de destino y las palabras en la oración de origen. De manera análoga, gracias a la disponibilidad de nuestro corpus paralelo de biblias en español, es posible alinear los versículos de las diferentes versiones para identificar equivalencias de palabras dentro de los versículos dados. Véase, por ejemplo, en (4), el versículo de Marcos 5: 34 para las biblias ESP (a), LAT (b), JUN (c) y NAC (d).

- (4) a. Él le dijo: Hija, tu fe te ha curado. Vete en paz y sigue sana de tu tormento (ESPMAR005034).
- b. Jesús le dijo: “Hija, tu fe te ha salvado; vete en paz y queda sana de tu enfermedad” (LATMAR005034).
- c. Y él díjola: “Hija, tu fe te ha salvado, ve a paz, sé sana de tu plaga” (JUNMAR005034).

- d. Y Él le dijo: Hija, tu fe te ha salvado; vete en paz y seas curada de tu mal (NACMAR005034).

Al observar las oraciones, se pueden identificar varios pares y agrupamientos de palabras, a saber: *él, Jesús; curado, salvado; sigue, queda; sana, curada, y tormento, enfermedad, plaga, mal*. Estos grupos exhiben correspondencia semántica, lo que indica que las palabras pueden sustituirse entre sí sin alterar significativamente el significado.

Una ventaja del algoritmo de alineación es que no depende de grandes cantidades de datos y puede generar agrupamientos incluso cuando la alineación de palabras es infrecuente, ya que compara las palabras en dos versículos de manera secuencial, estableciendo correspondencias entre ellas. Básicamente, identifica el costo mínimo asociado con cada operación requerida para transformar una frase en otra, incluidas las sustituciones de palabras, inserciones y eliminaciones, lo que se conoce como distancia de Levenshtein (Levenshtein, 1966). Conforme aumente el número de palabras iguales circundantes a un par de palabras diferentes, mayor será la similitud semántica del par en cuestión, lo que coincide con la hipótesis distribucional de Harris (1954), que postula que las palabras con un uso contextual similar tienden a compartir significados similares.

Para el ejemplo de Marcos 5: 34, en el alineamiento (5) formado por las biblias ESP y LAT, podemos ver a continuación que los pares de palabras diferentes de (a) y (b), que denominamos *pares correspondientes*, tienen varias palabras alrededor que son iguales en (a) y (b).

- (5) a. Él le dijo Hija tu fe te ha curado Vete¹⁰ en paz y sigue sana de tu tormento (ESPMAR005034).

- b. Jesús le dijo Hija tu fe te ha salvado vete en paz y queda sana de tu enfermedad (LATMAR005034).

¹⁰ En estos ejemplos, se mantiene el uso de mayúsculas acorde con la fuente de origen, pero se han omitido los signos de puntuación.

Véase que el par correspondiente *curado-salvado* tiene 7 palabras iguales a la izquierda y 4 a la derecha, por lo que puede considerarse que su ventana de similitud es de 11 palabras. Por su parte, los pares *Él-Jesús* y *sigue-queda* tienen una ventana de 7 palabras, mientras que *tormento-enfermedad* tiene una ventana de similitud de 3.

A continuación, se muestra una selección de grupos semánticos derivados de haber realizado la tarea de agrupar o unir los pares coincidentes identificados en el libro de Marcos a lo largo de las biblias ESP, JNM, NAC y LAT. Estos agrupamientos demuestran las asociaciones entre palabras que exhiben equivalencia semántica dentro de los textos analizados. Y son los siguientes:

- Mente, inteligencia, entendimiento
- Departamento, pieza
- Diestro, derecho
- Decir, preguntar, responder, contestar, asegurar, contar
- Mañana, madrugada
- Impuro, malo, inmundo
- Enojar, indignar
- Ir, venir, volver, regresar, llegar, bajar, pasar, acercar
- Permitir, dejar
- Luego, después
- Lago, orilla, mar
- Germinar, brotar
- Misericordia, compasión
- Tierra, suelo
- Hablar, conversar
- Príncipe, jefe, oficial
- Predicar, anunciar
- Hombre, gente, nación, gentío, pueblo, aldea, multitud

Se observa que ciertos grupos semánticos coinciden con vocablos que pudieran encontrarse relacionados en cualquier diccionario de sinónimos. Por ejemplo, sustantivos como *mente*, *inteligencia* y *entendimiento*, así como verbos como *enojar* e *indignar*. En otros casos, los grupos semánticos consisten en *sinónimos indirectos*, como en el caso de *departamento* y *pieza*, pues si bien no son sinónimos, ambos comparten *habitación* como sinónimo.

