

Comparación de Sistemas de Recuperación de Información: BM25 vs. DPR vs. Reranking

Eisler Francisco Valles Rodriguez, Rafael Acosta Márquez, and Jorge Alejandro Pichardo

Universida de La Habana, La Habana, Cuba

Abstract. Este trabajo presenta una comparación entre tres sistemas de recuperación de información: BM25, un modelo tradicional sin el uso de inteligencia artificial, Dense Passage Retrieval (DPR), un modelo moderno que integra técnicas de inteligencia artificial, y una estrategia de reranking que combina ambos enfoques. Se evaluaron los sistemas en términos de precisión y eficiencia, utilizando un conjunto de datos de prueba. BM25 se destacó por su simplicidad y velocidad, mientras que DPR mostró una mayor precisión en la recuperación de documentos semánticamente relevantes, aunque a un costo computacional más alto. El reranking permitió una mejora significativa al utilizar BM25 como filtro inicial y DPR para refinar los resultados más relevantes. Se concluye con una discusión sobre las limitaciones de estos enfoques y propuestas de mejora futuras.

Keywords: Recuperación de información, BM25, DPR, Reranking, Inteligencia artificial, Similitud semántica

1 Introducción

La recuperación de información es un campo fundamental en la ciencia de la computación, con aplicaciones que van desde motores de búsqueda hasta sistemas de recomendación. Este trabajo compara tres enfoques de recuperación de información: BM25, Dense Passage Retrieval (DPR) y un sistema híbrido de *reranking*.

BM25 es un modelo probabilístico tradicional que estima la relevancia de un documento basándose en la frecuencia de las palabras clave en la consulta y el documento [1]. A pesar de su simplicidad, BM25 es rápido y eficiente, pero tiene limitaciones al no capturar relaciones semánticas profundas entre los términos.

DPR (Dense Passage Retrieval), por otro lado, utiliza representaciones densas generadas por redes neuronales para capturar el significado semántico de las consultas y documentos [2]. Al representar tanto la consulta como los documentos como vectores en un espacio de embeddings, DPR es capaz de recuperar información de manera más precisa, particularmente en casos donde la relevancia semántica juega un papel crucial.

El **reranking** es una estrategia que combina lo mejor de ambos enfoques [3]. En este caso, BM25 se usa como un filtro inicial para reducir el número

de documentos candidatos, y posteriormente, DPR se emplea para realizar un reranking de los documentos filtrados. Esto permite mejorar la precisión sin incurrir en el costo computacional de aplicar DPR a todo el corpus.

2 Antecedentes

BM25 ha sido el estándar en recuperación de información durante varias décadas debido a su balance entre simplicidad y efectividad. Sin embargo, en los últimos años, métodos más avanzados como DPR han demostrado una mejor capacidad para capturar relaciones semánticas complejas. Al utilizar embeddings densos generados por redes neuronales, DPR ofrece mejores resultados en dominios donde la semántica es fundamental. El reranking ha surgido como una técnica que busca optimizar la recuperación de información al combinar BM25 y DPR, permitiendo un sistema más escalable y preciso.

2.1 Conjunto de Datos

Para llevar a cabo la comparación entre los sistemas de recuperación de información, se utilizó el [WikiQuestionAnswer Dataset](#), disponible en Kaggle. Este conjunto de datos está compuesto por pares de preguntas y respuestas de dominio abierto, lo que permite evaluar la efectividad de los diferentes modelos en la recuperación de información.

2.2 Descripción de BM25, DPR y re-ranking con ambos modelos

BM25

BM25 es un modelo de recuperación de información clásico y ampliamente utilizado que pertenece a la familia de los modelos probabilísticos de recuperación. Su objetivo principal es estimar la relevancia de un documento con respecto a una consulta dada, basándose en la frecuencia de términos tanto en la consulta como en el documento.

Cómo funciona BM25:

- **Frecuencia de términos (TF):** BM25 considera cuántas veces aparece un término de la consulta en un documento. Cuanto más frecuente sea el término, mayor será su contribución a la puntuación de relevancia.
- **Frecuencia inversa de documentos (IDF):** Este componente da más peso a los términos que son raros en la colección de documentos. Un término que aparece en muchos documentos es menos discriminativo.
- **Normalización por longitud del documento:** BM25 incluye un factor que penaliza o favorece documentos según su longitud, evitando que documentos más largos obtengan puntuaciones injustamente altas solo por contener más términos.

La fórmula básica de BM25 es:

$$\text{Score}(D, Q) = \sum_{q_i \in Q} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})}$$

Donde:

- $f(q_i, D)$ es la frecuencia del término q_i en el documento D .
- $|D|$ es la longitud del documento D .
- avgdl es la longitud promedio de los documentos en la colección.
- k_1 y b son parámetros de ajuste.

