UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE FILOLOGÍA
Departamento de Lingüística, Estudios Árabes, Hebreos, Vascos y de Asia Oriental

MÁSTER UNIVERSITARIO EN LETRAS DIGITALES:
ESTUDIOS AVANZADOS EN TEXTUALIDADES ELECTRÓNICAS

TRABAJO FIN DE MÁSTER

Propuesta de arquitectura para sistemas híbridos de Recuperación de Información en español

ESPECIALIDAD: Procesamiento Lenguaje Natural
APELLIDOS Y NOMBRE: Ardoiz Galaz, Alfonso
TUTORES: Ana María Fernández-Pampillón Cesteros y Miguel Ortega Martín
CURSO ACADÉMICO: 2021-2022
CONVOCATORIA: Septiembre 2022
CALIFICACIÓN: Sobresaliente (9)

Resumen

Este Trabajo de Fin de Máster presenta una nueva arquitectura para sistemas de Recuperación de Información en español. Esta arquitectura está basada en las técnicas que mejores resultados están obteniendo en inglés gracias a los avances del Procesamiento del Lenguaje Natural en los últimos años. En concreto se utiliza un sistema de recuperación híbrido que fusiona la rapidez de los modelos probabilísticos clásicos con la eficiencia de los modelos basados en redes neuronales modernas. Complementariamente, el trabajo propone un nuevo conjunto de datos, llamado "RISQAC", para evaluar los sistemas de Recuperación de Información en español.

Abstract

This Master Thesis presents a new architecture for Information Retrieval systems in Spanish. This architecture is based on the techniques that are obtaining the best results in English thanks to the advances in Natural Language Processing in the last years. In particular, a hybrid retrieval system that merges the speed of classical probabilistic models with the efficiency of models based on modern neural networks is used. Additionally, the paper proposes a new dataset, called RISQAC", to evaluate Information Retrieval systems in Spanish.

Palabras clave

PLN, Inteligencia Artificial, IA, Tecnologías Educativas, Recuperación de Información, RI

Keywords

NLP, Artificial Inteligence, AI, Educational Technologies, Information Retrieval, IR

Índice general

Sección 1

Introducción

Hoy en día el campo de la Recuperación de Información (RI) se ha convertido en una de las áreas del Procesamiento del Lenguaje Natural (PLN) en la que, debido al impacto que tienen estos sistemas en la vida real (motores de búsqueda como Google, sistemas de pregunta-respuesta, o sistemas de filtrado de contenido entre otras aplicaciones), se están produciendo más avances, centrados casi exclusivamente en el uso de Inteligencia Artificial (Khattab and Zaharia 2020; Ma et al. 2021; Qu et al. 2021).

Por desgracia, uno de los principales problemas es que en este campo encontramos un desequilibrio muy acentuado entre los desarrollos realizados para el inglés y el resto de lenguas (Hedderich et al. 2021). Esta situación se refleja también en el español, lengua para la cual se observa una carencia de nuevos Sistemas de Recuperación de Información (de ahora en adelante SRI) que sí existen para el inglés.

Con la motivación de intentar acortar esta brecha en nuestro idioma, el presente trabajo introduce una arquitectura actualizada para sistemas de Recuperación de Información en español basada en los avances en similitud semántica dentro del Procesamiento del Lenguaje Natural. Esta arquitectura surge de la combinación de dos técnicas de cálculo de similitud semántica en textos: (i) el algoritmo matemático Okapi BM25 (Robertson and Zaragoza 2009) y, (ii) la arquitectura de red neuronal SentenceTransformer propuesta por Reimers and Gurevych 2019. Reproduciendo, de este modo, el funcionamiento de los mejores sistemas de Recuperación de Información en inglés según el ranking del estándar de comparación BEIR (Thakur et al. 2021).

Complementariamente, para comprobar el funcionamiento de esta nueva arquitectura se ha construido un prototipo informático que la implementa en una aplicación de Recuperación de Información para ordenadores personales. Esta aplicación permite localizar documentos en una base de datos documental. Además, se incluye un conjunto de datos que ha sido construido explícitamente para evaluar SRI. Los motivos de esta implementación son dos: comprobar que la arquitectura RI creada para el español funciona correctamente y tomar medidas de su rendimiento (métricas de Cobertura y Mean Reciprocal Rank).

De este modo, este Trabajo de Fin de Máster aporta dos contribuciones a la comunidad académica: en primer lugar, una nueva arquitectura para sistemas de Recuperación de Información en español; y, en segundo lugar, el prototipo de un nuevo conjunto de datos (creado a partir del corpus "SQAC", Gutiérrez-Fandiño et al. 2022) para la evaluación y comparación de SRI en español.

Esta memoria del trabajo se divide en seis secciones. En esta primera sección se ha presentado el problema y la motivación del trabajo. Las siguientes secciones tratan: la hipótesis y los objetivos (sección dos); el estado de la cuestión de los SRI, su funcionamiento y su evaluación (sección tres); la presentación de la nueva arquitectura de SRI para el español y su propuesta de aplicación (sección cuatro); el proceso de evaluación de la arquitectura de SRI (sección cinco) y las conclusiones del trabajo y líneas de investigación futuras (sección seis).

Por último, la memoria cuenta con cuatro anexos que amplían cuestiones específicas que aparecen a lo largo del trabajo.

Sección 2

Hipótesis y Objetivos

Actualmente la mayoría de los avances en el panorama del PLN están centrados casi exclusivamente en inglés[1]. La consecuencia directa de este hecho es que el resto de idiomas quedan relegados a un segundo o tercer plano donde se acaba optando por usar modelos multilingües, una solución económica pero que, para muchas tareas, obtiene peores resultados en comparación con el uso de sistemas preparados para un idioma particular[2].

A pesar de que la mayoría de los esfuerzos en la investigación de SRI en idiomas distintos al inglés están dedicados a la construcción de sistemas multilingües, se plantea que puede optarse por otra solución basada en adaptar a cada idioma los modelos y técnicas que ofrezcan buenos resultados en inglés. Concretamente la hipótesis de este trabajo se formula de la siguiente forma:

Es posible mejorar los resultados de los SRI más usados en español actualmente (modelos multilingües) mediante la creación de un sistema híbrido de RI exclusivamente para el español basado en: (i) el algoritmo BM25 para obtener una representación semántica de los documentos y de la consulta del usuario y, (ii), una red neuronal de arquitectura tipo cross-encoder para el cálculo de la similitud entre la consulta del usuario y los documentos de la base de datos documental.

Esta hipótesis supone el punto de partida de este Trabajo de Fin de Máster cuyo objetivo es desarrollar una arquitectura de SRI para el español que mejore los resultados de los actuales SRI multilingües que incluyen este idioma[3].

[1] En la plataforma HuggingFace, un espacio digital centralizado de modelos de PLN, el número total de modelos para el inglés alcanza los seis millares, mientras que para el español solo hay seiscientos

[2] Vulić et al. 2021 afirma que los codificadores de los modelos multilingües funcionan peor que los codificadores monolingües para tareas relacionadas con la similitud semántica

[3] Incluimos en esta lista modelos basados en el algoritmo BM25, y los modelos multilingües mT5 (Xue et al. 2021) y mMiniLM (Wang et al. 2020)

Adicionalmente, para certificar la consecución de la hipótesis, resulta necesario disponer de un método de evaluación de la eficacia de los SRI en español que sea preciso. Pero, lamentablemente, en la actualidad los métodos disponibles no son de suficiente calidad. Esta carencia no permite verificar formalmente si los resultados obtenidos por el SRI que va a ser desarrollado en este trabajo consiguen superar a los resultados de los actuales SRI multilingües.

De este modo, se proponen los dos objetivos principales del trabajo:

1. El diseño e implementación de una arquitectura híbrida para la RI en español que combine la aplicación del algoritmo BM25 y una red neuronal cross-encoder de forma semejante a las arquitecturas actuales de los mejores sistemas de RI en inglés.

2. La evaluación de esta nueva arquitectura mediante el diseño de un nuevo conjunto de datos para la evaluación de los SRI en español.

Por último, de forma simultánea a los objetivos principales, se proponen una serie de objetivos específicos con los que se perfila la metodología que se va a seguir en este Trabajo de Fin de Máster. En concreto, se proponen los siguientes objetivos específicos:

1. La creación de un sistema híbrido de RI en español basado en el algoritmo BM25 y una red neuronal cross-encoder.

2. La creación de un módulo de preprocesamiento de texto en español y su ensamblado con el SRI híbrido en una nueva arquitectura.

3. La propuesta de un prototipo de nuevo conjunto de datos de evaluación para el español, y posterior evaluación de la arquitectura.

4. La comparación de los resultados respecto a los SRI más usados actualmente en español.

5. La implementación de la nueva arquitectura en un prototipo informático para evaluar su viabilidad en un caso de uso: la búsqueda, en tiempo real, de documentos en una base de datos documental alojada en un ordenador personal.

6. La evaluación empírica del funcionamiento del prototipo de aplicación con documentos con contenidos educativos del Máster en Letras Digitales: Estudios Avanzados en Textualidades Electrónicas.

De todos los objetivos específicos, el núcleo de este trabajo se conforma en los objetivos 1 y 2, la creación de la arquitectura del sistema híbrido de Recuperación de Información en español. El resto de objetivos específicos se dirigen a completar dicho sistema y a su evaluación. Antes de presentar el nuevo SRI para el español, en la siguiente sección se revisa el estado de la cuestión en RI.

Sección 3

Estado de la Cuestión – Recuperación de Información, Sistemas de Recuperación de Información y Evaluación

En esta sección se exponen las bases y antecedentes de la Recuperación de Información, una de las aplicaciones principales del Procesamiento del Lenguaje Natural. La sección está dividida en tres apartados.

Dentro del primer apartado se explica en qué consiste un sistema de RI y sus conceptos fundamentales.

En el segundo apartado se explican los fundamentos de los algoritmos y las redes neuronales en los que se basa la propuesta de nueva arquitectura y prototipo de SRI que se realiza en este trabajo.

Por último, el tercer apartado detalla los aspectos relacionados con la evaluación de los SRI.

3.1. La Recuperación de Información

La Recuperación de Información o Information Retrieval es una aplicación del Procesamiento del Lenguaje Natural que consiste en devolver al usuario la información (normalmente en forma de documentos de una base de datos) más relevante o cercana a la búsqueda o consulta que ha realizado (Robertson and Zaragoza 2009).

Esta definición distingue los elementos más importantes de esta tarea. El **usuario**, la **búsqueda** y la **información**. Sin embargo, cabe destacar otro concepto fundamental en relación con la búsqueda: **la necesidad de información** por parte del usuario, es decir, lo que motiva a los usuarios el usar un SRI.

Habitualmente, los usuarios sienten la necesidad de cambiar la búsqueda para mejorar los resultados obtenidos. De aquí surge otro de los conceptos más importantes en esta tarea: **la relevancia** de la información que devuelve el sistema. Siguiendo las palabras de Robertson y Zaragoza (2009, pp. 336-337) un resultado es relevante cuando puede satisfacer la necesidad de información del usuario. De este hecho surgen dos implicaciones:

- La relevancia está relacionada únicamente con la necesidad de información. No con el sistema de búsquedas ni con la base de datos.
- La relevancia es una propiedad binaria. Algo es relevante o no lo es[4].

Ahora bien, debido a la naturaleza de la relevancia, un Sistema de Recuperación de Información no sabe con certeza si un documento es relevante o no, simplemente **asume esa propiedad mediante un proceso probabilístico**.

Para calcular las probabilidades de que un documento sea relevante se usan las características léxicas del propio documento y de la búsqueda del usuario para computar la similitud semántica entre ambas.

Como explica el Probability Ranking Principle (Robertson 1977), si la ordenación final de los documentos corresponde a cómo deberían estar ordenados siguiendo el criterio de relevancia, podemos afirmar que el sistema los ha devuelto correctamente. En consecuencia, la evaluación de un SRI estará enfocada a comparar la ordenación de los documentos que devuelve el sistema con la ordenación en la que tendrían que estar dichos documentos conforme a la necesidad de información del usuario.

Como nota final de esta introducción a la RI, muchos de los avances en este campo están teniendo en cuenta otro factor relevante a la hora de recuperar información: el perfil del usuario (Saoud and Kechid 2016; Zhou et al. 2017; Azad and Deepak 2019). Este factor tiene una relación directa con el concepto de necesidad de información, ya que lo matiza teniendo en cuenta las muchas correlaciones implícitas entre el tipo de búsqueda que puede hacer un perfil de usuario concreto. En este trabajo se ha decidido aplicar un enfoque simple, centrado exclusivamente en la propia tarea de la Recuperación de Información y sin incluir este factor.

[4]Este punto es debatible. Blair 1990 expone que los límites de la relevancia no están bien definidos, por lo que infiere que pueda ser una propiedad gradual y no binaria

3.2. Sistemas de Recuperación de Información

Se conoce como SRI el conjunto de programas que se encargan de calcular la similitud semántica entre la consulta del usuario y el corpus de documentos para devolver al usuario los primeros k (*top-k*) documentos más relevantes. La figura 1 representa el esquema de un SRI básico.

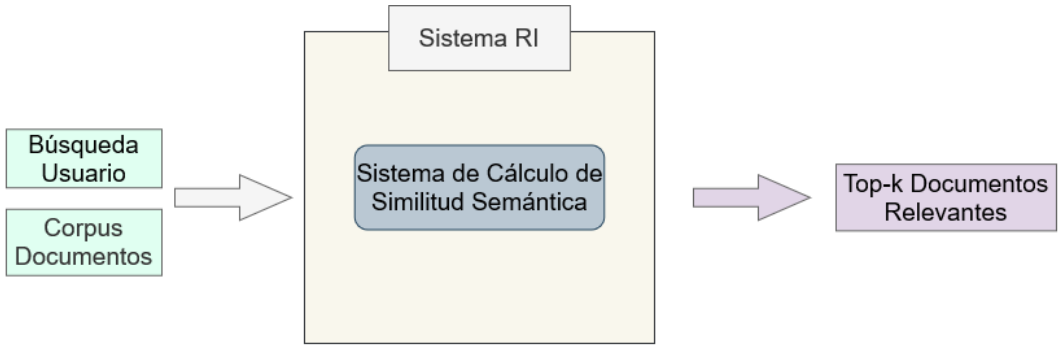

Figura 1: *Esquema de sistema de Recuperación de Información.*

Es importante señalar la distinción entre sistema y arquitectura. Una **arquitectura** (figura 2) comprende la unión de un SRI con uno o más módulos de preprocesamiento o post-procesamiento textual para optimizar la salida del SRI.