Además, se presentan casos en los que los agrupamientos semánticos comprenden cohipónimos, como *príncipe*, *jefe* y *oficial* para sustantivos, o implicación léxica, como *decir*, *preguntar*, *responder*, *contestar*, *asegurar* y *contar* para verbos. Estas agrupaciones dan cuenta de relaciones de inclusión o implicación, donde una palabra abarca o implica el significado de otra dentro del mismo grupo.

Sonia Morett (2006) realizó un estudio sintáctico de un corpus de pares de definiciones en el área de física para extender el método de alineamiento con el fin de ir más allá de la alineación consecutiva de palabra por palabra y mejorar el rendimiento de obtención de grupos semánticos. A partir de su estudio de las relaciones y funciones de los constituyentes de la oración, propuso realizar un preprocesamiento del corpus para considerar, entre otros, perífrasis gramaticales (locuciones, términos compuestos y perífrasis verbales), las palabras que funcionan como nexos (conjunciones, preposiciones y verbos copulativos), las relaciones léxicas (sinonimia, hiponimia e hiperonimia) y la negación. Un ejemplo de locución y relación semántica se tiene en el alineamiento (6), de Marcos 1: 13, entre ESP (a) y LAT (b):

- (6) a. Estuvo en el desierto cuarenta días: Satanás lo ponía a prueba, estaba con las fieras y los ángeles le servían (ESPMAR001013).
- b. Estuvo cuarenta días en el desierto y fue tentado por Satanás. Vivía entre los animales salvajes y los ángeles le servían (LATMAR001013).

Por su parte, Natalia Guerrero (2022) observó otra serie de fenómenos lingüísticos que ocurren entre pares de versículos y que se detectan con el algoritmo de agrupación semántica, tales como la derivación y la composición, la sustitución léxica (hiperónimos e hipónimos, merónimos y holónimos, antónimos). Algunos de estos fenómenos son particularmente frecuentes en algunas versiones de la Biblia. Un ejemplo es la enclisis en el siguiente pasaje (7) de Marcos 1: 17, muy común en JNM.

- (7) a. Jesús les dijo: “Sígueme y yo los haré pescadores de hombres” (LATMAR001017).

- b. Y díjoles Jesús: “Venid en pos de mí, y haréos ser pescadores de hombres” (JNMMAR001017).

APLICACIÓN PARA EVALUACIÓN DE LLM PARA PREGUNTA-RESPUESTA

Los modelos grandes de lenguaje (LLMs, por sus siglas en inglés) han transformado el acceso y procesamiento de la información, especialmente en sistemas de pregunta/respuesta. Estos modelos entienden y responden consultas en lenguaje natural, lo que facilita el diseño de interfaces intuitivas y accesibles para los usuarios. Su capacidad para analizar el contexto y producir respuestas relevantes y coherentes es crucial en todas las áreas de conocimiento.

Los LLM, entrenados con grandes conjuntos de datos, poseen la capacidad de responder consultas específicas extrayendo y sintetizando conocimientos de sus bases de datos internas. Sin embargo, surge un desafío significativo cuando la información requerida es restringida, es decir, contiene detalles específicos que podrían no estar en su base de datos de entrenamiento, o bien, diferir, en algunos aspectos, de la almacenada. Por ejemplo, en el caso de manuales técnicos, las regulaciones de construcción que presentan diferentes estipulaciones que varían de un país a otro; o cuando se trata de políticas y procedimientos de empresas, los cuales pueden diferir sustancialmente de las regulaciones generales.

En este sentido, el desafío radica en la capacidad de estos modelos para interpretar y aplicar su conocimiento a casos que, aunque superficialmente se parezcan a situaciones familiares, difieren en aspectos críticos que pueden afectar la precisión de las respuestas generadas. La capacidad de ajustarse a estas diferencias sutiles pero fundamentales es lo que pone a prueba la versatilidad y efectividad de los LLMs en aplicaciones prácticas.

Considerando lo anterior, en esta sección se reportan distintos experimentos realizados con el fin de evaluar la adecuación de los LLMs para inferir respuestas basadas exclusivamente en documentos restringidos, limitando el uso de información previamente almacenada en sus bases de conocimiento. En estas pruebas se ha intentado encapsular el funcionamiento de los modelos de lenguaje, aislándolos de su conocimiento anterior. Los LLMs usados en

nuestros experimentos fueron Llama2 (Chowdhery *et al.*, 2022), ChatGPT (Brown *et al.*, 2020) y PaLM (Touvron *et al.*, 2023).