Uso de BM25: BM25 se utiliza comúnmente en motores de búsqueda y sistemas de recuperación de información para ordenar documentos en función de su relevancia respecto a una consulta. Es eficiente computacionalmente y ofrece un buen equilibrio entre precisión y simplicidad.

Dense Passage Retrieval (DPR)

DPR es un enfoque moderno para la recuperación de información basado en modelos de aprendizaje profundo, específicamente en representaciones densas aprendidas mediante redes neuronales. A diferencia de métodos tradicionales como BM25, que utilizan representaciones dispersas basadas en términos, DPR utiliza *embeddings* de alta dimensión para representar consultas y documentos.

Cómo funciona DPR:

- **Representación Densa:** Tanto las consultas como los pasajes de documentos se representan mediante vectores densos en un espacio de alta dimensión.
- **Modelos BERT Independientes:** DPR utiliza dos modelos BERT separados para las consultas y los pasajes, permitiendo capturar características específicas de cada uno.
- **Función de Similitud:** La relevancia entre una consulta y un pasaje se mide mediante la similitud coseno o producto escalar entre sus vectores.
- **Entrenamiento Supervisado:** DPR se entrena utilizando pares de consultas y pasajes relevantes, optimizando para que los pasajes relevantes estén más cerca de la consulta en el espacio vectorial.

Uso de DPR: DPR es efectivo en tareas de búsqueda semántica y recuperación de respuestas en sistemas de Preguntas y Respuestas (QA). Captura relaciones semánticas más allá de la coincidencia exacta de términos, lo que permite recuperar documentos relevantes que no comparten necesariamente palabras con la consulta.

Re-ranking usando BM25 y DPR

El proceso de *re-ranking* combina las fortalezas de ambos modelos para mejorar la precisión de la recuperación de información.

Cómo es el proceso:

1. **Recuperación Inicial con BM25:** Se utilizó BM25 para realizar una recuperación inicial de documentos o pasajes. BM25 es rápido y eficiente, lo que permite filtrar rápidamente un conjunto amplio de documentos y obtener un conjunto candidato de, por ejemplo, los top 100 documentos más relevantes.
2. **Re-ranking con DPR:** Se aplicó DPR al conjunto candidato obtenido de BM25. Se calcularon las representaciones densas de las consultas y los pasajes candidatos y luego se recalcularon las puntuaciones de relevancia basándose en la similitud de los *embeddings*. Finalmente se ordenaron nuevamente los documentos según las nuevas puntuaciones obtenidas con DPR.

Ventajas de este enfoque:

- **Eficiencia:** BM25 filtra rápidamente los documentos menos relevantes, reduciendo la carga computacional para DPR.
- **Precisión:** DPR refina el *ranking* inicial considerando relaciones semánticas profundas, mejorando la relevancia de los resultados finales.
- **Combina lo Mejor de Ambos Mundos:** Aprovecha la rapidez de los métodos tradicionales y la precisión de los modelos neuronales.

3 Consideraciones Implementadas

Al implementar estos sistemas, se tomaron en cuenta las siguientes consideraciones:

- **Complejidad vs. Precisión:** BM25 fue seleccionado por su simplicidad y bajo costo computacional, mientras que DPR fue elegido por su capacidad para capturar relaciones semánticas complejas. El reranking se utilizó para equilibrar ambas características, aplicando BM25 como un primer filtro y DPR para refinar los resultados más relevantes.
- **Escalabilidad:** Se evaluó la escalabilidad de ambos métodos, considerando el tiempo de respuesta y el uso de recursos, especialmente en grandes volúmenes de datos. El reranking permitió mejorar la escalabilidad al reducir el número de documentos que DPR debía procesar.
- **Contexto de la Consulta:** DPR y el reranking se priorizaron en consultas donde la interpretación del contexto era crucial, mientras que BM25 se utilizó en consultas más simples y directas.

4 Evaluación Cuantitativa y Cualitativa

4.1 Evaluación Cuantitativa

A continuación, se presentan los resultados obtenidos de la comparación entre los sistemas de recuperación de información BM25, DPR, y Reranking. Los resultados se expresan en términos de Precisión Media (Average Precision) y Recall Medio (Average Recall).

Table 1. Resultados de Precisión Media

Sistema	Precisión Media
BM25	0.324
DPR	0.531
Reranking	0.4969

Table 2. Resultados de Recall Medio

Sistema	Recall Medio
BM25	0.4053
DPR	0.5853
Reranking	0.4496

Se observa que el modelo DPR supera a BM25 y Reranking en ambas métricas, con una precisión media de 0.531 y un recall medio de 0.5853. Aunque Reranking mejora respecto a BM25 en precisión, su rendimiento es inferior en términos de recall comparado con DPR.