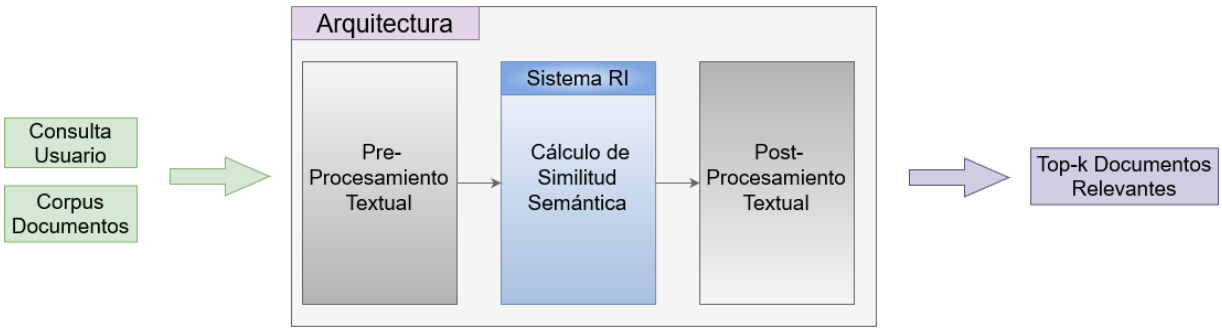

Figura 2: *Esquema de arquitectura de Recuperación de Información.*

Actualmente, los sistemas de Recuperación de Información que calculan la similitud semántica entre la consulta del usuario y los documentos se clasifican en 3 tipos, según el tipo de representación semántica que utilizan para los documentos y consultas (Luan et al. 2021):

- Sistemas basados en **representaciones dispersas** o "sparse retrieval systems"
- Sistemas basados en **representaciones densas** o "dense retrieval systems"
- **Sistemas híbridos** que fusionan ambos enfoques o "hybrid systems"

3.2.1. *Sparse Retrieval Systems*

La RI basada en representaciones dispersas (o "*Sparse Information Retrieval*") de los documentos[5] se basa en representar dichos documentos como vectores numéricos en base a un vocabulario, conjunto de términos o bolsa de palabras (en inglés "*Bag of Words*" o *BOW*).

Partiendo de este vocabulario (que, normalmente, contiene cada uno de los términos que aparecen en todos los documentos), cada documento puede ser representado por un vector de n-dimensiones (tantas como términos haya en el vocabulario) que incluirá el número de apariciones de cada término del propio documento, y otorgará un valor de 0 a todos aquellos términos que no aparezcan en el mismo. La figura 3 ilustra el proceso de transformación del texto a las representaciones vectoriales de las siguientes tres oraciones: "Mi perro no tiene hambre", "Mi primo tiene hambre" y "Tiene hambre y tiene un perro".

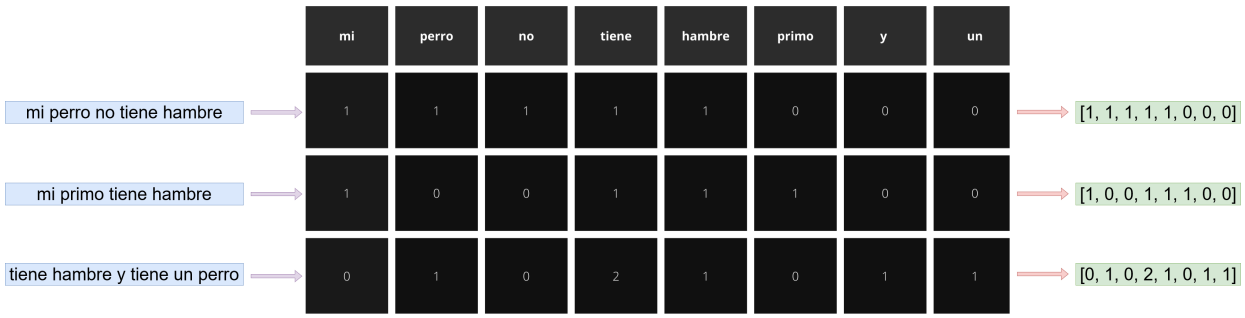

Figura 3: *Representación vectorial como BOW de tres oraciones.*

Como podemos ver, cada oración se convierte en un vector que representa tanto las palabras que incluye, como las que no. Convertir el texto a vectores hace posible que los sistemas informáticos puedan realizar las operaciones de comparación entre la consulta del usuario (que también se transforma en un vector) y los documentos a gran velocidad.

La denominación de "disperso" o "*sparse*" hace referencia al elevado número de ceros en cada vector que se corresponden con los términos no incluidos en el documento. El vocabulario de términos que se utiliza es muy amplio (del orden de decenas de miles) puesto

[5]Las consultas de los usuarios también se representan utilizando este mismo modelo de forma que, posteriormente, puedan compararse para calcular su similitud a nivel semántico.

que debe contener a todas las palabras que contienen miles o millones de documentos. Esto provoca que las representaciones distintas a 0 aparezcan de forma dispersa.

Este modelo de representación, si bien es sencillo de implementar y fácil de entender, implica la necesidad de recoger todas las formas distintas de cada lema que aparezca en la colección de documentos. Por este motivo, para reducir el número de términos, los SRI suelen incorporar una fase previa de preprocesamiento de documentos en sus arquitecturas. Este preprocesamiento está habitualmente constituido por:

(i) la eliminación de las llamadas palabras vacías o "*stopwords*"[6], (ii) la conversión del texto a minúsculas, y (iii) la conversión de cada término a su lema o raíz.

El preprocesamiento consigue que, además de facilitar el cálculo de las representaciones semánticas, el coste de almacenamiento de los textos se reduzca hasta un 30 % (Martínez Méndez 2004).

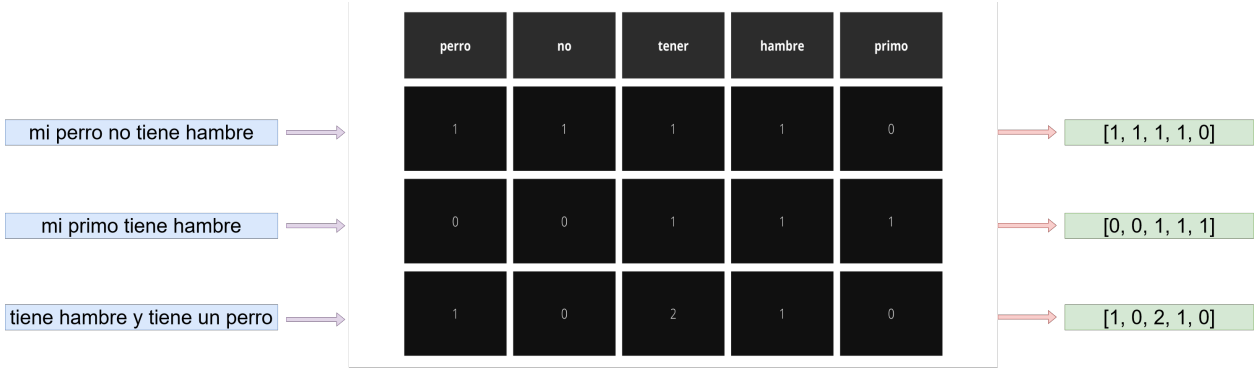

Figura 4: *Representación vectorial como BOW preprocesada.*

Para el cálculo de la similitud semántica, los SRI basados en representaciones vectoriales dispersas utilizan diversos algoritmos matemáticos como el **Okapi BM25** (Robertson and Zaragoza 2009) o datos estadísticos como el **tf-idf** (Sparck Jones 1972). En concreto, el cálculo de la similitud gira alrededor de un concepto fundamental llamado **peso**. Cada documento tiene un peso único en función de su relevancia respecto a la búsqueda del usuario. La idea base es que cuantas más **palabras coincidan** entre la búsqueda y un documento, más relevante será este y más peso tendrá. Es importante señalar que ambos métodos tienen en consideración más variables; por ejemplo, en el caso del tf-idf se tiene en cuenta el número total de palabras del documento, mientras que el algoritmo BM25 usa la longitud media de todos los documentos.

El funcionamiento interno de los SRI dispersos ofrece dos ventajas fundamentales. Por

[6]Se conocen como *stop-words* o palabras vacías a aquellas palabras sin significado léxico.

un lado, los SRI dispersos son agnósticos al idioma, únicamente calculan si las palabras de la búsqueda se repiten a lo largo del corpus. Por otro lado, el cálculo de similitud es un proceso extremadamente rápido. Esta es la principal razón de que hoy en día estos métodos se sigan utilizando, ya que sus resultados continúan siendo robustos y apenas tienen costes computacionales (véase sección 5.3).

A pesar de estas ventajas, los SRI dispersos sufren lo que se conoce como diferencia léxica o "***lexical gap***" (Berger et al. 2000). El *lexical gap* se refiere al problema de que las palabras introducidas en la búsqueda del usuario pueden no coincidir con las de los documentos a pesar de referirse al mismo significado (sinónimos, hipero o hipónimos directos, variantes léxicas, errores ortográficos, etc); en estos casos, los SRI que utilizan representaciones dispersas son incapaces de encontrar documentos con las formas alternativas de las palabras de la búsqueda del usuario y, por lo tanto, devolver adecuadamente los documentos similares.

Otra desventaja a tener en cuenta es que los documentos se almacenan como un vector de palabras plano, que no tiene en cuenta la ordenación de las palabras en el texto ni las distintas dependencias sintácticas dentro de cada oración, con lo que se pierde gran cantidad de información contextual.

3.2.2. *Dense Retrieval Systems*

Las limitaciones de los SRI dispersos han llevado a muchos autores a buscar alternativas. En concreto, se ha optado por sistemas que utilizan representaciones más completas semánticamente.

Este tipo de SRI hace frente a los problemas de los sistemas dispersos, *lexical gap* y orden de las palabras, incorporando en las representaciones factores como el contexto o significado léxico de las palabras. En este sentido, este tipo de SRI ha sido potenciado en los últimos años mediante los avances de la Inteligencia Artificial. Especialimente, desde el nacimiento de las redes neuronales de arquitectura[7] *Transformer* (Vaswani et al. 2017).

Definición

Se conoce a este tipo de búsqueda como la "**basada en representaciones densas**" o "*Dense Information Retrieval*" ya que está fundamentada en las **representaciones semánticas complejas** de las oraciones (tanto de la búsqueda del usuario como de los documentos) como medida de similitud.

Estas representaciones, al igual que las representaciones dispersas, buscan codificar el

[7]Cabe destacar que al hablar de redes neuronales, el término "arquitectura" adquiere un significado diferente al usado en el contexto de la Recuperación de Información

significado de los documentos en un vector numérico de n-dimensiones. La diferencia con las representaciones dispersas está en que las densas utilizan los rasgos semántico-sintácticos de cada palabra de la oración, y no necesitan un vocabulario preestablecido.

Una de las primeras propuestas de representación vectorial densa de documentos es el denominado "Word2Vec" (Mikolov et al. 2013). Esta propuesta consiste en la codificación de las representaciones de cada palabra en base a la semántica distribucional y a los rasgos distintivos que posee esa palabra.

La figura 5[8] ilustra cómo son las representaciones vectoriales del "Word2Vec".

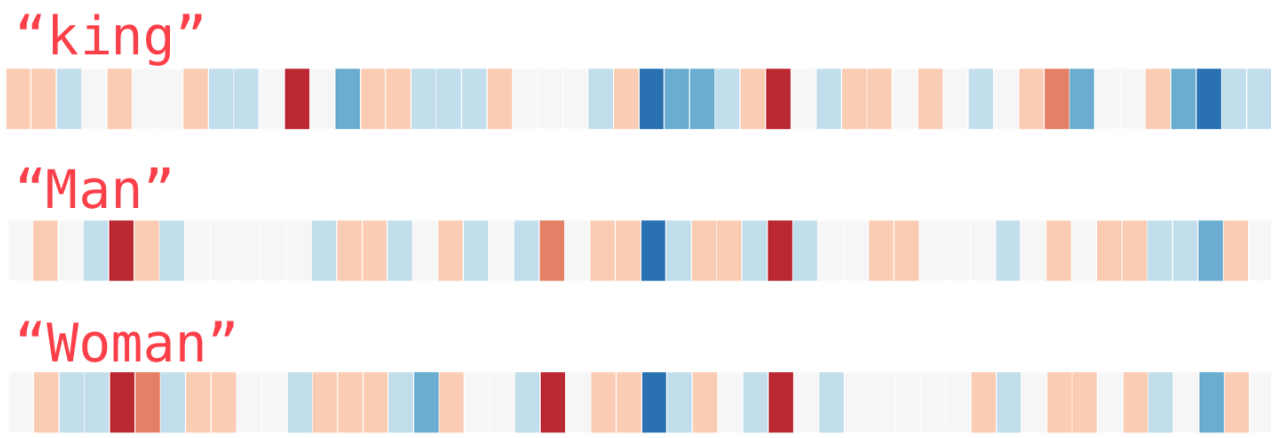

Figura 5: *Representaciones vectoriales de las palabras "rey", "hombre" y "mujer".*

Como se puede observar, cada palabra está representada por un conjunto de dimensiones que representan cierto rasgo léxico; de este modo palabras con significado parecido tendrán una representación similar al estar compuestas por rasgos semánticos cercanos[9] ("mujer" comparte más rasgos con "hombre" que con "rey"). También, a la hora de posicionarlas en un espacio vectorial, se encontrarán cerca unas de otras (como se puede apreciar en la Figura 6).

Actualmente, estas representaciones vectoriales son llamadas ***embeddings***. Jurafsky and Martin 2018 definen los embeddings como la representación semántica vectorial de un lema formada por un conjunto de rasgos lingüísticos extraídos automáticamente de gigantescos corpus textuales. Estos rasgos incluyen significado, propiedades morfosintácticas, relaciones o propiedades asociadas a la semántica distribucional de cada palabra.

[8]Alammar, J. (2019). The illustrated Word2Vec. Recuperado de https://jalammar.github.io/illustrated-word2vec/. Los rectángulos corresponden a cada una de las dimensiones del vector.

[9]Estos rasgos están representados por un valor numérico, que en la imagen se representa por un color determinado. Cuantos más colores coincidan, más parecidas serán sus representaciones.

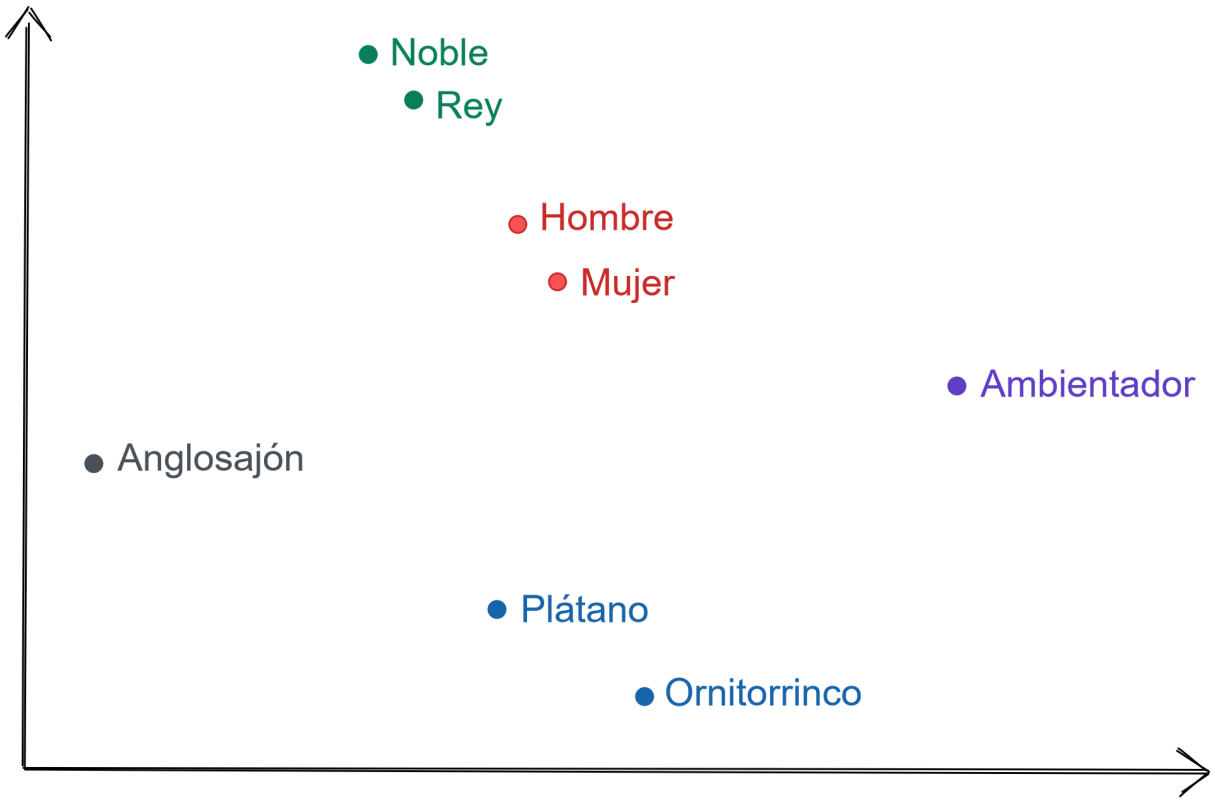

Figura 6: *Espacio vectorial que distribuye distintos lemas en función de su representación vectorial. Los lemas con más rasgos similares ocupan espacios cercanos en el espacio.*

Cada rasgo extraído se codifica con un valor numérico en una dimensión de un vector n-dimensional que en su conjunto representa cada palabra de un idioma. Distintos modelos de representación usan un número de dimensiones diferentes, siendo los más comunes 300 como Word2Vec, 512 como GloVe (Pennington et al. 2014) o 768 como WordPiece (Schuster and Nakajima 2012).