Como estudio pivote, se seleccionó el dominio bíblico como tema central, y se determinó que las respuestas a las preguntas debían basarse no en el conocimiento general, sino estrictamente en capítulos específicos de la Biblia; en concreto, de la versión LAT de nuestro corpus. El enfoque del experimento consiste en probar la eficiencia de los modelos de lenguaje utilizados para esta tarea, no en la interpretación religiosa de los textos. Estas preguntas requieren que los LLMs consideren la Biblia como una fuente de información, no como directrices religiosas o como información que contradice, o no, el conocimiento preentrenado del modelo. Los experimentos tienen como objetivo evaluar si los LLMs pueden:

- a) limitarse a la información y a los *prompts*¹¹ proporcionados;
- b) ignorar el contexto y corregir la respuesta usando su conocimiento preentrenado, y
- c) verificar si la información proporcionada afecta la exactitud de la respuesta.

Para tal efecto, en primera instancia, se escribieron una serie de preguntas sobre diversos pasajes de la Biblia, tanto del Antiguo como del Nuevo Testamento. En total, se seleccionaron 36 preguntas para la versión final del experimento. La mayoría de ellas referidas a eventos, nombres, momentos y lugares específicos. Dividimos estas cuestiones en dos grandes grupos: A) aquellas preguntas que pueden ser contestadas correctamente con la información existente en la Biblia, y B) aquellas para las que el texto fuente no proporciona información suficiente que permita llegar a una solución válida.

El primer tipo de preguntas, que corresponde a la mayor parte de nuestro cuestionario, engloba 31 *inputs* (tabla 2). En general, la respuesta se puede localizar en la información contenida en un capítulo, como sucede con las consultas 1 y 3, que corresponden a Éxodo 14 y 1 Samuel 17. En muy pocos casos, la respuesta se llega a encontrar en un solo versículo. Un ejemplo de esto último corresponde a la pregunta 10, cuya solución se localiza en Daniel 1:

¹¹ Los *prompts* son informaciones o consultas simples proporcionadas a un LLMs para obtener una respuesta.

1. Con el fin de probar el razonamiento de información factual de un LLM, algunas preguntas requieren de interpretación, como la pregunta 16.

TABLA 2. PREGUNTAS DEL GRUPO A, QUE PUEDEN SER RESPONDIDAS DIRECTAMENTE CON INFORMACIÓN DE LA BIBLIA

#	Pregunta	Respuesta
1	¿Qué mar fue abierto por Moisés y para qué?	El Mar Rojo
2	¿Qué ídolo erróneamente veneran los israelitas y por qué?	Un becerro de oro
3	¿Cómo mató David a Goliat?	Con una piedra de su honda
4	¿Quién tuvo el sueño de las vacas gordas y las vacas flacas?	El Faraón de Egipto
5	¿A quién le dijo Rut las palabras: donde tú vayas, iré yo, y donde tú vivas, viviré yo; tu pueblo será mi pueblo y tu Dios será mi Dios?	A su suegra Noemí
6	¿Cómo se llamaba el jefe del ejército a quien derrotaron los israelitas bajo el mando de la jueza Débora?	Sísara
7	¿Quién mató a Holofernes?	Judit
8	¿Quién despojó de su cabellera a Sansón y por qué lo hizo?	Su esposa, Dalila, para que perdiera su fuerza
9	¿Qué comían los israelitas en el desierto?	Maná
10	¿Quién era Nabucodonosor?	El rey de Babilonia
11	¿Qué oficio tenía Melquisedec?	Sacerdote
12	¿Qué construyó Noé?	Un arca
13	¿De dónde era Ciro?	Ciro era de Persia
14	¿Quién tentó a Jesús en el desierto?	Satanás
15	¿En qué monte fue crucificado Jesús?	En Gólgota, también llamado Calvario
16	¿Qué pareja acompañó a Pablo en algunos de sus viajes?	Aquila y Priscila
17	¿Quién se quedó sin oreja la noche que murió Jesús?	Malco

TABLA 2. (CONT.)

