Sistema RAG para Verificación de Hechos

Procesamiento del Lenguaje Natural

Rubén Cid Costa
Guillermo Grande Santi
Jaime Ruiz Salcedo
Jesús Martín Trilla

Repositorio de GitHub: https://github.com/BiggestGuille/Fact-Verification-System-LLM

Descripción de las decisiones de diseño, tecnologías y funcionalidades.

Se ha usado el conjunto de datos *Climate-Fever*, el cual contiene afirmaciones sobre el cambio climático junto a evidencia relevante para cada una. En total, se disponen 1535 afirmaciones y 7675 piezas de evidencia. Estos textos contienen información sobre su validez, el consenso científico y procedencia.

Se ha utilizado **ChromaDB** como base de datos vectorial por su simpleza de uso. Los embeddings se han creado mediante el modelo preentrenado transformer **all-MiniLM-L6-v2** debido a que obtiene un buen rendimiento a pesar de su tamaño. Se han generado dos colecciones, una para las afirmaciones *(claims)* y otra para las evidencias *(evidences)*.

Para la interacción con el Large Language Model (LLM), se ha utilizado la librería **LlamaIndex**. El LLM principal seleccionado es **Llama3.2**, pudiendo ser sustituido por **GPT-4o-mini** en el caso de que no esté disponible, garantizando la disponibilidad del servicio mientras trabajábamos.

Para asegurar una buena recuperación de fuentes, se utiliza **fragmentación de texto** de manera que si una entrada del usuario consta de diferentes hechos a verificar se separe en afirmaciones atómicas y se realice una consulta por cada una de ellas. De esta manera se evita una tosca valoración general de la oración, logrando una verificación más precisa y matizada de cada hecho.

Se ha utilizado ***Few-Shot Prompting*** para guiar al LLM a lo largo de todo el proceso: la descomposición en afirmaciones atómicas, la verificación de las afirmaciones y la consolidación de la respuesta final. Esta última sigue el siguiente patrón:

- Valoración general de la afirmación: Verdadera, falsa, inconclusa (si existe información contradictoria) o sin evidencia (si no hay ninguna fuente al respecto en la base de datos).
- Explicación detallada de por qué las fuentes sugieren dicha valoración. Si la entrada del usuario puede fragmentarse en afirmaciones atómicas se ofrece una valoración propia para cada una. Además, se referencian las fuentes con formato Vancouver.
- Fuentes ordenadas por orden de aparición a lo largo de la respuesta.

Como funcionalidad adicional se ha implementado una **traducción** a varios idiomas. La entrada del usuario siempre se traduce al inglés, ya que es el idioma en el que está escrita la base de datos. La respuesta final del LLM se traduce al idioma original de la entrada, sea cual sea este. Esta traducción se hace mediante el propio modelo principal elegido (Llama3.2 ó GPT-4o-mini) por lo que soporta un amplio rango de idiomas. Otra funcionalidad adicional añadida ha sido incluir un valor de confianza de cuán seguro está el modelo sobre la veracidad de la respuesta, basándonos en una modificación del

Fact Score. Gracias a un modelo de “Natural Language Inference” comparamos las fuentes recuperadas con la afirmación que estamos demostrando, obteniendo un valor de confianza en nuestra respuesta.

Evaluación del sistema

Respecto a la evaluación del sistema, se han tenido en cuenta diferentes aspectos. Por un lado la capacidad del sistema para encontrar evidencia relevante y adecuada para emitir un veredicto se ha evaluado mediante la métrica **Recall@K**, eligiendo un valor de 5 para el parámetro K. Por otro lado se ha utilizado el **BERTScore** para comprobar la similaridad entre el texto generado por el LLM y las evidencias originales. También se ha evaluado la veracidad de los veredictos finales etiquetando 75 consultas a mano y comprobando la **precisión**, sensibilidad (***recall***) y **F1** de los mismos. Por último, se ha medido el **tiempo medio** que tarda el sistema en generar una respuesta.

A continuación se muestran los resultados obtenidos para las métricas empleadas:

Métricas de los veredictos finales				
Class	**Precision**	**Recall**	**F1-Score**	**Instances**
Other (No evidence / Not enough evidence)	0.96	0.75	0.84	32
Refute	0.80	0.95	0.87	21
Support	0.88	1.00	0.94	22
Average	0.88	0.9	0.88	75

Similaridad e implicación entre respuesta y fuentes		
Score	**Métrica**	**Valor**
BERT Score	Precisión Promedio	0.871
	Recall Promedio	0.930
	F1 Promedio	0.898
FACT Score	0.657	

Tiempo de Respuesta	
Tiempo Promedio Sin Traducción	2,77 s
Tiempo Promedio Con Traducción	10,25 s

Conclusiones y limitaciones

El sistema logra unos buenos resultados en cuanto a recuperación de documentos relevantes, generación de respuestas adecuadas a las evidencias y tiempo de respuesta. Cabe destacar que, excepto el tiempo de respuesta, estas mediciones se han realizado sobre consultas exclusivamente en inglés, ya que estas métricas podrían tener valores diferentes para cada idioma. Además, se ha considerado la traducción una función adicional no esencial.

Se observa una menor precisión en la clase "refute", lo cual constituye nuestra principal limitación. Esto se debe al modelo de embeddings utilizado, ya que, aunque existen afirmaciones en la base de datos, no se logran recuperar evidencias al buscar su negación. Asimismo, la clase "other" presenta la sensibilidad más baja debido a la limitación de nuestro conjunto de datos. Al haber utilizado un dataset con todo tipo de frases relacionadas con el cambio climático, el modelo tiende a identificar información inconclusa con mayor frecuencia que a acertar o rechazar la afirmación.