Sin embargo, es muy importante destacar que estas representaciones han sido pre-calculadas, es decir, son representaciones estáticas que no incluyen contexto. Esto implica que en muchas ocasiones un *embedding* puede no corresponder al significado de un lema en un contexto concreto; por ejemplo, si codificamos la expresión "ojo de buey" es bastante probable que la representación vectorial de "ojo" se asemeje más a "parte de animal" que a "ventana".

Para solventar este problema, los *embeddings* se pueden "actualizar" con el contexto de la oración de la que forman parte. Esta técnica fue potenciada cuando, en 2017, apareció por primera vez un nuevo tipo de red neuronal capaz de computar grandes cadenas de texto utilizando el contexto global en cada uno de sus elementos. Este tipo de red neuronal

bautizada como "***Transformer***" (Vaswani, 2017) es capaz de codificar cada elemento de una oración teniendo en cuenta el contexto que le rodea mediante un mecanismo llamado atención.

Posteriormente, surgió una nueva red especializada en la detección de la similitud semántica entre dos oraciones. Esta red, conocida como ***SentenceTransformers*** (Reimers and Gurevych 2019), permite codificar de forma eficaz el significado de todo un texto en representaciones vectoriales oracionales mediante el uso de **modelos de lenguaje pre-entrenados**. Estos modelos de lenguaje se consideran el núcleo de estas redes neuronales, ya que son los encargados de sintetizar las características semánticas y fusionar el resultado con el contexto de cada palabra, para crear, de este modo, cada representación vectorial de los elementos de la oración.

Una vez se han calculado todos los *embeddings*, para calcular una representación vectorial oracional el sistema utiliza varias estrategias, entre las más destacadas encontramos el *MaxPooling* y el *MeanPooling*.

El *MaxPooling* crea una representación vectorial oracional partiendo de los valores más elevados de cada dimensión de los *embeddings* de las palabras; podemos pensar en esta técnica como **la unión de los rasgos más distintivos** de cada *embedding* de la oración para la creación del vector o *embedding* oracional.

El *MeanPooling* aplica una operación matemática a todas las representaciones de las palabras que conforman la oración para obtener la **media aritmética** de cada dimensión de los vectores de palabras. Obteniendo en este caso un *embedding* oracional más equilibrado que busca sintetizar el significado promedio de cada uno de sus integrantes.

Para el cálculo de la similitud semántica entre dos oraciones, Reimers propone dos tipos de arquitecturas de redes neuronales:

(i) ***Los bi-encoder***, que pre-calculan los embeddings de dos oraciones por separado para más adelante compararlos mediante los algoritmos de similitud semántica.

(ii) ***Los cross-encoder***, que concatenan las dos oraciones para generar su representación vectorial y calcular la posibilidad de que ambas oraciones sean similares.

Como puede suponerse esta arquitectura de redes neuronales ha contribuido de manera excepcional al desarrollo de nuevos sistemas de Recuperación de Información.

A modo de recapitulación: los sistemas de representaciones densas utilizan, por lo tanto, las **representaciones vectoriales** de los documentos como **peso** a la hora de devolver los documentos en orden de similitud. Para ello, codifican el corpus de documentos y la consulta del usuario en *embeddings* y aplican operaciones de cálculo de similitud semántica

entre estos.

Fases en la Recuperación de Información en los SRI densos

Para calcular el peso de un documento respecto a la búsqueda del usuario, los sistemas basados en representaciones densas dividen el proceso en dos fases:

- El procesamiento de los documentos y de la búsqueda del usuario para obtener sus representaciones vectoriales mediante un modelo pre-entrenado de lenguaje.
- El cálculo de similitud entre la representación de la búsqueda y la representación de los documentos mediante algoritmos matemáticos como la distancia de coseno o el producto escalar.

El corpus de documentos suele ser convertido a representaciones vectoriales de antemano, de este modo, en la inferencia los tiempos de respuesta a las consultas del usuario son asequibles. Excepcionalmente, es posible realizar la codificación del corpus de forma síncrona al momento de consulta, pero la carga computacional de estos sistemas es mucho más elevada.

Por último, se debe destacar que el cálculo de similitud entre representaciones se computa mediante diversas operaciones matemáticas, como pueden ser la distancia de coseno, el producto escalar o, la recientemente propuesta, distancia Arias (Rodríguez 2022). Este tipo de operaciones entre *embeddings* apenas suponen costes temporales o de procesamiento (véase sección 5.3).

En resumen, los **SRI densos** surgen como una **solución a los problemas léxicos** que sufren los SRI dispersos. El poder introducir el contexto de las palabras en sus representaciones vectoriales ayuda a que el proceso de búsqueda no dependa de palabras clave, sino que **utiliza el significado de las consultas** de los usuarios para encontrar documentos afines. Sin embargo, estos sistemas necesitan una **capacidad de procesamiento superior**, ya que se necesita calcular las representaciones oracionales de todo el corpus y de cada consulta que hagan los usuarios. La tabla 1 resume las principales características y carencias de ambos sistemas.

3.2.3. Sistemas híbridos

Para solventar los problemas computacionales de los modelos basados en representaciones densas, durante estos dos últimos años (Wang et al. 2020, Gao et al. 2021) han surgido modelos híbridos que combinan la rapidez y eficiencia de los modelos dispersos con la precisión de los modelos densos.

La figura 7 representa el funcionamiento de este tipo de sistemas. Se diferencian claramente dos fases consecutivas: (i) la **fase de recuperación de documentos** relevantes en

	SRI Disperso	SRI Denso
Documentos	Representaciones como bolsa de palabras	Representaciones como vectoriales contextuales
Cálculo similitud semántica	Algoritmos matemáticos como el Okapi BM25	Distancia Coseno Producto Escalar
Ventajas	Rapidez y simpleza. Es agnóstico al idioma	Precisión y búsquedas semánticas
Desventajas	Lexical Gap	Coste computacional muy elevado. Dependencia de modelos de lenguaje pre-entrenados

Tabla 1: *Comparación SRI dispersos vs. SRI densos*

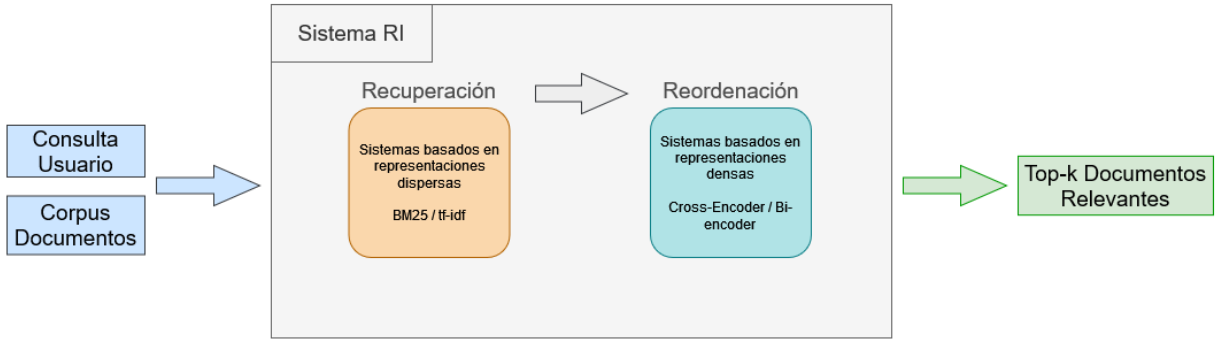

Figura 7: *Esquema de sistemas de Recuperación de Información híbridos.*

función a la salida de un sistema de representación dispersa, y, (ii) la **fase de reordenación** en base a la similitud semántica (calculada mediante un sistema de representación densa) entre los documentos devueltos y la búsqueda del usuario.

La propuesta de este trabajo va a utilizar como componente principal de su diseño y desarrollo a este tipo de sistemas. En concreto, se va a partir del trabajo de Wang et al. 2020, donde se presenta un modelo pre-entrenado de lenguaje llamado "MiniLM". Este modelo ha sido utilizado como núcleo de la fase de reordenación de un SRI conocido como BM25+CE (BM25 + *Cross-Encoder*). Este sistema utiliza el algoritmo BM25 base para la fase de recuperación de información, y una red neuronal de arquitectura *cross-encoder* con el modelo "MiniLM" como núcleo para la fase de reordenación. Los resultados obtenidos por este sistema lo colocan en la primera posición del estándar de comparación BEIR.

Este enfoque está logrando posicionarse como el mejor método para la Recuperación de Información, ya que combina la rápida recuperación de los documentos similares gracias a los modelos dispersos con la alta eficiencia en la ordenación de los modelos densos.

3.3. Evaluación de los sistemas de Recuperación de Información

Como se ha mencionado anteriormente, el aspecto más importante para la evaluación de los SRI es la corrección de la relevancia de los documentos recuperados. Esto significa que se debe evaluar si el SRI devuelve ciertos documentos en el orden correcto respecto a las consultas de los usuarios. Para realizar este proceso de evaluación se necesitan dos elementos fundamentales: un **conjunto de datos** como base, y unas **métricas específicas** para medir los resultados.

3.3.1. Los conjuntos de datos de evaluación

Para evaluar los SRI normalmente se opta por diseñar, manualmente, un conjunto de búsquedas sobre documentos preestablecidos, y se define, para cada búsqueda, la lista de los k documentos más relevantes. Estos datos se organizan en pares documento-búsqueda para obtener lo que se denomina un conjunto de datos de evaluación. Este conjunto de datos se utiliza para **comparar sistemáticamente el resultado obtenido con el esperado**.

Entre los conjuntos de datos más destacados actualmente en inglés (idioma predominante de la RI) encontramos el "MS MARCO" (Bajaj et al. 2018), el "NQ" (Kwiatkowski et al. 2019) o el "TREC" (con cualquiera de sus variantes). Mientras que para el español solo destaca el "mMARCO" (Bonifacio et al. 2021) que es un conjunto de datos traducido automáticamente del mencionado "MS MARCO" avarios idiomas, incluido el español. Esta traducción ha generado, como se puede comprobar en la sección 5.1, errores al asociar las consultas traducidas automáticamente con documentos que han dejado de ser relevantes o no son los más relevantes.

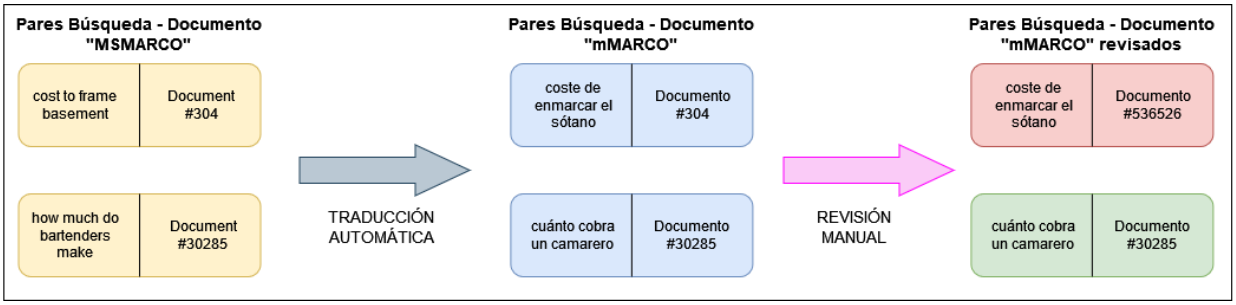

Figura 8: *Ejemplos de errores en el conjunto de datos causados por la traducción automática.*

La figura 8 muestra dos casos del efecto provocado por la traducción del conjunto de datos "MS MARCO". Se trata de dos consultas traducidas automáticamente al español que conservan los documentos asociados originales como los más relevantes. Como se observa, una revisión manual muestra que, mientras que en el segundo caso se sigue manteniendo la

relación de relevancia, no ocurre lo mismo para el primer caso, donde el documento asociado ya no es el más relevante.

En la sección cinco se profundizará sobre este problema, y sobre la necesidad de contar con otro conjunto de datos para el español.

3.3.2. Métricas

Otra de las cuestiones que hay que tener en cuenta cuando se evalúa un SRI es la necesidad de cuantificar la eficiencia de la fase de recuperación de documentos y la fase de ordenación por separado.

Actualmente, en el ámbito de la RI encontramos principalmente tres métricas para calcular la eficiencia de los SRI evaluados: una para la fase de recuperación y otras dos para la fase de ordenación de documentos:

1. En la fase de recuperación, la única medida utilizada para evaluar su eficacia es la **cobertura** o *recall*. Esta medida se calcula como la fracción de documentos que se han recuperado correctamente dentro de los *top-k* primeros en función a un umbral de relevancia. En otras palabras, esta medida calcula si todos los documentos relevantes han sido localizados por el sistema sin tener en cuenta la posición que ocupen.

2. En la fase de ordenación de los documentos se suelen usar dos medidas, la **nDCG** o *Normalized Discounted Cumulative Gain* (Järvelin and Kekäläinen 2002) y el **MRR** o *Mean Reciprocal Rank* (Voorhees and Tice 2000).

   La nDCG calcula una puntuación especialmente útil para los conjuntos de evaluación en los que cada consulta tiene más de un documento asociado. En concreto, se penaliza severamente a los sistemas que ordenan en baja posición cualquiera de los documentos relevantes.

   La MRR se centra en comparar la posición del documento asociado como más relevante con la posición que ocupa en la ordenación. Es decir, solo tiene en cuenta la posición del documento más relevante, por lo que es ideal para conjuntos como el "MS MARCO" en los que solo suele haber un documento relevante por cada consulta.

En la sección cinco, las medidas de cobertura y MRR son usadas para evaluar la nueva arquitectura propuesta.

Sección 4

Nueva arquitectura de Sistemas de Recuperación de Información para el español

Esta sección presenta la aportación central de este trabajo: la creación de una arquitectura de sistemas híbridos de RI para el español y su evaluación empírica mediante un prototipo informático.

4.1. Diseño de la nueva arquitectura

La mayoría de los avances en RI presentan sistemas independientes, que suelen necesitar elementos extra para alcanzar su máximo potencial. En este trabajo se ha pretendido llegar más lejos presentando una arquitectura completa, que incluye todos los procesos textuales que este sistema de Recuperación de Información necesita para funcionar de manera óptima.