#	Pregunta	Respuesta
18	¿Quién estaba siendo juzgado junto con Jesús por los romanos?	Barrabás
19	¿Qué le hizo Juan Bautista a Jesús?	Lo bautizó
20	¿Por cuántas monedas Judas traiciona a Jesús?	Treinta piezas de plata
21	¿Quién hizo que decapitaran a Juan Bautista?	La hija de Herodías
22	¿Qué amigo le escribe dos cartas a Timoteo?	Pablo de Tarso
23	¿Quién niega a Jesús? ¿Cuántas veces?	Pedro. Tres
24	¿A qué hora murió Jesús?	A las tres de la tarde
25	¿En qué ciudad hizo Pablo de Tarso su discurso "Al Dios no conocido"?	Atenas
26	¿En qué fueron grabados los diez mandamientos y cuáles son esos?	Fueron dados en dos tablas de piedra y son: 1. Amarás a Dios sobre todas las cosas. Sólo hay un Dios, creador y todopoderoso, al cual adorar. 2. No tomarás el nombre de Dios en vano. 3. Santificarás las fiestas. 4. Honrarás a tu padre y a tu madre. 5. No matarás. 6. No cometerás actos impuros. 7. No robarás. 8. No darás falso testimonio ni mentirás. 9. No consentirás pensamientos ni deseos impuros. 10. No codiciarás los bienes ajenos.
27	¿Cuántos hijos tuvo Jacob? ¿Cómo se llamaban?	Doce. Rubén, Simeón, Leví, Judá, Isacar, Zabulón, Dan, José, Benjamín, Neftalí, Gad, Aser
28	¿Cuál era el oficio de Mateo antes de unirse a los seguidores de Jesús? ¿Y de Pedro?	Cobrador de impuestos (publicano). Pescador
29	¿Cuántos candeleros hay en Apocalipsis y a qué se refiere?	Siete. A las siete iglesias
30	¿A quién se tragó el gran pez?	A Jonás
31	¿A quién le fue revelado el libro del Apocalipsis?	A Juan

El grupo B incluye solamente 5 preguntas (tabla 3). Para contestarlas se requiere más información que la proporcionada por el contexto, pues –en este caso– se precisa de interpretaciones y opiniones formuladas por académicos o autoridades eclesiásticas. Por tanto, se espera que los LLMs no puedan proporcionar respuestas, ya que éstas no se encuentran explícitamente en los libros de la Biblia.

En algunas ocasiones, el modelo ha sido capaz de contestar estas cuestiones, y esto prueba cómo hace uso de su conocimiento previo; se trata, por tanto, de una malfunción del sistema. En algunas ocasiones, en cambio, se podría defender que un sistema inferencial suficientemente potente sería capaz de deducir la respuesta. Por ejemplo, para la pregunta 33, el sistema podría llegar a la conclusión de que David es el autor de los Salmos, debido a las constantes referencias a este rey que se ponen en boca de Dios en los textos cuando el autor y el creador están en diálogo, como en los versículos de (8).

- (8) a. Por mi propia santidad lo juré una vez, ¿cómo podría yo mentir a David? (SAL089035)
- b. Señor, ¿dónde están tus favores pasados que juraste a David por tu fidelidad? (SAL089049)
- c. Señor, tenle en cuenta a David su fervoroso celo (SAL132001)

TABLA 3. PREGUNTAS DEL GRUPO B, QUE REQUIEREN MÁS INFORMACIÓN QUE LA PROPORCIONADA POR EL TEXTO PARA SER RESUELTAS

#	Pregunta	Respuesta
32	¿Qué libro de la Biblia cuenta el amor de los esposos?	El Cantar de los Cantares
33	¿Quién es considerado el autor de los Salmos?	El rey David
34	¿Qué son Isaías, Jeremías, Ezequiel y Daniel?	Profetas: los profetas mayores
35	¿Qué profeta escribió el libro de Lamentaciones?	Jeremías
36	¿Cuál era el más escéptico de los discípulos de Jesús?	Tomás

Como se mencionó anteriormente, el cuestionario está diseñado para probar las capacidades de cada modelo más allá de extraer respuestas explícitas del texto. El proceso para llegar a la extracción tiene dos pasos: el primero consiste en la *recuperación de contextos*, y el segundo, en la *extracción de la respuesta correcta de cada modelo*. A continuación, se describen ambos procesos.

Para llevar a cabo la recuperación de contextos, se seleccionó un enfoque de *generación aumentada por recuperación* (RAG) (Lewis *et al.*, 2021). La metodología RAG es ampliamente usada como método para mejorar las respuestas de los LLM a una pregunta dada. Su funcionamiento consiste en expandir el conocimiento de un modelo de lenguaje mediante el análisis de una pregunta determinada, alimentando así el LLM con información relacionada, la cual se extrae de los datos sobre los que actúa.

Para garantizar su consistencia y eliminar el ruido que puede derivarse del proceso de recuperación de cada modelo, la parte de recuperación se realizó mediante *embeddings*¹² estandarizados. Las preguntas y los capítulos se representaron utilizando un modelo de *embeddings* consistente para todos los LLMs, seguido de una recuperación por similitud. Los textos recuperados fueron utilizados posteriormente por cada LLM para responder las preguntas correspondientes. A continuación, se implementó el modelo para la extracción de las respuestas. Para medir la incidencia del RAG en el resultado, se hicieron dos experimentos diferentes. Primero, se formularon las preguntas sin proporcionar contexto adicional, confiando únicamente en los datos de entrenamiento de los LLMs. Esto sirvió como una línea base para comparar con el sistema RAG completo.