4.2 Evaluación Cualitativa

Cualitativamente, se observó que el reranking permitía una recuperación más precisa, particularmente en consultas complejas donde la semántica del contexto era crucial. BM25, aunque menos preciso, se destacó en consultas simples y fue significativamente más rápido. El reranking balanceó ambos enfoques, mejorando la precisión sin aumentar excesivamente el costo computacional.

5 Declaración Autocrítica y Propuestas de Mejora

A pesar de los resultados obtenidos, se reconocen varias limitaciones en la implementación:

- **Costo Computacional de DPR:** Aunque DPR fue más preciso, su alto costo computacional lo hace menos práctico en entornos con recursos limitados. El reranking ayudó a mitigar este problema, pero se sugiere explorar técnicas de optimización adicionales, como la reducción de la dimensionalidad de los embeddings.
- **Limitaciones de BM25:** BM25 no es capaz de capturar relaciones semánticas complejas, lo que limita su efectividad en ciertas consultas. El reranking ayudó a mejorar este aspecto, pero podría evaluarse la posibilidad de aplicar otros métodos de inteligencia artificial para consultas más complejas.
- **Generalización en Diferentes Dominios:** Los experimentos se realizaron en un conjunto de datos específico. Se propone realizar pruebas en diferentes dominios y con otros tipos de consultas para evaluar la generalización de los resultados.
- **Limitaciones computacionales:** Con una infraestructura más adecuada se podrían haber ejecutado pruebas más complejas utilizando incluso datasets más grandes y complejos

6 Conclusión

Este trabajo comparó tres enfoques de recuperación de información: BM25, DPR y una estrategia híbrida de reranking. Mientras que BM25 mostró una gran eficiencia, DPR destacó por su precisión en consultas complejas. El reranking permitió aprovechar lo mejor de ambos enfoques, mejorando la precisión sin sacrificar significativamente la eficiencia computacional. Un enfoque híbrido como el de reranking parece ser una solución prometedora para futuros sistemas de recuperación de información.

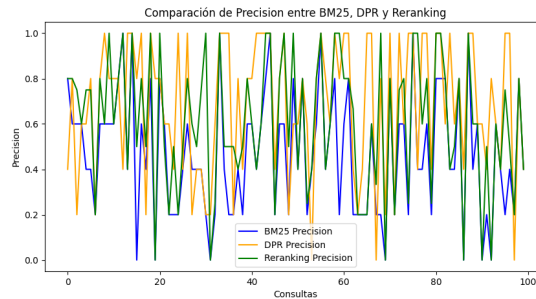


Fig. 1. Uso de CPU

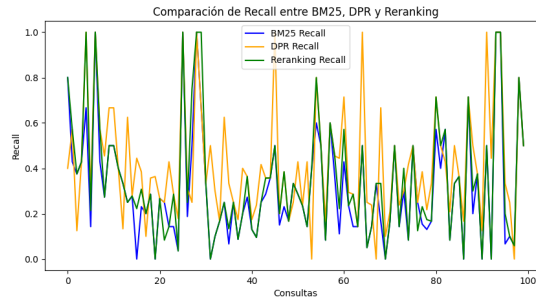


Fig. 2. Uso de CPU

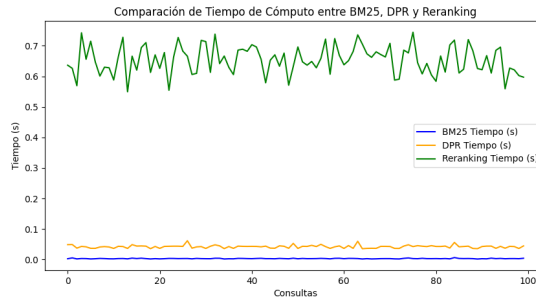


Fig. 3. Uso de CPU

7 Bibliografía

References

1. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, **3**(4), 333–389 (2009). [BM25](#)
2. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., et al.: Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781 (2020). [DPR](#)
3. Jia-Huei Ju, Jheng-Hong Yang, and Chuan-Ju Wang. 2021. Text-to-text multiview learning for passage re-ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. [Reranking](#)
4. Yung-Sung Chuang, Wei Fang, Shang-Wen Li, Wen-tau Yih, and James Glass. 2023. Expand, Rerank, and Retrieve: Query Reranking for Open-Domain Question Answering. [arXiv:2305.17080](#).