La figura 9 representa el diseño de la arquitectura propuesta. Esta arquitectura se compone de dos componentes: (i) el sistema de preprocesamiento previo del texto, y (ii) el SRI híbrido de dos capas para la recuperación de documentos y su ordenación.

El primer componente, un módulo previo de procesamiento textual, se encarga de normalizar y reducir el tamaño del texto de los documentos y de las consultas con el fin de simplificar y acelerar los procesos del SRI.

Complementariamente, el segundo componente está inspirado en el funcionamiento del mejor sistema del estándar de comparación BEIR (Thakur et al. 2021), el "BM25+CE". Este componente se divide en una primera capa encargada de la fase de recuperación de documentos mediante el uso de una **variación del algoritmo Okapi BM25**, y una segunda capa para la fase de reordenación basada en el uso de una **red neuronal de arquitectura**

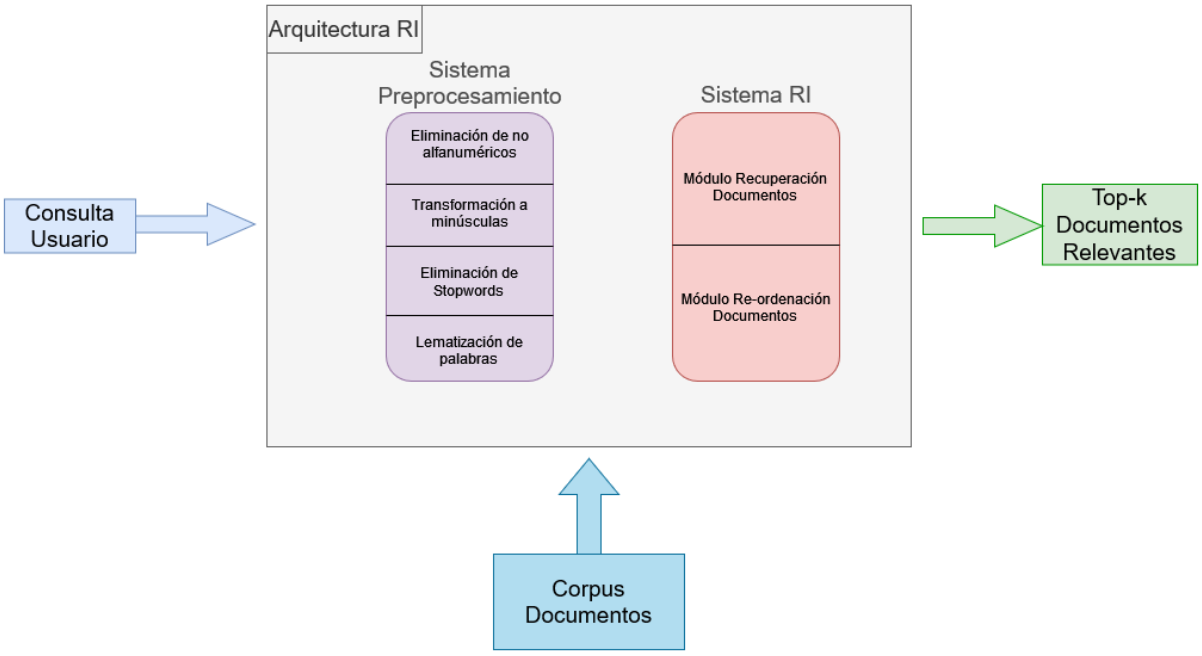

Figura 9: *Arquitectura general propuesta para el sistema híbrido RI del español.*

Cross-encoder.

En las siguientes subsecciones se explican cada uno de los módulos.

4.1.1. Sistema de preprocesamiento

Para agilizar los procesos del SRI híbrido, esta arquitectura incorpora un sistema previo de preprocesamiento textual cuyo objetivo es normalizar y simplificar la consulta y los documentos. Este sistema realiza varias tareas habituales de preprocesamiento textual que combinará para optimizar los procesos del SRI.

Entre estas técnicas se encuentran:

- Eliminación de los caracteres no alfanuméricos
- Transformación del texto a minúsculas
- Omisión de las palabras vacías
- Lematización de las palabras

Es importante destacar que esta nueva arquitectura preprocesa la consulta del usuario y el corpus de documentos dos veces, una para cada fase del SRI:

Para la fase de recuperación de documentos, todas estas técnicas se utilizan para procesar el corpus de documentos y la consulta del usuario, aligerando la carga computacional y generando inferencias más rápidas.

Para la fase de reordenación, solo se aplican las dos primeras técnicas, ya que el proceso de creación de representaciones oracionales densas requiere que las oraciones originales no sean modificadas[10].

Estudio de la importancia de las técnicas de preprocesamiento

De cara a entender el impacto de estas técnicas de procesamiento textual en el rendimiento de la arquitectura, se ha creído conveniente dedicar una sección del trabajo a estudiar cómo afecta cada una de las técnicas al resultado final de la evaluación.

Con este objetivo se ha realizado un **estudio de extirpación de los componentes** ("*ablation study*", en inglés), para cuantificar el efecto global de cada una de las técnicas de preprocesamiento.

Este estudio ha revelado que la técnica que más afecta a la fase de recuperación de documentos es la **lematización** de las palabras. A su vez, la **eliminación de los caracteres no alfanuméricos** ha supuesto una mejora considerable en el SRI. Se cree que el motivo de esta mejora es que ambas técnicas han ayudado a normalizar la entrada textual del sistema. Los resultados en detalle y análisis del estudio se encuentran disponibles en el Anexo A.

4.1.2. El sistema híbrido de Recuperación de Información

Este sistema se encarga de computar la puntuación de similitud entre la consulta del usuario y los documentos, calculando cuáles son relevantes y en qué orden. Para ello se utilizan dos módulos independientes: el módulo de recuperación inicial de documentos mediante el cálculo de similitud Okapi BM25 y el módulo de ordenación final mediante el cálculo de similitud de una red neuronal *cross-encoder*.

El resultado es un SRI híbrido que mezcla representaciones dispersas con representaciones densas, aprovechando la rapidez de las primeras y la eficiencia de las segundas.

[10]Los modelos Transformers (Vaswani et al. 2017) calculan la atención de cada elemento de la oración en base al contexto y a las relaciones que forma con el resto de elementos de la oración. Por lo que es conveniente no modificar el texto original.

El módulo de recuperación de documentos mediante el cálculo de similitud Okapi BM25

Este módulo se encarga de la recuperación de documentos, en concreto, su trabajo es codificar la información de entrada del sistema (tanto la consulta del usuario como el corpus de documentos) para calcular posteriormente la similitud semántica entre la consulta y cada uno de los documentos del corpus. Este cálculo de similitud se realiza mediante una modificación del algoritmo Okapi BM25 previamente configurada. Por último se devuelve al SRI híbrido los *top-k* documentos más relevantes según la consulta del usuario.

El proceso que se realiza en este módulo es el siguiente:

1. Entrada inicial del corpus de documentos
2. Codificación del corpus en representaciones dispersas en forma de bolsas de palabras (BOW) e indexado en el sistema
3. Entrada de la consulta del usuario
4. Codificación de la consulta en BOW (necesitando previamente el vocabulario del corpus de documentos)
5. Cálculo de similitud entre los vectores BOW de la consulta y el corpus mediante una variación del algoritmo matemático Okapi BM25
6. Salida de los k documentos más relevantes ordenados por el resultado del paso anterior

Como se puede observar, hay una diferencia fundamental con el sistema que ha servido de inspiración. El cálculo de similitudes se realiza mediante una **variante del algoritmo Okapi BM25** (la adición de un parámetro extra y la modificación de dos parámetros previos) para calcular el grado de similitud (score) entre un documento D y una consulta Q. La fórmula que sigue este algoritmo es la siguiente:

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_1) \cdot \left[\frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})} + \delta \right]$$

Donde se utiliza la siguiente notación:

- la puntuación de similitud (score)
- el documento (D)
- la consulta del usuario (Q) y cada uno de sus términos (qi)

Algoritmo BM25	**k1**	**b**	**delta (δ)**
Este Enfoque	1.5	0.75	0.25
MiniLM	0.9	0.4	0

Tabla 2: *Comparativa de los parámetros entre nuestro sistema de recuperación de documentos y el usado en el BM25+CE.*

- el sumatorio de las frecuencias inversas de cada uno de los n términos de la consulta (IDF(qi))
- la frecuencia (f) de cada término en términos probabilísticos respecto a la distribución de Poisson
- la longitud del documento (|D|)
- la media de las palabras del corpus de documentos (avgdl)
- los parámetros modificables k1, b y delta (δ).

En resumen, el algoritmo BM25 utiliza la suma del *score* de cada término de la consulta para calcular cómo de relevante es el documento que se está comprobando. Para ello, se tienen en cuenta **factores** como la **longitud del documento** (penalizando a documentos cuya longitud es menor que la media del corpus) o la **frecuencia de cada término** (cuanto más veces aparezca en el documento, más relevante será este). Finalmente, la fórmula se puede **optimizar cambiando los parámetros** k1 (el peso que tiene el número de repeticiones de los términos; cuanto más bajo, menos importa), b (el peso asociado a la longitud de los documentos; cuanto más alto, más se tiene en cuenta a los documentos cortos del corpus) y delta (el peso asociado a la varianza de la longitud de documentos; cuanto más bajo, más se penalizan los documentos largos).

En este caso, se han utilizado los valores de los parámetros k1, b y delta que se muestran en la tabla 2. Estos valores se han determinado empíricamente con el objetivo de penalizar al mínimo los cambios de longitud de documentos, ya que el sistema ha sido diseñado para la recuperación de textos académicos, los cuales pueden tener una longitud muy variable. Esta tabla muestra, también, los valores del SRI que utiliza el modelo "MiniLM" propuesto por Wang et al. 2020.

Para recapitular, en la fase de recuperación de documentos el sistema híbrido se comporta de la misma forma que un SRI disperso, en primer lugar **transforma la búsqueda** del usuario a su representación como bolsa de palabras, y posteriormente **compara** esta representación con el **corpus de documentos** para encontrar los *top-k* resultados más relevantes mediante nuestra variante del algoritmo Okapi BM25.

Por último, es importante señalar:

(i) el sistema puede devolver cualquier documento que contenga las mismas palabras que la búsqueda del usuario, por lo que es conveniente poner un **límite de documentos**. En nuestro caso, este umbral es de siete documentos al ser una arquitectura pensada para corpus pequeños.

(ii) el usar un sistema de representaciones dispersas como forma de recuperar documentos **imposibilita superar el "*lexical gap*"** que estos sistemas arrastran. Sin embargo, las ventajas computacionales que ofrecen estos sistemas son demasiado convenientes.

El módulo de ordenación SentenceTransformer

El segundo módulo usa una red neuronal *SentenceTransformer* de arquitectura ***cross-encoder*** para **reorganizar la ordenación** de los documentos relevantes devueltos por el módulo anterior.

Como se ha mencionado en la sección 3.2 , este tipo de red neuronal **concatena la consulta** del usuario **con cada uno de los documentos** devueltos por el módulo anterior para **calcular una puntuación que represente el grado de similitud** entre ambos textos. En este proceso es necesario que la red *cross-encoder* cuente con un modelo pre-entrenado de lenguaje que pueda codificar el texto a su representación densa.

El cálculo de similitud semántica sigue el esquema de la figura 10. En primer lugar se introduce en el *cross-encoder* la consulta y el documento, luego se concatenan ambas, y un clasificador genera una puntuación de similitud semántica en función del *embedding* resultante de las oraciones concatenadas. Esta puntuación fluctúa entre 0 y 1 (siendo 0 oraciones nada similares, y 1 oraciones idénticas). [11]

Para este trabajo se ha realizado un estudio complementario que ha analizado todos los modelos pre-entrenados *SentenceTransformer* disponibles para el español (véase Anexo B). Tras valorar los resultados se ha seleccionado el **modelo "mmarco-mMiniLMv2-L12-H384-v1"** como núcleo de la red neuronal *cross-encoder*.

El uso de este modelo de lenguaje diferencia nuestro enfoque del sistema BM25+CE, en cuanto a que el modelo que hemos elegido ha sido entrenado mediante *fine-tuning*[12] para la detección de similitud semántica en español. Mientras que el "MiniLM" no ha sido entrenado para esta tarea. En nuestra opinión este hecho constituye una ventaja frente a la propuesta original porque permite que la red neuronal se ajuste mejor a la tarea de cálculo de similitud.

[11] Para más información sobre el funcionamiento interno de este módulo a nivel de código acudir a https://github.com/UKPLab/sentence-transformers

[12] Se conoce como "fine-tuning" al proceso de adaptación de un modelo pre-entrenado de lenguaje a una tarea específica

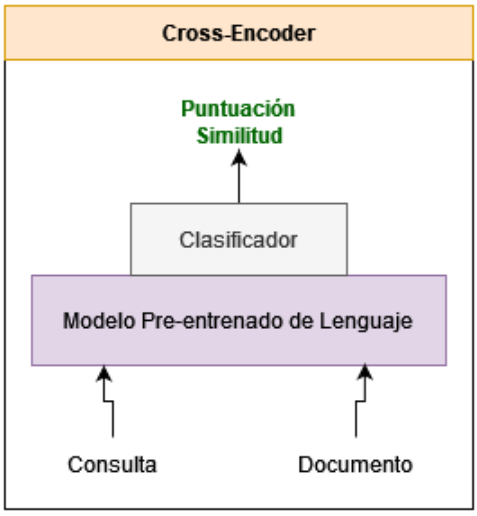

Figura 10: *Arquitectura interna de un cross-encoder*

El motivo por el que este módulo se encarga únicamente de la reordenación de los resultados (y no de todo el proceso) es doble. En primer lugar, los recursos computacionales necesarios para el funcionamiento de una red *SentenceTransformer* son muy costosos, pero además, una arquitectura *Cross-Encoder* requiere calcular la similitud semántica de n pares búsqueda-documento en el momento de la búsqueda (no se pueden pre-calcular las representaciones vectoriales porque no se dispone de las búsquedas que realizarán los usuarios), volviendo al sistema virtualmente inaccesible[13].

Por ello, es el módulo basado en BM25 quien realiza el filtro previo de recuperación de documentos, limitando así la lista de documentos relevantes que son comparados con la consulta. En otras palabras, se trata de que el *cross-encoder* refine el cálculo de similitud que realiza el módulo BM25 sin sobrecargar el sistema.

En segundo lugar, aunque actualmente los vectores o *embeddings* oracionales consiguen reflejar, en cierta forma, los rasgos semánticos de las palabras que componen la oración, aún no se ha conseguido recoger por completo el significado de las oraciones en estas representaciones. Este hecho se refleja en el escaso número de sistemas que utilizan únicamente las representaciones semánticas densas como método para la Recuperación de Información (a pesar de que uno de esos sistemas, el ColBERT -Khattab and Zaharia 2020-, está posicionado

[13]El tiempo de inferencia para una búsqueda en un corpus de 4000 documentos usando 4 GPU Nvidia Tesla v100 32GB es de 17 segundos. El script usado para este experimento se encuentra disponible en el repositorio Github del trabajo.

como segundo mejor sistema del estándar de comparación BEIR).

Ambos factores se han tenido en cuenta a la hora de diseñar esta arquitectura, optando finalmente por una mezcla entre un sistema de representaciones dispersas como método de filtrado, y un sistema de representaciones densas como un método de reordenación.