La segunda evaluación incluyó los pares de texto y pregunta recuperados, utilizando *prompts* diseñados para instruir al modelo a responder basándose únicamente en el contexto proporcionado. Cuando éste resultaba insuficiente, se instruyó al modelo a indicarlo con un comentario como “información no disponible”. La tabla 4 muestra los *prompts* utilizados para cada modelo LLM con RAG.

Cada par conformado por pregunta y contexto se ingresó en el modelo de manera iterativa, generando una sola respuesta por iteración. Las respuestas, junto con las preguntas y los contextos correspondientes, se recopilaron para

¹² Vectores numéricos que representan palabras y que capturan información léxica.

la evaluación. Los resultados se representan como tuplas de tres elementos: **pregunta**, **contexto** y **respuesta**.

Teniendo en cuenta los dos tipos de evaluaciones, con contexto y sin contexto, y los tres LLM utilizados (LLAMA, ChatGPT y PaLM), es posible analizar los resultados que se generan para cada uno de los grupos de preguntas: grupo A y grupo B. Los resultados obtenidos se muestran en la tabla 5, donde FI (**falsa inferencia**) significa que el LLM ha recurrido a su conocimiento externo, y 0 significa que no hubo respuesta.

TABLA 4. PROMPTS UTILIZADOS PARA CADA MODELO

PaLM	LLaMA	GPT
Responde la pregunta dado el siguiente contexto {{context:}}. \n Si la respuesta \n no está dentro del {{context:}} y no estás seguro, por favor responde "Información no disponible".	Dado el siguiente contexto: '{retrieved_info}', responde: '{question}'. \n \n Si la respuesta no está dentro de la información proporcionada y no estás seguro de la respuesta, por favor responde "Información no disponible en el contexto".	Si la respuesta no está dentro del {{context:}} y no estás seguro, por favor responde "Información no disponible".

TABLA 5. RESULTADOS PARA LAS 31 PREGUNTAS DEL GRUPO A

	LLAMA CONTEXTO	LLAMA NO CONTEXTO	CHATGPT CONTEXTO	CHATGPT NO CONTEXTO	PALM CONTEXTO	PALM NO CONTEXTO
Correctas	19	15	17	17	19	23
Incorrectas	8	16	3	6	6	2
FI	4	-	1	-	1	-
0	-	-	10	8	5	6

En el caso de las consultas con contexto, para que una respuesta se considere correcta, no sólo es necesario que sea acertada, sino que haya sido adecuadamente inferida a partir del RAG. En muchas ocasiones el RAG no arroja un contexto de donde se pueda extraer la solución. Si se falla en este primer paso, y el algoritmo funciona de forma adecuada, la solución coherente es la de que no existen suficientes datos (0). En algunas ocasiones, en cambio, la respuesta es la que se esperaba, aunque no se haya inferido del RAG. Esto

quiere decir que el LLM ha acudido a su conocimiento más allá del texto fuente. Para los propósitos de este experimento, se considera una solución incorrecta, que se marca como FI.

La tabla 6 muestra el ejemplo de la pregunta 18: “¿Quién estaba siendo juzgado junto con Jesús por los romanos?” En este caso, el RAG ofrece todo el capítulo 15 de Marcos, que incluye la información que se buscaba. Por tanto, el contexto es óptimo para que los tres modelos puedan encontrar una respuesta satisfactoria. En cambio, ni LLAMA ni ChatGPT lo hacen, ya que confunden el pasaje del juicio por parte de Pilatos con el de la crucifixión. En cambio, PaLM RAG encuentra la información solicitada sin problemas.

TABLA 6. RAG Y CONTESTACIONES DE LOS MODELOS PARA LA PREGUNTA 18: “¿QUIÉN ESTABA SIENDO JUZGADO JUNTO CON JESÚS POR LOS ROMANOS?”

RAG	LLAMA RAG	ChatGPT RAG	PaLM RAG
MARCOS 15	En cuanto a tu pregunta, según el texto que me has proporcionado, Jesús fue juzgado junto con otros dos ladrones, uno a su derecha y otro a su izquierda, por los romanos. El texto especifica que los tres fueron crucificados juntos en un lugar llamado Gólgota o Calvario	Información no disponible	Barrabás
Correcto	Incorrecto	Incorrecto	Correcto

La tabla 7 muestra los resultados con RAG para la pregunta 5: “¿A quién le dijo Rut las palabras: donde tú vayas, iré yo; y donde tú vivas, viviré yo; tu pueblo será mi pueblo y tu Dios será mi Dios?” En esta ocasión, el RAG obtenido corresponde a un texto completamente alejado de la pregunta. En cambio, LLAMA es capaz de encontrar la respuesta correcta: *Noemí*, a pesar de que la identifica como madre y no como suegra; aunque esto no puede desprenderse de la información del RAG. Por tanto, la etiqueta asignada es FI (**falsa inferencia**). Aquí se observa que el modelo no ha funcionado porque claramente ha salido de su dominio. En cuanto a ChatGPT y PaLM, dicen no tener información disponible, por lo que se les adjudica la etiqueta 0; en cambio, esto es coherente con lo que reporta el RAG.

TABLA 7. RAG Y CONTESTACIONES DE LOS MODELOS PARA LA PREGUNTA 5: "¿A QUIÉN LE DIJO RUT LAS PALABRAS: DONDE TÚ VAYAS, IRÉ YO; Y DONDE TÚ VIVAS, VIVIRÉ YO; TU PUEBLO SERÁ MI PUEBLO Y TU DIOS SERÁ MI DIOS?"

RAG	LLAMA RAG	ChatGPT RAG	PaLM RAG
Lev 18	La persona a quien Rut le dijo las palabras "donde tú vayas, iré yo; y donde tú vivas, viviré yo; tu pueblo será mi pueblo y tu Dios será mi Dios" fue Noemí, su madre	Información no disponible	No hay información disponible en el contexto
Incorrecto	F1	0	0

Por lo que respecta a las preguntas del grupo B, no se espera que el sistema sea capaz de inferir las contestaciones, ya que no se desprenden directamente del texto. La tabla 8 muestra los resultados que indican que hubo claramente un mejor rendimiento en aquellos experimentos donde la respuesta no dependía del RAG sobre aquellos para los que sí se requería, donde tanto LLAMA y PaLM, con dos respuestas correctas, como ChatGPT, que no consiguió responder ninguna, tienen un desempeño muy pobre.

TABLA 8. RESULTADOS PARA LAS 5 PREGUNTAS DEL GRUPO B

	LLAMA contexto	LLAMA no contexto	ChatGPT contexto	ChatGPT no contexto	PaLM contexto	PaLM no contexto
Correctas	2	4	-	4	2	5
Incorrectas	2	1	1	-	3	-
F1	1	-	-	-	-	-
0	-	-	4	1	-	-

Una observación clave que se desprende de este estudio es la demostración de la capacidad de estos modelos para responder preguntas basadas en un contexto dado. La variabilidad en las respuestas correctas entre los modelos *con* y *sin* contexto depende de factores como la correcta recuperación

de información, el tamaño del modelo y la disponibilidad de datos. Cabe destacar que no podemos garantizar la ausencia de la Biblia en los datos de preentrenamiento de ninguno de los tres modelos.

Esta metodología facilita la adaptación al dominio de manera sencilla y, con constantes actualizaciones de los modelos, se puede utilizar sin necesidad de un ajuste fino.

Es importante mencionar la relevancia de contar con modelos de código abierto, ya que –de otra forma– el costo computacional y económico de generar modelos de este tipo es accesible sólo para unas pocas de las empresas más grandes. Con la metodología RAG, podemos aprovechar estos sistemas para evitar la necesidad de entrenar un modelo de este tamaño.

REFLEXIONES FINALES

La elaboración de corpus, conjuntos de datos y recursos de diferentes tipo, es un elemento fundamental y previo para la lingüística de corpus, el PLN y, en definitiva, para las tecnologías del lenguaje que, finalmente, nutren los sistemas de inteligencia artificial, que en los últimos años han venido para quedarse en la sociedad, las relaciones, las formas de comunicación y la vida cotidiana.

La mayor parte de los corpus de investigación se encuentran en inglés, y cada vez resulta más difícil la implementación de tecnologías del lenguaje en otras lenguas. El español, a pesar de ser una lengua extendida, apenas puede competir con los avances que se encuentran basados en las estructuras, el léxico y la cultura inglesa.

El corpus alineado de traducciones de la Biblia en español, elaborado por el Grupo de Ingeniería Lingüística (GIL) del Instituto de Ingeniería de la UNAM, es un recurso pensado para posibilitar el avance de la lingüística computacional en español. El corpus presenta un mismo texto en once variantes cercanas semánticamente equivalentes. La elaboración de este corpus ha sido un trabajo colaborativo de los investigadores del GIL con estudiantes, alumnos de servicio social y becarios de investigación.

A partir de este corpus, se pueden llevar a cabo estudios realizados por becarios de investigación, con fondos del proyecto CF-2023-G de CONAHCyT y en diferentes áreas de la lingüística y del PLN, como similitud,

extracción léxica, detección de paráfrasis, estudio de sesgo, cambio lingüístico y dialectología. En este trabajo, se han presentado, en líneas generales, tres aplicaciones: agrupación semántica, evaluación de sistemas de pregunta-respuesta y detección de paráfrasis. El corpus se ofrece pública y gratuitamente para que se puedan evaluar en él diferentes algoritmos para todo tipo de tareas de PLN y, de esta manera, impulsar el desarrollo de tecnologías del lenguaje en español.