En definitiva, el funcionamiento del módulo de reordenación es el siguiente:

1. Se introducen los k documentos, en su versión original, que devuelve la primera fase del sistema.
2. Se introduce la consulta del usuario, también en formato texto.
3. Por cada uno de los documentos, se concatena este y la consulta del usuario.
4. Se utiliza la red *cross-encoder* para calcular una puntuación de similitud semántica.
5. Se reordena a los documentos en función de esta puntuación
6. Se devuelven al usuario en forma de la salida final del SRI

4.2. Propuesta de prototipo de Sistema de Recuperación de Información

Este diseño de arquitectura ha **cumplido la primera parte del objetivo (i)** de este trabajo de fin de máster. Queda, sin embargo, evaluar la viabilidad de esta propuesta. Para ello se implementa la arquitectura en un prototipo de SRI para el español que trabajará sobre un conjunto de documentos concreto: los apuntes de las diferentes asignaturas del *Máster en Letras Digitales*. De esta forma se aborda la segunda parte del primer objetivo general del TFM.

El diseño final que se va a usar como base se muestra en la figura 11. En este esquema se puede observar cómo está conectada la búsqueda del usuario y el corpus de documentos con el sistema de preprocesamiento y con el sistema híbrido de Recuperación de Información.

Para comprobar el funcionamiento de este diseño se ha creado una aplicación informática que usa la arquitectura híbrida propuesta como motor de búsqueda sobre un corpus de apuntes, y que añade un módulo de post-procesamiento para facilitar la visualización de los resultados por el usuario. En concreto, se ha decidido recopilar, procesar y almacenar distintos documentos académicos de diversas asignaturas del Máster de Letras Digitales. El código de la aplicación se encuentra disponible en un repositorio online de libre acceso[14].

[14] https://github.com/aardoiz/ir_system

Además de servir como evaluación empírica, esta aplicación, bautizada como "Recuperador de Apuntes", nace de la motivación provocada por la frustración que genera el disponer de cantidades inmensas de documentos y no poder realizar búsquedas sencillas sobre ellos para su estudio. El "Recuperador de Apuntes" busca ocupar este hueco permitiendo realizar búsquedas transversales sobre colecciones personales de documentos sin invadir la privacidad de sus usuarios.

Todos los componentes del "Recuperador de Apuntes" han sido programados siguiendo esta idea. En concreto, se ha desarrollado un módulo de post-procesamiento, que utiliza la salida del SRI y añade al texto de los documentos etiquetas semánticas de lenguaje HTML para resaltar los términos que coinciden con la consulta del usuario.

Para optimizar la experiencia de usuario, el "Recuperador de Apuntes" incluye un sistema adicional de conversión de archivos (gracias al cual se ha podido recopilar el corpus de documentos académicos), permitiendo añadir documentos personales a la base de datos de nuestra arquitectura.

Por último, para facilitar el uso de esta aplicación, se ha desplegado un prototipo de muestra[15] en la web utilizando la herramienta HerokuAPP, disponible a través del siguiente enlace[16].

La interfaz gráfica del prototipo se ha desarrollado utilizando el *framework Vue.js* para facilitar el uso de la aplicación a aquellos usuarios que no tengan conocimientos previos de programación. Además, se incluye una documentación, donde se explica el funcionamiento básico del "Recuperador de Apuntes". Actualmente, los usuarios pueden realizar búsquedas relacionadas con los dos corpus de documentos disponibles:

- El corpus de apuntes de Letras Digitales

- El corpus de evaluación "RISQAC"

Las Figuras 12-14 muestran el proceso de uso para buscar un documento relacionado con "tipos de fuente" en el corpus de apuntes. En primer lugar, en la parte superior de la aplicación aparece un recuadro de búsqueda (con la consulta predeterminada "Tecnologías de la Información"), y cinco botones (figura 12).

Para utilizar esta aplicación hay que introducir la consulta deseada en el recuadro específico para ello y pulsar el botón de búsqueda sobre el corpus de apuntes (figura 13).

[15]Una versión reducida del "Recuperador de Apuntes" con una sección de las bases de datos.

[16]`https://aardoiz-tfm.herokuapp.com/app` - Por desgracia a partir de noviembre de 2022 HerokuAPP cesó sus servicios gratuitos y el enlace ya no está disponible

Finalmente, la aplicación devuelve en tiempo real los resultados, resaltando en amarillo (gracias al módulo de post-procesamiento) las partes de cada título o documento que ha considerado relevantes (figura 14). Como se puede observar, la aplicación ha devuelto correctamente documentos de la asignatura *Edición* asociados al campo de la tipografía. Resultado que confirma el funcionamiento adecuado de nuestra arquitectura SRI.

En la siguiente sección se discute el proceso de evaluación cuantitativa de la eficacia de la arquitectura híbrida SRI propuesta en esta sección.

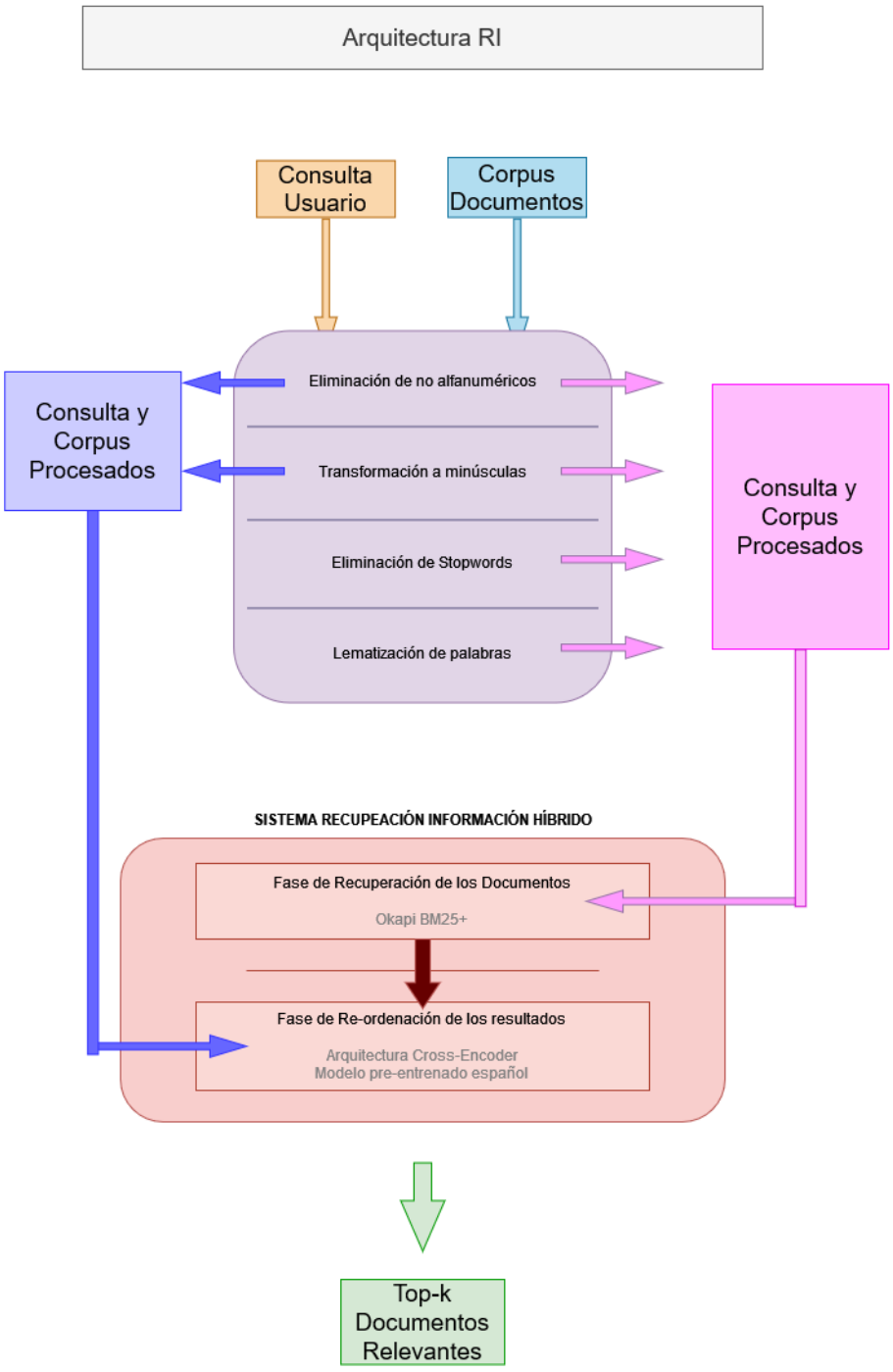

Figura 11: *Diseño final de la arquitectura propuesta*

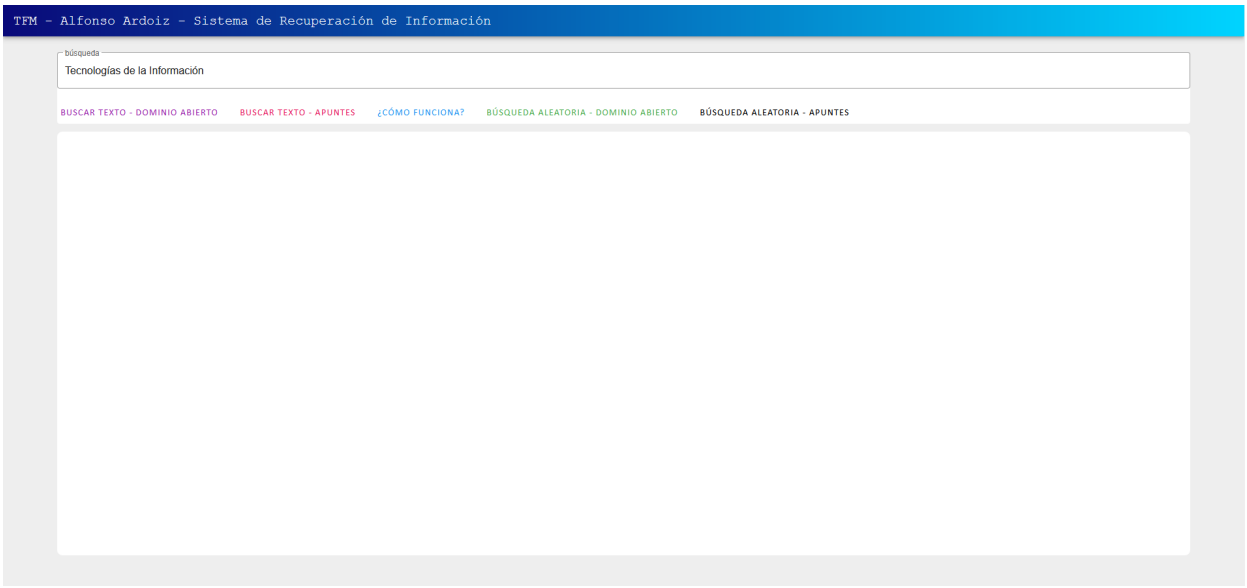

Figura 12: *Interfaz de usuario de la aplicación*

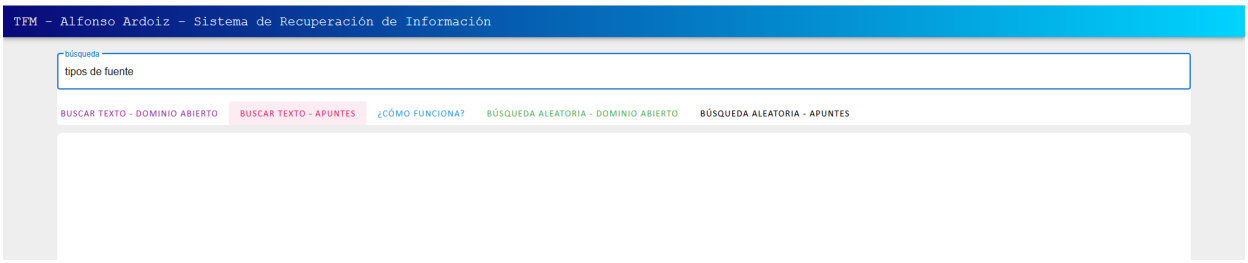

Figura 13: *Ejemplo de uso de la aplicación*

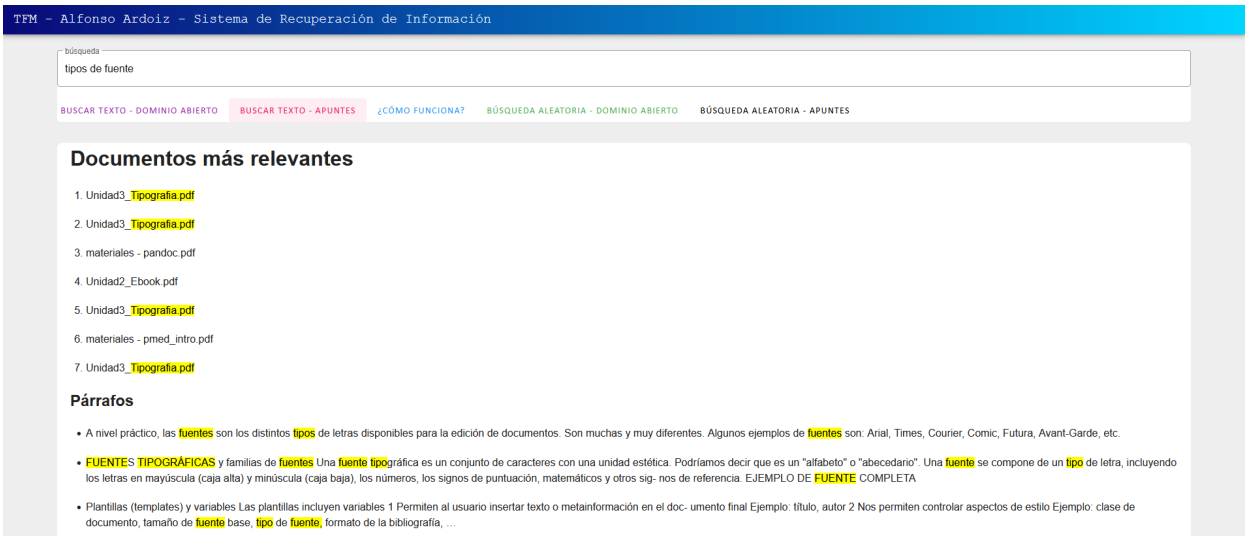

Figura 14: *Resultados final de la aplicación*

Sección 5

Evaluación de eficacia de la nueva arquitectura de Sistemas de Recuperación de Información

Esta sección presenta la evaluación cuantitativa de la eficacia del SRI híbrido para el español. Además, incluye la segunda aportación novedosa y objetivo (ii) de este Trabajo de Fin de Máster: la creación de un prototipo de nuevo conjunto de datos para dicha evaluación.

5.1. La necesidad de un nuevo conjunto de datos

En la Recuperación de Información se han utilizado distintos métodos y conjuntos de datos para la evaluación de los sistemas de búsqueda (desde los primeros proyectos Cranfield en 1953, pasando por las conferencias TREC, y llegando a los conjuntos de datos actuales como el NQ). Estos conjuntos de datos, suelen estar preparados para el inglés, y no es habitual que sean traducidos a otras lenguas. Sin embargo, el éxito del conjunto de datos "MS MARCO" (formado por un corpus con más de un millón de pares búsqueda-documento reales, extraídos de diversos motores de búsqueda en línea como Bing, y utilizado por la mayoría de los SRI actuales como base para la evaluación de su rendimiento) ha dado cabida, en el año 2021, al desarrollo de una versión multilingüe mediante traducción automática, el conjunto "mMARCO". Su objetivo es disponer de un conjunto de datos de referencia en 13 idiomas (entre ellos el español) de modo que haga posible la comparación directa entre distintos enfoques en idiomas diferentes.