BIBLIOGRAFÍA

- Brown, Tom B. *et al.* (2020), “Language Models are Few-Shot Learners”, pp. 1-75, consultado el 24 de agosto de 2023. 
- Chowdhery, Aakanksha *et al.* (2022), “PaLM: Scaling Language Modeling with Pathways”, pp. 1-87, consultado el 24 de agosto de 2024. 
- Christodouloupoulos, Christos y Mark Steedman (2015), “A massively parallel corpus: the bible in 100 languages”, *Language Resources and Evaluation*, vol. 49, núm. 2, pp. 375–395, consultado el 24 de agosto de 2024. 
- Das, Dipanjan y Noah A. Smith (2009), “Paraphrase identification as probabilistic quasi-synchronous recognition”, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 468–476, consultado el 24 de agosto de 2024. 
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee y Kristina Toutanova (2019), “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186, consultado el 24 de agosto de 2024. 
- Dolan, William B. y Chris Brockett (2005), “Automatically Constructing a Corpus of Sentential Paraphrases”, *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, pp. 9-16, consultado el 24 de agosto de 2024. 
- Eckhoff, Hanne M. *et al.* (2018), “The PROIEL treebank family: a standard for early attestations of Indo-European languages”, *Language Resources & Evaluation*, vol. 52, pp. 29-65.
- Guerrero Cerón, Natalia (2023), Análisis de paráfrasis en traducciones de la Biblia en español, tesis de licenciatura en Lengua y Literaturas Hispánicas, México,

- Facultad de Estudios Superiores Acatlán, Universidad Nacional Autónoma de México, consultado el 24 de agosto de 2024. 
- Harris, Zellig S. (1954), “Distributional structure”, *Word*, vol. 10, núm. 2-3, pp. 146-162, consultado el 24 de agosto de 2024. 
- Klein, William W., Craig B. Blomert y Robert L. Hubbard Jr. (2017), *Introduction to Biblical Interpretation* (3ª. ed.), California, Zondervan.
- Kong, Leilei, Zhongyuan Han, Yong Han y Haoliang Qi (2020), “A deep paraphrase identification model interacting semantics with syntax”, *Complexity*, vol. 2020, pp. 1-14, consultado el 24 de agosto de 2024. 
- Levenshtein, Vladimir I. (1966), “Binary codes capable of correcting deletions, insertions, and reversals”, *Soviet physics doklady*, vol. 10, núm. 8, pp. 707–710.
- Lewis, Patrick *et al.* (2021), “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”, consultado el 24 de agosto de 2024. 
- Mayer, Thomas y Michael Cysouw (2014), “Creating a massively parallel Bible corpus”, *International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3158–3163, consultado el 24 de agosto de 2024. 
- McCarthy, Arya D. *et al.* (2020), “The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration”, *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 2884-2892, consultado el 24 de agosto de 2024. 
- Morett Álvarez, Sonia Elisa (2006), Inclusión de una perspectiva sintáctica a la obtención de pares semánticos, tesis de licenciatura en Lengua y Literaturas Hispánicas, México, Universidad Nacional Autónoma de México, consultada el 24 de agosto de 2024. 
- Mota Montoya, Margarita A., Iria Da Cunha y Fernanda López-Escobedo (2016), “Un corpus de paráfrasis en español: metodología, elaboración y análisis”, *RLA. Revista de lingüística teórica y aplicada*, vol. 54, núm. 2, pp. 85-112, consultado el 24 de agosto de 2024. 
- Reimers, Nils e Iryna Gurevych (2019), “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, consultado el 24 de agosto de 2024. 
- Resnik, Philip, Mari Broman Olsen y Mona Diab (1999), “The Bible as a parallel corpus: Annotating the ‘book of 2000 tongues’”, *Language Resources and Evaluation*, vol. 33, pp.129-153.

- Sierra, Gerardo y John McNaught (2000), “Extracting semantic clusters from MRDs for an onomasiological search dictionary”, *International Journal of Lexicography*, vol. 13, núm. 4, pp. 264-286, consultado el 24 de agosto de 2024. 
- Sierra, Gerardo *et al.* (2024), “An aligned corpus of Spanish bibles”, *Language Resources & Evaluation*, consultado el 24 de agosto de 2024. 
- Touvron, Hugo *et al.* (2023), “LLaMA: Open and Efficient Foundation Language Models”, pp. 1-27, consultado el 24 de agosto de 2024. 
- Zhang, Yuan, Jason Baldridge y Luheng He (2019), “PAWS: Paraphrase Adversaries from Word Scrambling”, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 1298–1308, consultado el 24 de agosto de 2024. 

Corpus y repositorios en línea

- Enrique-Arias, Andrés y Francisco Javier Pueyo Mena (2008), “Biblia medieval”, *Corpus Biblia Medieval*, consultado el 24 de agosto de 2024. 
- Grupo de Ingeniería Lingüística – Universidad Nacional Autónoma de México (s.f.), *Spanish Paraphrase Corpora*, consultado el 23 de octubre de 2024. 
- Iyer, Shankar, Nikhil Dandekar y Kornél Csernai (s. f.), *First Quora Dataset Release: Question Pairs*, consultado el 23 de octubre de 2024. 
- Sierra G., Bel-Enguix G., Díaz-Velasco A., Guerrero-Cerón N., Bel N. (2024). Corpus paralelo de Biblias, consultado el 23 de octubre de 2024. 
- Wikipedia the Free Encyclopedia (s.f.), *Bible translations*, consultado el 10 de agosto de 2024. 
- Wikipedia the Free Encyclopedia (s.f.), *Wikimedia Downloads*, consultado el 30 julio 2024. 

Biblias que conforman el corpus

- Casidoro de Reina, Jerónimo (trad.) (1569), *La Biblia, que es, los Sacros Libros del Viejo y Nuevo Testamento*, [s.e.], Basilea.
- de Ausejo, Serafín (trad.) (1975), *La Biblia*, Barcelona, Editorial Herder.
- de Valera, Casidoro (trad.) (1960), *La Santa Biblia*, Amsterdam, Editorial Sociedades Bíblicas Unidas. (Original publicado en 1602).
- Escuela Bíblica de Jerusalén (trads.) (1967), *La Biblia de Jerusalén*, París, Desclée de Brouwer. (Original publicado en 1956).

- Jünemann, Wilhem (1928), *La Sagrada Biblia*. Concepción (Chile), Editorial Diocesana de Concepción.
- Lockman Foundation (trad.) (1986), *La Biblia de las Américas*, La Habra (California), Lockman Foundation.
- Martín Nieto, Evaristo (trad) (1961), *La Santa Biblia*, Madrid: San Pablo.
- Nácar, Eloíno y Alberto Colunga (trads.) (1944). *Sagrada Biblia*. Madrid, Biblioteca de Autores Cristianos.
- Ricchiardi, Ramón y Hurault, Bernardo (trads.) (1972), *La Nueva Biblia. Edición pastoral para Latinoamérica*. San Pablo/Verbo Divino.
- Schökel, Luis Alonso, y Juan Mateos (trads.) (1975), *Nueva Biblia Española*, Madrid, Ediciones Cristiandad.
- Torres Amat, Félix (trad.) (1825), *La Sagrada Biblia*, Montaner y Simón Editores, Barcelona (Original publicado en 1823).

GERARDO E. SIERRA-MARTÍNEZ: Investigador titular del Instituto de Ingeniería en la UNAM, en donde es fundador y líder del Grupo de Ingeniería Lingüística. Realizó su doctorado en lingüística computacional, en UMIST, Inglaterra. Ha sido responsable de continuos proyectos patrocinados de tecnologías del lenguaje. Sus áreas de interés se enfocan a la lexicografía computacional, terminótica, extracción conceptual, corpus lingüísticos, minería de textos y lingüística forense. Es autor del libro *Introducción a los corpus lingüísticos*; coautor de los libros *Tratamiento de información textual y generación de taxonomías* y *Lingüística computacional en México: Investigación y desarrollo*, y ha publicado más de doscientos artículos en revistas, capítulos de libros y artículos arbitrados de congresos.

GEMMA BEL-ENGUIX: Doctora en Lingüística por la Universitat Rovira i Virgili (Tarragona, Catalunya). Ha trabajado como investigadora en las universidades de Georgetown y Milano-Bicocca y Aix-Marseille. Ha sido investigadora Ramón y Cajal en la URV (Rovira i Virgili). Es experta en modelos bioinspirados para el lenguaje natural y en aproximaciones formales a la lingüística. Desde 2016 trabaja en la UNAM, en el Grupo de Ingeniería Lingüística,

desarrollando distintos temas de investigación dentro del procesamiento de lenguaje natural: estudio de redes sociales, detección de paráfrasis, redes léxicas, análisis de sentimientos, entre otros. Es coautora de ocho libros y tiene numerosas publicaciones en el campo de la lingüística matemática y computacional.

D. R. © Gerardo E. Sierra Martínez, Ciudad de México, enero-junio, 2024.

D. R. © Gemma Bel-Enguix, Ciudad de México, enero-junio, 2024.