Sin embargo, antes de utilizar este conjunto de datos como el nuevo estándar de comparación en español es importante tener en cuenta dos problemas. Por un lado, esta traducción se ha realizado de manera automática, es decir, el conjunto de datos ha adquirido el estado

de "*Silver Standard*"[17] perdiendo de este modo la fiabilidad que poseía el "MS MARCO" como "*Gold Standard*"[18]. En consecuencia, la traducción automática ha generado una pérdida de calidad en los datos reflejada en la anotación incorrecta de pares consulta-documento que, al traducir la búsqueda y los documentos, dejan de ser los más relevantes. Es decir, en muchos casos el documento más relevante para cierta búsqueda ya no coincide con el que tiene enlazado en el conjunto de datos original. Además, el uso de la traducción automática ha provocado errores en el texto en español que se plasman en la pérdida de significado original de la consulta traducida, como en el hecho de que son consultas artificiales, que un usuario real no realizaría. El Anexo C incluye ejemplos relativos a estos errores detectados en un análisis manual del conjunto de datos traducido al español.

Por otro lado, el "MS MARCO" ha sido criticado por ser un conjunto de datos muy restrictivo respecto al número de documentos asociados por consulta. Por ejemplo, para muchas búsquedas, el conjunto de datos solo incluye un único documento relevante, marcando a los demás documentos como no relevantes. Según los experimentos de Qu et al. 2021, el conjunto de datos "MS MARCO" anota como no relevantes hasta un 70 % de documentos sí relevantes por búsqueda.

Este problema se repite en el conjunto de datos "mMARCO" que, al unirse con los problemas que ha generado la traducción automática, acaba resultando en un conjunto de datos que contiene un elevado número de pares búsqueda-documento pero sin suficiente calidad.

En nuestra opinión, estos problemas desaconsejan usar el corpus obtenido mediante traducción automática al español, "mMARCO", como conjunto de datos para evaluar los SRI en español. Como solución, se propone la creación de un nuevo conjunto de datos de categoría "*Gold Standard*" para la evaluación de los SRI en español.

5.2. Creación de un nuevo conjunto de datos de evaluación

Con el objetivo de crear el nuevo conjunto de datos se han explorado distintas opciones, entre ellas, la creación desde cero o la adaptación de otro conjunto de datos disponible para otra finalidad. Teniendo en cuenta el tiempo y los recursos limitados de un TFM y que el objetivo principal de este trabajo es la propuesta de una nueva arquitectura para SRI en español, la primera idea fue desechada a favor de buscar un conjunto de datos existente con

[17] En Machine Learning, se conoce como "*Silver Standard*" a aquellos datos no curados, es decir, que no han sido anotados manualmente. Se incluyen datos extraídos sin supervisión humana, o traducidos desde un conjunto previamente anotado.

[18] En contraposición, "*Gold Standard*" denota a aquellos datos que han sido anotados manualmente y que son de calidad.

el potencial de ser modificado para evaluar tareas de RI.

Esta búsqueda se realiza en base a cuatro requisitos:

- que el conjunto de datos incluya información que pueda corresponderse con los pares búsqueda-documentos
- que el conjunto de datos haya sido anotado originalmente en español
- que el conjunto contenga al menos 1000 pares de datos
- que el conjunto no corresponda a un dominio cerrado

El objetivo de estos requisitos es conseguir un conjunto de datos heterogéneo de calidad que pueda ser utilizado por el mayor número de SRI posibles.

La búsqueda de un conjunto de datos de este tipo concluyó con un resultado que se ajustaba a los cuatro criterios presentados. Es el caso del conjunto de datos para evaluar los sistemas pregunta-respuesta "SQAC", anotado manualmente y desarrollado por el *Barcelona Supercomputing Center* en el año 2022 (Gutiérrez-Fandiño et al. 2022).

Este conjunto de datos incluye 6247 documentos y 18817 preguntas enlazadas mediante la relación contexto-pregunta, adicionalmente incluye respuestas a todas esas preguntas y el título de los documentos que forman el contexto. Al ser un conjunto de datos pensado para sistemas pregunta-respuesta, la relevancia entre pares está implícita, y al estar anotado manualmente, cumple con el criterio de ser un conjunto de tipología "*Gold Standard*".

Para adaptar a este conjunto de datos a la tarea de Recuperación de Información se ha seguido la siguiente metodología:

- Descargar el conjunto de datos SQAC.

- Convertir cada elemento en una tripleta documento(título)- documento(contenido) - consulta(pregunta).

- Eliminar la repetición de documentos y asignar un nuevo ID a cada tripleta. Este paso ha reducido significativamente el tamaño del conjunto de datos.

El resultado final es un prototipo que se ha bautizado como "RISQAC". Es un conjunto de datos formado por 3823 documentos de diversos dominios, originalmente pensado para el español y anotado manualmente. La consecuencia negativa de esta metodología es que no ha sido posible asignar varios documentos a cada búsqueda (repitiendo uno de los fallos criticados del "mMARCO"), lo cual se plantea como una línea de trabajo futuro.

En el Anexo D se encuentra una pequeña muestra del nuevo conjunto de datos.

5.3. Evaluación de la eficacia de la nueva arquitectura

La construcción del nuevo conjunto de datos de evaluación ha posibilitado la evaluación de la capacidad de recuperación y ordenación del "Recuperador de Apuntes".

Este proceso de evaluación se ha dividido en dos pasos independientes que evalúan tanto la velocidad como la eficiencia de cada una de los dos fases del sistema de Recuperación de Información híbrido:

- Por un lado, se calcula la medida de **Cobertura** o Recall para **cuantificar la eficiencia de recuperación de documentos** del módulo que implementa el algoritmo BM25 en español. Para ello se comprueba si el sistema ha seleccionado correctamente el documento esperado dentro de los *top-k* resultados devueltos para cada pregunta. El resultado ha sido que la Cobertura de nuestra arquitectura en el conjunto "RISQAC" dentro de los primeros 20 documentos es un: **0.792** sobre 1; es decir, el sistema ha filtrado correctamente el documento más relevante del casi 80 % de consultas. Adicionalmente, también se ha calculado las medidas de cobertura dentro de los 10 y 3 primeros documentos. Nuestros experimentos muestran que para un corpus de tamaño de 3823 documentos el **tiempo de inferencia** de esta parte corresponde a **2 a 3 milisegundos**.

- En segundo lugar, se calcula la medida de **MRR** o Mean Reciprocal Rank para **valorar el resultado final de la fase de reordenación** de los documentos. Usar esta medida permite valorar objetivamente la eficiencia de la arquitectura, ya que estudia en qué posición se encuentra el documento más relevante de cada consulta. La medida de MRR obtenida en el conjunto "RISQAC" es **0.697** sobre 1 cuando se reordenan los 20 documentos devueltos por la primera fase. Los experimentos adicionales muestran que para una reordenación de 7 documentos, el tiempo de inferencia de esta arquitectura en un entorno sin GPU es de entre **5 a 6 segundos**.

5.3.1. Discusión de resultados

Para poder discutir estos resultados y comprobar la hipótesis del trabajo hace falta conocer las medidas actuales de los sistemas multilingües usados en español, los modelos "mT5" y "mMiniLM" (Bonifacio et al. 2021). Sin embargo, a pesar de que la comunidad investigadora encargada del desarrollo de los modelos que se usan como núcleo de estos sistemas ha abierto su acceso, permitiendo el uso a particulares[19], estos modelos por sí solos no están preparados para la evaluación de nuevos corpus.

Ambos modelos necesitan una arquitectura propia, que por culpa de problemas técnicos[20]

[19]Disponibles en https://github.com/unicamp-dl/mMARCO

[20]A fecha de la finalización de este trabajo, las librerías de Python necesarias tienen errores que siguen sin corregirse, e.j. https://github.com/castorini/pygaggle/issues/267

Nombre	Recall@1000	MRR@10
mT5	0.740	0.297
mMiniLM	0.740	0.292

Tabla 3: *Métricas relativas a la eficiencia de estos modelos desplegados en un sistema híbrido en base al conjunto de datos "mMARCO".*

Nombre	Recall@20	Recall@10	Recall@3	MRR@7	Tiempo Inferencia
"Pyserini"	0.791	**0.752**	0.654	0.599	**2.5 ms**
Nuestra arquitectura	**0.792**	0.751	**0.663**	**0.697**	6000 ms

Tabla 4: *Comparación de medidas de los SRI más usados en español con nuestra arquitectura. Datos obtenidos de nuestros experimentos.*

ha sido imposible de utilizar. Por desgracia, estos modelos no pueden ser aplicados en una red *cross-encoder*, ya que son incompatibles con este tipo de arquitectura, y por lo tanto no pueden ser usados en la arquitectura presentada en este trabajo.

Como solución, se propone un cambio de estrategia, y se pretende comparar la eficiencia de la arquitectura propuesta con las métricas de rendimiento de un SRI disperso de una de las arquitecturas más utilizadas a nivel industrial en los últimos años: "Pyserini"[21] (Lin et al. 2021).

De manera complementaria, para dar una pequeña aproximación al posible resultado que tendrían los modelos multilingües en el nuevo conjunto de evaluación se ha incluido una tabla (tabla 3) con los resultados obtenidos en el conjunto "mMARCO".

Para la comparación con "Pyserini", se ha adaptado "RISQAC" al formato necesario para utilizarlo en su mejor sistema RI para el español (basada en sistemas de representación dispersos utilizando BM25). La tabla 4 representa una comparación de las métricas de eficiencia de ambos sistemas. Observar esta tabla permite analizar las diferencias en ambos sistemas:

1. Por un lado, la puntuación de cobertura de nuestra arquitectura supera en dos de los tres casos analizados a la obtenida por "Pyserini". Si bien es cierto que la diferencia no es muy significativa, el mero hecho de haber superado la eficiencia de una arquitectura consagrada para la RI es un resultado satisfactorio respecto a la posible confirmación de la hipótesis del trabajo: *Mejorar la eficacia de los actuales SRI multilingües del español.*

[21]"Pyserini" es una de las librerías especializadas en RI más utilizadas a nivel industrial. El artículo donde se presenta esta librería ha sido citado más de 100 veces en los dos últimos años.

2. Por otro lado, se observa una diferencia notable en el MRR. El valor obtenido por nuestra arquitectura es casi una décima superior al obtenido por "Pyserini", lo que favorece considerablemente a nuestra arquitectura. Sin embargo, en el futuro es necesario realizar una comparación directa con los sistemas multilingües.

3. Finalmente, ha sido imposible comparar los tiempos de inferencia de los sistemas multilingües frente a nuestro sistema híbrido. Respecto a la velocidad de nuestro SRI híbrido, sabemos que la carga computacional de nuestra arquitectura depende por completo del SRI denso. Esta diferencia entre los sistemas evidencia esa pérdida de rapidez pero incremento de precisión que se ha mencionado a lo largo del trabajo.

Nuestra arquitectura ha alcanzado **resultados excepcionales en las medidas de Cobertura y MRR**. A pesar de no haber podido comparar estas métricas con las obtenidas por los sistemas multilingües, se ha desvelado que es posible mejorar los resultados de estos modelos, y que merece la pena continuar con esta línea de investigación en el futuro.

Sección 6

Resumen, conclusiones y trabajo futuro

6.1. Resumen y conclusiones

En este Trabajo de Fin de Máster se han presentado dos contribuciones al campo de la Recuperación de Información en español. Concretamente:

(1) Se ha explorado, de forma novedosa, la posibilidad de diseñar una arquitectura para SRI que reproduzca las ideas de los sistemas híbridos más eficaces del inglés. Estas arquitecturas funcionan mediante la combinación de algoritmos de cálculo de similitud de representaciones semánticas vectoriales dispersas, con redes neuronales encargadas de la reordenación de los documentos relevantes de forma más precisa pero, también, con mayor coste computacional.

(2) Se ha presentado el prototipo de un nuevo conjunto de datos para la evaluación de los SRI en español, llamado "RISQAC", que pretende corregir los errores existentes en el conjunto de datos "mMARCO" provocados por su traducción automática desde el inglés.

Este trabajo se basa en el mejor sistema del ranking BEIR, el sistema híbrido "BM25+CE" para el diseño de la nueva arquitectura RI. Sin embargo, aún partiendo de la misma idea:

Sistema	BM25+CE	Nuestra arquitectura
Fase de Recuperación	Okapi BM25	Variación Okapi BM25
Parámetros k1, b, δ	0.9, 0.4, 0	1, 0.75, 0.25
Fase de Reordenación	Basada en *Cross-Encoder*	Basada en *Cross-Encoder*
Modelo pre-entrenado	MiniLM	Adaptación al español de MiniLM
Módulos extra	No	Preprocesamiento

Tabla 5: *Resumen comparativo entre el sistema híbrido RI para el inglés "BM25+CE" y la arquitectura propuesta en este trabajo.*

usar un sistema disperso para recuperar documentos y usar un sistema denso para reordenarlos, se ha adaptado este diseño al español. Además, en el diseño de la nueva arquitectura se ha incluido un módulo de preprocesamiento para la normalización y simplificación del texto. La tabla 4 resume los cambios realizados y las diferencias entre ambas propuestas.

La evaluación de la viabilidad y la eficacia de la nueva arquitectura apunta a que la hipótesis inicial era correcta. Asimismo, a lo largo del trabajo se han logrado todos los objetivos específicos propuestos. Para finalizar, se puede enumerar las siguientes contribuciones:

1. Propuesta de una nueva arquitectura para sistemas híbridos de Recuperación de Información en español
2. Creación de una aplicación informática que utiliza esta arquitectura para recuperar documentos de un corpus de apuntes
3. Propuesta de un nuevo conjunto de datos propio del español en el campo de la Recuperación de Información
4. Evaluación de la arquitectura propuesta mediante el nuevo conjunto de datos y comparación de los resultados obtenidos con los modelos predominantes en nuestra lengua

Este trabajo se puede considerar como uno de los primeros pasos en la adaptación de los avances de los últimos años en sistemas de Recuperación de Información al español, y supone una puerta abierta para nuevas investigaciones que puedan mejorar los resultados obtenidos.

6.2. Líneas de trabajo futuras

A pesar de los resultados satisfactorios obtenidos en la evaluación de la arquitectura, a lo largo del trabajo se han visto las limitaciones que, por desgracia, los sistemas actuales en los sistemas de Recuperación de Información aún no pueden superar.

En concreto, no ha sido posible superar las limitaciones ocasionadas por el "*lexical gap*" de los sistemas basados en representaciones dispersas, ni ha sido posible explotar por completo el potencial de los sistemas de representaciones densas en la nueva arquitectura debido a la falta de recursos computacionales.

Adicionalmente, al presentar un nuevo conjunto de datos de referencia en español para la evaluación de SRI, no ha sido posible establecer una comparación objetiva entre nuestra nueva arquitectura y los sistemas multilingüe usados en español. Estos sistemas no han podido ser evaluados sobre el nuevo conjunto de datos, y no ha sido posible obtener un sistema informático con capacidad suficiente como para poder probar el funcionamiento de la nueva arquitectura sobre un corpus de más de un millón de documentos.

Para superar estas limitaciones, se proponen las siguientes líneas de trabajo e investigación futuras:

1. Un estudio detallado sobre las capacidades reales de los embeddings oracionales y su aplicación en la Recuperación de Información.

2. La ampliación del conjunto de datos RISQAC, y la creación de nuevos conjuntos de datos para otro tipo de evaluaciones de SRI en español; ej. un conjunto de datos con errores ortográficos.

3. La evaluación comparativa de nuestra arquitectura y de los SRI multilingües mediante el uso del mismo conjunto de datos de evaluación.

Bibliografía

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., and Wiebe, J. (2015). SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Azad, H. K. and Deepak, A. (2019). Query expansion techniques for information retrieval: A survey. *Information Processing & Management*, 56(5):1698–1735. arXiv:1708.00247 [cs].

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., and Wang, T. (2018). MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv:1611.09268 [cs]*. arXiv: 1611.09268.

Berger, A., Caruana, R., Cohn, D., Freitag, D., and Mittal, V. (2000). Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 192–199, New York, NY, USA. Association for Computing Machinery.

Blair, D. C. (1990). *Language and representation in information retrieval*. Elsevier Science Publishers ; Distributors for the U.S. and Canada, Elsevier Science Pub. Co., Amsterdam, New York. OCLC: 20670761.

Bonifacio, L. H., Campiotti, I., Jeronymo, V., Lotufo, R., and Nogueira, R. (2021). mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset. *arXiv:2108.13897 [cs]*. arXiv: 2108.13897.

Gao, L., Dai, Z., and Callan, J. (2021). COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3030–3042, Online. Association for Computational Linguistics.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Gonzalez-Agirre, A., Armentano-Oller, C., Rodriguez-Penagos, C., and Villegas, M. (2022). MarIA: Spanish Language Models. arXiv:2107.07253 [cs].

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., and Klakow, D. (2021). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. arXiv:2010.12309 [cs].

Jurafsky, D. and Martin, J. H. (2018). Speech and language processing (draft). *preparation [cited 2020 June 1] Available from: https://web. stanford. edu/ jurafsky/slp3*.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Khattab, O. and Zaharia, M. (2020). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48, Virtual Event China. ACM.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J., Lee, K., Toutanova, K. N., Jones, L., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q., and Petrov, S. (2019). Natural Questions: a Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics*.

Lin, J., Ma, X., Lin, S.-C., Yang, J.-H., Pradeep, R., and Nogueira, R. (2021). Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 2356–2362, New York, NY, USA. Association for Computing Machinery.

Luan, Y., Eisenstein, J., Toutanova, K., and Collins, M. (2021). Sparse, Dense, and Attentional Representations for Text Retrieval. arXiv:2005.00181 [cs].

Ma, J., Korotkov, I., Yang, Y., Hall, K., and McDonald, R. (2021). Zero-shot Neural Passage Retrieval via Domain-targeted Synthetic Question Generation. arXiv:2004.14503 [cs].

Martínez Méndez, F. J. (2004). *Recuperación de información: modelos, sistemas y evaluación*. Murcia : Kiosko, 2004. Accepted: 2011-08-21T19:58:51Z Publication Title: http://digitum.um.es/xmlui/handle/10201/4316.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. arXiv:1310.4546 [cs, stat].

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W. X., Dong, D., Wu, H., and Wang, H. (2021). RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. arXiv:2010.08191 [cs].

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv:1908.10084 [cs]*. arXiv: 1908.10084.

Robertson, S. (1977). The Probability Ranking Principle in IR. *Journal of Documentation*, 33:294–304.

Robertson, S. and Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Rodríguez, I. A. (2022). *El modelo cortical HTM y su aplicación al conocimiento lingüístico*. http://purl.org/dc/dcmitype/Text, Universidad Complutense de Madrid. Pages: 1.

Saoud, Z. and Kechid, S. (2016). Integrating social profile to improve the source selection and the result merging process in distributed information retrieval. *Inf. Sci.*

Schuster, M. and Nakajima, K. (2012). Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. ISSN: 2379-190X.

Sparck Jones, K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28(1):11–21. Publisher: MCB UP Ltd.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (2021). BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *arXiv:2104.08663 [cs]*. arXiv: 2104.08663.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. arXiv:1706.03762 [cs].

Voorhees, E. M. and Tice, D. M. (2000). The TREC-8 Question Answering Track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Vulić, I., Baker, S., Ponti, E. M., Petti, U., Leviant, I., Wing, K., Majewska, O., Bar, E., Malone, M., Poibeau, T., Reichart, R., and Korhonen, A. (2021). Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity. *Computational Linguistics*, 46(4):847–897.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. arXiv:2002.10957 [cs].

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. arXiv:2010.11934 [cs].

Zhou, D., Wu, X., Zhao, W., Lawless, S., and Liu, J. (2017). Query Expansion with Enriched User Profiles for Personalized Search Utilizing Folksonomy Data. *IEEE Transactions on Knowledge and Data Engineering*.

Anexo A

Estudio y comparación del impacto de las técnicas de preprocesado

Este anexo se corresponde con el estudio del impacto de la extracción de cada una de las técnicas concretas de procesamiento textual que se han implementado dentro del módulo de preprocesado de nuestra arquitectura.

La finalidad de este estudio es comprender y medir la importancia que estas técnicas han tenido en el funcionamiento del SRI. Complementariamente, se pretende evaluar la eficiencia del módulo de transformación a lemas de las librerías Spacy y NLTK.

De cara al primer objetivo, se ha creído conveniente focalizar el estudio de estas técnicas en la fase de recuperación de documentos del SRI híbrido. La motivación de esta decisión responde al mero hecho del número de técnicas que se usan en esta fase (4) frente al número de técnicas que se usan en la fase de ordenación (2).

La metodología de esta comparación es la siguiente:

1. Establecer las métricas obtenidas por el sistema como base para la comparación
2. Adaptación del script de evaluación usado en el SRI a estos experimentos
3. Cálculo de las métricas sin usar ninguna técnica de procesamiento
4. Cálculo de las métricas con la adición de las técnicas por separado
5. Cálculo de las métricas con la elisión de las técnicas por separado
6. Comparación de todos los resultados

A.1. Resultados del experimento

Para la fase de recuperación de documentos, el diseño original de la arquitectura propuesta en este trabajo aplica las cuatro técnicas de procesamiento mencionadas en la sección 4.1.2 para transformar el texto antes de usar el algoritmo matemático Okapi BM25. En los experimentos que se van a realizar, se quiere cuantificar el impacto de cada una de las cuatro técnicas por separado en el resultado final y ordenarlas en función de su relevancia.

La medida con la que parte este experimento es una cobertura base de **0.792** usando las cuatro técnicas (usando el lematizador de Spacy). Aprovechando este entorno de pruebas se quiere comparar la eficiencia de esta librería frente a la librería NLTK. Para ello, se ha sustituido el lematizador de Spacy por el componente correspondiente de NLTK y se han aplicado conjuntamente el resto de técnicas. Finalmente, la cobertura obtenida baja hasta un puntaje de 0.744, demostrando la necesidad de usar Spacy en las aplicaciones de PLN moderno.

Tras este pequeño hiato en el experimento del impacto de los componentes, se procede a realizar el paso 3 de la metodología descrita, el cálculo de las métricas del sistema al **eliminar todas las técnicas de preprocesado**. En este caso, la cobertura se reduce hasta una puntuación de **0.712**.

El módulo de procesamiento diseñado consigue mejorar la métrica de cobertura del SRI más de un 10 % respecto al SRI base. Este hecho recalca la importancia de la implementación de técnicas de preprocesamiento en cualquier aplicación que utilice texto plano como datos.

El paso 4 de la metodología implica calcular la cobertura del modelo cuando se aplica únicamente una de las técnicas. El primer experimento de este paso pretende aplicar únicamente la técnica de eliminación de caracteres no alfanuméricos. En este caso la cobertura aumenta a 0.724.

Por otro lado, si únicamente se transforma el texto a minúscula se obtiene un resultado muy similar al base, con una cobertura de 0.714.

Al usar la elisión de stops-words como única técnica el resultado se mantiene en 0.718. Finalmente, si analizamos el peso que tiene la conversión de cada palabra a su lema, obtenemos una cobertura de 0.780 con la librería Spacy y un 0.730 con la librería NLTK. Estos resultados posicionan a esta técnica como la más relevante del preprocesamiento textual.

La figura 15 representa el resultado final de cada uno de estos experimentos y permite evaluar el impacto de cada técnica por separado.

El siguiente paso realiza el ejercicio opuesto, medir el comportamiento del sistema si se

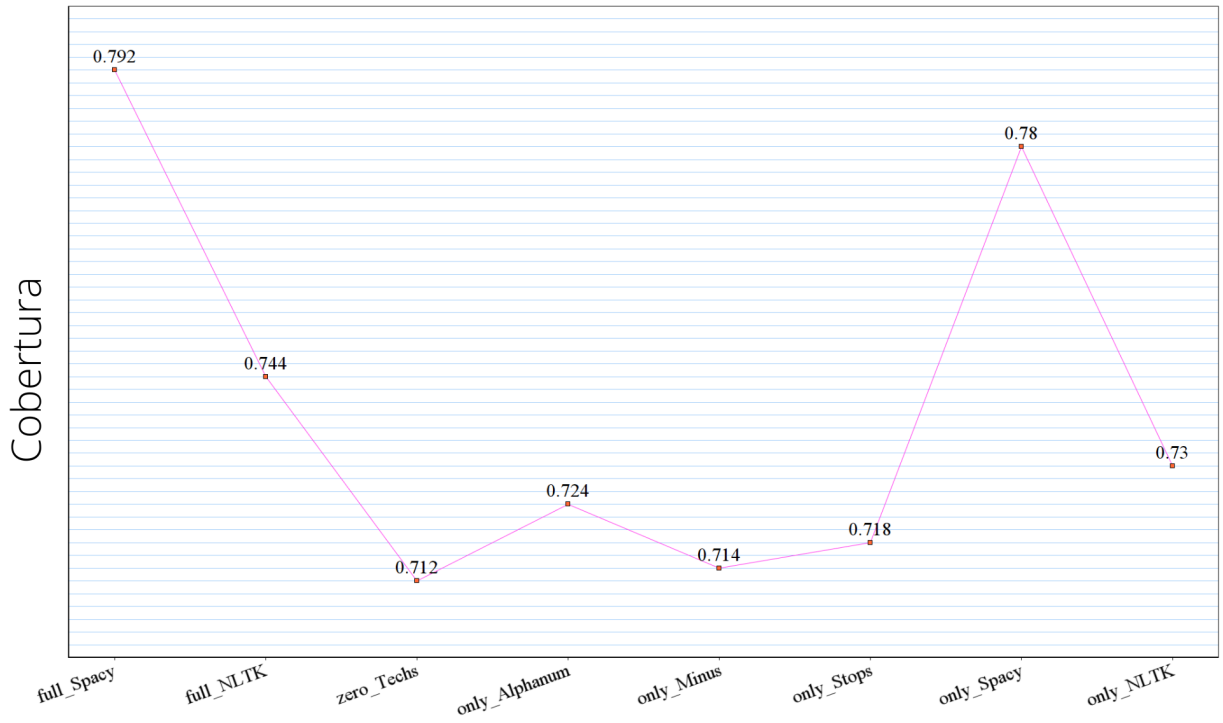

Figura 15: *Resultados del uso de una única técnica de preprocesamiento.*

extirpa una de las técnicas pero se mantiene el resto. En líneas generales este paso repite los experimentos del paso anterior.

El primer experimento consiste en la no eliminación de los caracteres alfanuméricos utilizando las otras tres técnicas. La cobertura resultante baja hasta un 0.748.

Si por otro lado, permitimos que el texto mantenga sus mayúsculas, la métrica de cobertura alcanza un **0.791**, lo cual casi iguala el resultado base. Desvelando que esta técnica apenas tiene repercusión sobre el sistema de recuperación de documentos.

Si no se eliminan las stopwords en los textos mientras se usa el resto de técnicas, el resultado es un 0.786.

Finalmente, si no se convierten las palabras del texto a su lema, la cobertura baja drásticamente hasta un valor de 0.736, confirmando la idea de que la lematización es la técnica que más influye en la eficiencia del SRI.

En este caso, los resultados finales se pueden apreciar en la figura 16.

Finalmente, el estudio concluye con el análisis de la comparación de la relevancia de las técnicas. Dado los resultados obtenidos en los experimentos, el orden las técnicas más

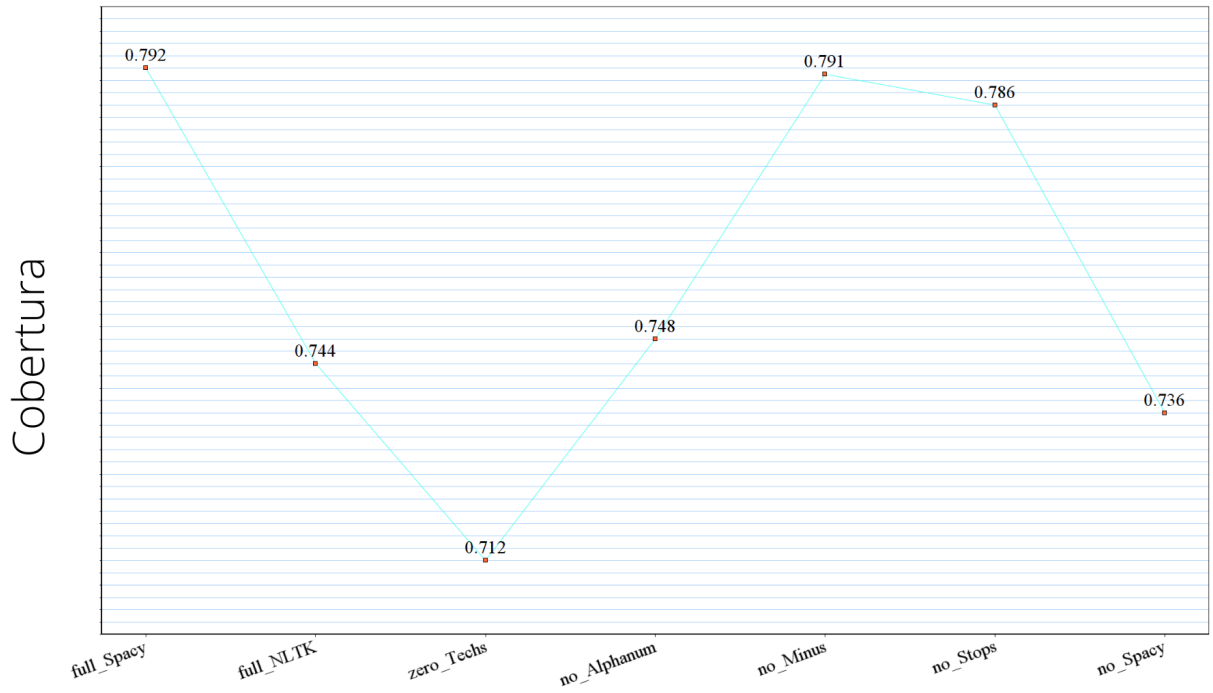

Figura 16: *Resultados de la combinación de tres de las técnicas de preprocesamiento.*

importantes del módulo de preprocesamientos es el siguiente:

(i) Lematización de cada palabra del texto

(ii) Eliminación de los caracteres no alfanuméricos

(iii) Elisión de las stop-words

(iv) Conversión a minúsculas del texto

Complementariamente, se han comparado dos de las librerías de PLN más usadas en español. Cuyo resultado ha demostrado la eficiencia de la librería Spacy como base para los procesos de lematización frente a los pobres resultados obtenidos por el uso de la librería NLTK.

Anexo B

Evaluación de los modelos pre-entrenados de lenguaje especializados en el cálculo de similitud semántica para el español

Este anexo expone brevemente la metodología de evaluación usada para seleccionar el mejor modelo pre-entrenado de lenguaje para redes de arquitectura cross-encoder en el cálculo de similitud semántica para el español.

En el momento de realizar este proceso, se seleccionaron los seis modelos pre-entrenados del lenguaje para la detección de similitud semántica en español más populares de la plataforma HuggingFace. La tabla 6 sintetiza la información de estos modelos, y les asigna un ID para mayor comodidad.

La metodología para este experimento de evaluación es la siguiente:

1. Descarga de los modelos mediante la librería de python "Transformers"
2. Creación de un script de evaluación para medir similitud semántica entre oraciones
3. Uso del conjunto de datos STS (Agirre et al. 2015) para la evaluación de cada modelo
4. Comparativa de los resultados y selección del modelo final

Como se puede observar, se ha decidido utilizar un conjunto de datos existente para agilizar la evaluación de los modelos. En esta evaluación, se ha calculado el índice de similitud entre dos oraciones, y se ha comparado el resultado obtenido con el presente en el conjunto de similitud.

Modelo	**ID**
distiluse-base-multilingual-cased-v1	DBM1
paraphrase-multilingual-MiniLM-L12-v2	PA1
paraphrase-multilingual-mpnet-base-v2	PA2
distiluse-base-multilingual-cased-v2	DBM2
mmarco-mMiniLMv2-L12-H384-v1	MLM
sentence-similarity-spanish-es	HIA

Tabla 6: *Modelos evaluados*

La figura 17 muestra la media (redondeada a la unidad) de los resultados de todas las iteraciones realizadas (140 pares de oraciones comprobados en subgrupos de 14 pares). En concreto, se muestran los resultados erróneos en rojo (más de un punto de discrepancia entre la similitud esperada y la resultante), los resultados aceptables en amarillo (un punto de discrepancia) y los resultados acertados en verde (menos de 0.5 puntos de discrepancia).

Finalmente, el modelo pre-entrenado seleccionado es el MLM, este modelo apenas ha obtenido fallos en el cálculo de similitud y es el que más resultados correctos ha devuelto. Durante el entrenamiento de este modelo pre-entrenado, se realizó un proceso de fine-tuning para preparar al modelo como el núcleo de una red *cross-encoder* para español, por lo que en nuestra opinión, se adapta perfectamente a las necesidades del nuevo SRI.

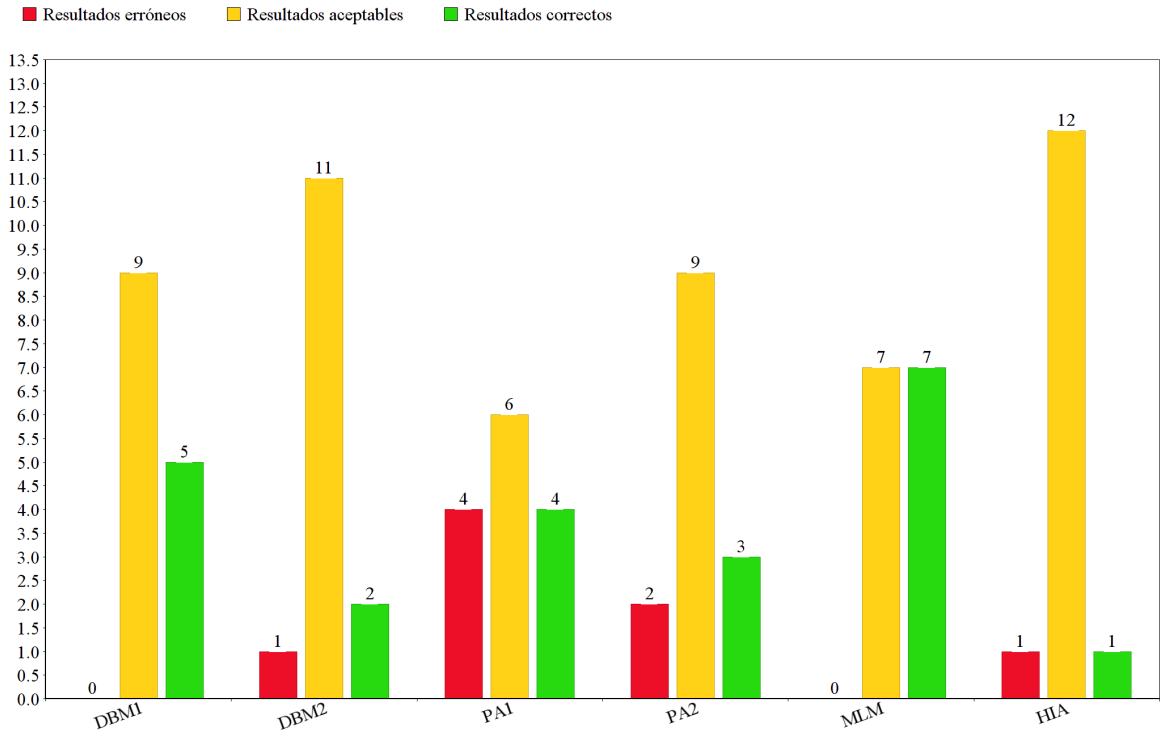

Figura 17: *Resultado de la evaluación de los modelos.*

Anexo C

Compendio de datos erróneos encontrados en mMARCO

En esta sección se describe el proceso que se ha seguido para descartar el uso del conjunto de datos "mMARCO" como estándar para la evaluación de SRI en español.

En primer lugar, se ha realizado una exploración manual previa del corpus. En esta exploración se han encontrado consultas particularmente preocupantes para su uso como estándar de comparación en español.

De cara a cuantificar el número de fallos en estas consultas se ha decidido seleccionar al azar conjuntos de 10 consultas durante 20 iteraciones. La hipótesis es que si el conjunto de datos es de calidad, los fallos vistos en la exploración inicial han tenido que ser un caso aislado y no representan el estado global del conjunto. Si se siguen encontrando fallos de manera generalizada en el experimento, se puede inferir que la calidad de este corpus se ha reducido debido a la traducción automática.

Con cada iteración se va a comprobar la calidad de la traducción de las consultas (entendiendo por consulta de calidad a aquella búsqueda que un usuario real puede hacer en nuestro idioma).

Para ello se ha desarrollado un programa de Python que descarga automáticamente 10 consultas en ambos idiomas y las guarda en una base de datos. Este proceso será repetido 20 veces, y una vez terminado se procederá a evaluar manualmente las consultas.

El análisis manual de las consultas revela que de 200 consultas analizadas, 20 contenían errores de traducción y no reflejan consultas reales de nuestro idioma. Para ejemplificar esta afirmación, a continuación se presenta un compendio de las seis consultas más extrañas:

'id': 391836, 'text': 'si t stat es menor que t crit igual varianza'
'id': 797201, 'text': '¿Por qué es causado el smog?'

'id': 738368, 'text': 'que es delta in2ition'

'id': 203218, 'text': 'dirección de highmark'

'id': 473204, 'text': 'costo por pies cuadrados en california para la construcción del inquilino'

'id': 263889, 'text': 'cuanto tiempo es si tu discapacidad'

Este análisis desvela que un 10 % de las consultas del pequeño muestreo que se ha estudiado contienen errores y no deberían usarse en un estándar de evaluación de SRI ya que (i) el valor que aporta la evaluación de una consulta errónea es nulo, (ii) muchas de estas consultas tienen asociados documentos como los más probables con los que se ya no se guarda relación.

Por ejemplo el documento asociado a la consulta "¿Por qué es causado el smog?" ha traducido correctamente "smog" por "niebla tóxica" perdiendo la palabra clave principal de la consulta del documento, inhabilitando que los SRI dispersos o SRI densos monolingües puedan devolver este documento correctamente en la evaluación de esta consulta.

Anexo D

Ejemplo del corpus “RISQAC”

En este Anexo se incluye, a modo de ilustración, una pequeña muestra del nuevo conjunto de datos de evaluación RISQAC".

Se han recopilado 5 tripletas título - documento - contenido que en nuestra opinión, ejemplifican a la perfección la esencia de este conjunto de datos:


```
{'title': 'Historia de Japón',
 'content': 'La historia de Japón (日本の歴史 o 日本史, Nihon no rekishi / Nihonshi?) es la sucesión de hechos acontecidos d
entro del archipiélago japonés. Algunos de estos hechos aparecen aislados e influenciados por la naturaleza geográfica de
Japón como nación insular, en tanto que otra serie de hechos, obedece a influencias foráneas como en el caso del Imperio c
hino, el cual definió su idioma, su escritura y, también, su cultura política. Asimismo, otra de las influencias foráneas
fue la de origen occidental, lo que convirtió al país en una nación industrial, ejerciendo con ello una esfera de influenc
ia y una expansión territorial sobre el área del Pacífico. No obstante, dicho expansionismo se detuvo tras la Segunda Guer
ra Mundial y el país se posicionó en un esquema de nación industrial con vínculos a su tradición cultural.',
 'question': '¿Qué influencia convirtió Japón en una nación industrial?'}
```


Figura 18: *Tupla Título-Documento-Consulta ID:0*


```
{'title': 'Las familias de los ocupantes del desaparecido vuelo MH370 planean buscar por su cuenta el avión',
 'content': '4 de marzo de 2017 La búsqueda oficial de este avión que, como recordamos, desapareció en marzo de 2014 en el
Océano Índico, fue cancelada en enero después de que las autoridades no pudieron encontrar los restos principales del avió
n. Sin embargo, las familias de quienes viajaban a bordo del vuelo MH370 de Malaysia Airlines no se resignan a quedarse si
n saber que fue les pasó a sus seres queridos y por ello han lanzado una campaña para recaudar 15 000 000 de euros para pa
gar una búsqueda privada del avión desaparecido. El Boeing 777 cubría una ruta de Kuala Lumpur a Beijing cuando desapareci
ó con 239 personas a bordo. Se cree que se estrelló en una parte remota del sur del Océano Índico, pero la mayor búsqueda
en la historia de la aviación no ha logrado encontrar los principales restos del avión. Esto se debería a que, según se pi
ensa, el avión se habría desecho y que por ello solo se han hallado trozos sin mayor importancia de la aeronave. Los detal
les del plan de las familias para una búsqueda privada fueron anunciados en un memorial justo antes del tercer aniversario
de la desaparición del avión, que es este próximo 8 de marzo. Este memorial se realizó hoy en Kuala Lumpur, en donde el mi
nistro de Transporte de ese país, de Malasia, Liow Tiong Lai, aseguró que "el Gobierno de Malasia está y siempre estará de
vuestra parte" en un discurso tras guardar un minuto de silencio. Jiang Hui, de quien su madre volaba en el avión, descubr
ió un trozo del MH370 en Madagascar en 2016. – Jiang Hui, hija de una pasajera del MH370 – Jiang Hui – Grace Nathan, aboga
da de los familiares de los ocupantes del vuelo desaparecido, durante una rueda de prensa con motivo del tercer aniversari
o del siniestro Se han presentado varias teorías acerca de lo que le pudo ocurrir a la aeronave, incluyendo un incendio a
bordo, un secuestro o una trama terrorista, una acción deshonesta del piloto o también, fallas mecánicas o estructurales.
Este mismo año va a salir el informe final sobre la desaparición del avión.',
 'question': '¿A qué compañía aérea pertenecía el vuelo MH370?'}
```


Figura 19: *Tupla Título-Documento-Consulta ID:38*

```
{'title': 'Aves',
 'content': 'Ecología \nLas aves ocupan un amplio espectro de nichos ecológicos. Mientras algunas aves son generalistas, o
tras están altamente especializadas en su hábitat o en su alimentación. Incluso en un solo hábitat, como por ejemplo un bo
sque, los nichos ecológicos ocupados por diferentes aves varían; algunas especies se alimentan en la copa de los árboles,
otras por debajo del dosel arbóreo, y algunas en el suelo del bosque. Las aves forestales pueden ser insectívoras, frugívo
ras y nectarívoras. Las aves acuáticas por lo general se alimentan pescando, comiendo plantas acuáticas, o como cleptopará
sitas. Las aves de presa están especializadas en cazar mamíferos, otras aves y otros animales, mientras que los buitres so
n aves carroñeras especializadas.',
 'question': '¿Dónde comen las aves?'}
```


Figura 20: *Tupla Título-Documento-Consulta ID:560*


```
{'title': 'Detienen dirigentes sindicales griegos por protestar contra impuesto a la propiedad',
 'content': '27 de noviembre de 2011 El presidente del Sindicato de Trabajadores Electricistas de Grecia (GENOP), Nikos Fo
topoulos, junto con 14 dirigentes sindicales fueron detenidos por la Policía tras cuatro días de ocupación de las instalac
iones de la Compañia Pública de Electricidad (DEH). Los trabajadores, que se encontraban protestando por el cobro de un im
puesto creado por el Parlamento griego a la propiedad inmobiliaria en las facturas de energía eléctrica a fin de reunir fo
ndos para palear la crisis económica, se enfrentaron a 80 policías antimotines en las entradas de la entidad, motivo por e
l cual Evangelos Venizelos (ministro de finanzas), advirtió a la población que si se rehúsan a pagar, se enfrentarían a la
suspensión del servicio. Al ser detenido Fotopoulos, este declaró su protesta ante los medios en defensa de las personas q
ue se verían desfavorecidas con la medida: En tanto, GENOP anunció una huelga de 48 horas para el martes próximo mientras
las fuerzas políticas y sindicales como la Federación Sindical de Empleados públicos (ADEDY), La Federación de Empleados j
udiciales (ODYE) y la Confederación General de Trabajadores de Grecia (GSEE) deploraron la acción de la policía. El dirige
nte del Partido Comunista helénico, Spyros Halvatzis, señala que "una vez más se ha probado que es necesario usar la polic
ía para imponer políticas antipopulares" invitando a los trabajadores a un paro que se realizará el 1 de diciembre. Por su
parte, el dirigente de la Coalición de la Izquierda Radical, Alekos Alavanos, criticó al gobierno de Lucas Papademos al ar
gumentar que "tras la sonrisa del primer ministro tecnócrata se esconden los colmillos de la banca" y concluye que las luc
has por venir terminarán en una victoria popular. En tanto el primer mandatario griego está trabajando a contrarreloj para
implementar medidas austeras ordenadas por la Unión Europea. Se estima que al comenzar el 2012, las tarifas de energía se
incrementarán entre un 10% y un 13%, por lo cual los clientes de estratos medio y bajo se vean afectados, razón por el cua
l las protestas en Grecia continuarán.',
 'question': '¿Por qué protestaban los sindicalistas griegos arrestados?'}
```


Figura 21: *Tupla Título-Documento-Consulta ID:707*


```
{'title': 'Gregorio Nacianceno',
 'content': 'Legado \nLas contribuciones teológicas más significativas de Gregorio surgen de su defensa de la doctrina nic
ena de la Trinidad. Destaca especialmente por sus contribuciones en el campo de la pneumatología, esto es, la teología ref
erente a la naturaleza del Espíritu Santo.[nota 1] A este respecto, Gregorio es el primero que usó la idea de procesión.[n
ota 2] para describir la relación entre el Espíritu y las demás personas de la Trinidad: «El Espíritu Santo es verdaderame
nte Espíritu, viniendo en verdad del Padre pero no de la misma manera que el Hijo, pues no es por generación sino por proc
esión, puesto que debo acuñar una palabra en beneficio de la claridad».[nota 3] Aunque Gregorio no desarrolla plenamente e
l concepto, la idea de procesión permanecería en la mayor parte del pensamiento posterior sobre el Espíritu Santo.[nota
4]',
 'question': '¿Cómo se define la pneumatología?'}
```


Figura 22: *Tupla Título-Documento-Consulta ID:1872*