Sistemas de recuperación aumentada (RAG): Una propuesta de investigación para potenciar las búsquedas semánticas y el contexto interactuando con inteligencia artificial generativa

PONCE, María Paula; MIGO, Gabriel Alejandro; ISTVAN, Romina

Universidad Tecnológica Nacional, Facultad Regional La Plata

Laboratorio de Ingeniería en Sistemas de Información, LINES

Av. 60 s/n° esquina 124, CP 1900, La Plata, Buenos Aires, Argentina.

{mpaulaponce; gmigo; ristvan}@frlp.utn.edu.ar

RESUMEN

Los modelos de lenguaje a gran escala (LLM), han demostrado un rendimiento sobresaliente en una variedad de tareas de procesamiento de lenguaje natural (PLN), incluyendo generación de texto, traducción automática y respuesta a preguntas.

Sin embargo, estos modelos enfrentan desafíos significativos en términos de eficiencia computacional y manejo de información externa. La arquitectura RAG (Sistema de Recuperación Aumentada) emerge como una solución con gran potencial para abordar estas limitaciones.

Este proyecto de investigación se enfoca en mejorar la eficiencia y precisión de la recuperación de información basada en texto mediante el uso de RAG, los cuales combinan técnicas de recuperación de información vectoriales con modelos de lenguaje avanzados para mejorar la relevancia de los resultados de búsqueda. El objetivo es explorar diferentes enfoques y técnicas dentro de los RAG para abrir nuevas oportunidades en la búsqueda semántica y el descubrimiento de conocimiento. Al integrar estos sistemas en investigaciones particulares, se espera poder interactuar con grandes modelos de lenguajes (LLM) enviando un contexto para acotar su ámbito de conocimiento para la IA generativa.

Palabras clave: Sistemas de Recuperación Aumentada (RAG), Búsqueda Semántica, Contexto, Inteligencia Artificial Generativa, Grandes Modelos de Lenguaje (LLM), Base de datos Vectoriales.

CONTEXTO

En el campo de la inteligencia artificial generativa, los modelos de lenguaje de gran tamaño (LLM) se entrenan con grandes volúmenes de datos y usan miles de millones de parámetros para generar resultados originales

en tareas como responder preguntas, traducir idiomas y completar frases.

En el último tiempo, se ha observado un creciente interés en mejorar la eficacia utilizando contextos como dato de entrada de los grandes modelos de lenguajes (LLM).

Los RAG extienden las ya poderosas capacidades de los LLM a dominios específicos o a la base de conocimientos interna de una organización, todo ello sin la necesidad de volver a entrenar el modelo. Se trata de un método para mejorar los resultados de los LLM de modo que sigan siendo relevantes, precisos y útiles en diversos contextos.

Utilizan técnicas avanzadas de procesamiento de lenguaje natural (PLN) y modelos vectoriales para representar documentos de texto de manera semánticamente significativa. La búsqueda en bases de datos vectoriales permite realizar comparaciones de similitud semánticas, aplicando distintos métodos matemáticos, siendo la similitud coseno uno de los más utilizados. Con ella se compara el vector de una consulta de búsqueda con los vectores de documentos almacenados, aquellos que tengan una similitud coseno más alta con la consulta se consideran más relevantes y, por lo tanto, se presentan como resultados de búsqueda prioritarios. Por lo que una comprensión más profunda del contenido de los documentos facilita la recuperación de información relevante incluso en casos de ambigüedad o variabilidad léxica.

Dentro de este contexto, los sistemas de recuperación aumentada (RAG) representan una innovadora evolución. Estos sistemas combinan los enfoques vectoriales de recuperación de información con la capacidad de comprensión de los modelos de lenguaje (LLM). La integración de RAG en el proceso de búsqueda semántica promete mejorar aún más la relevancia y la precisión de los resultados de los LLM al capturar mejor la intención del usuario y el contexto del documento.

1. INTRODUCCIÓN

La creciente disponibilidad de datos textuales y la demanda de sistemas de recuperación de información más efectivos han llevado al desarrollo de innovadoras técnicas que buscan mejorar la eficiencia y precisión en la IA Generativa.

Se ha demostrado que los grandes modelos de lenguaje previamente entrenados almacenan conocimiento fáctico en sus parámetros y logran resultados de última generación cuando se ajustan con precisión en tareas posteriores de PLN. Sin embargo, su capacidad para acceder y manipular con precisión el conocimiento aún es limitada. Además, la problemática de determinar la procedencia de sus decisiones introduce imprevisibilidad en las respuestas del LLM y la difícil y costosa tarea de actualizar su conocimiento del mundo siguen siendo problemas de investigación abiertos, por lo que los datos de entrenamiento del LLM suelen ser estáticos e introducen una fecha límite en los conocimientos que tienen.

Entre los desafíos o potenciales riesgos conocidos de los LLM se incluyen:

- Presentar información falsa cuando no tiene la respuesta (alucinaciones).
- Presentar información desactualizada o genérica cuando el usuario espera una respuesta específica y actual.
- Crear una respuesta de fuentes no autorizadas.

- Crear respuestas inexactas debido a una confusión terminológica, en la que diferentes fuentes de entrenamiento utilizan la misma terminología para hablar de cosas diferentes.
- Sufrir filtraciones de datos privados.
- Generar contenido inapropiado, dañino o engañoso.

Sin el RAG, el LLM toma la información del usuario y crea una respuesta basada en la información con la que fue entrenado. Con el RAG, se introduce un componente de recuperación de información que utiliza la entrada del usuario para extraer primero la información de un nuevo origen de datos. La consulta del usuario y la información relevante se proporcionan al LLM, quién hace uso de la cuestión generativa para armar la mejor respuesta basada en el contexto proporcionado.

Esta investigación no solo busca avanzar en el campo de la recuperación de información basada en texto, sino también abrir nuevas oportunidades en la búsqueda semántica, el descubrimiento de conocimiento y la generación de texto de la inteligencia artificial generativa. Al integrar sistemas RAG en nuestras investigaciones, esperamos no solo mejorar la calidad de los resultados de búsqueda, sino también facilitar múltiples tareas, como la realización de preguntas y respuestas sobre el contenido de interés.

Entre algunos beneficios de utilizar estos sistemas se incluyen:

1. Información actualizada: Incluso si los orígenes de datos de entrenamiento originales para un LLM son adecuados para sus necesidades, es difícil mantener la relevancia. Mediante el RAG se puede entonces proporcionar la información más reciente a los usuarios.
2. Mayor confianza en la solución de IA Generativa: El RAG permite presentar información precisa con la atribución de la fuente. La salida puede incluir citas o referencias a fuentes.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

La línea de investigación propuesta se centra en el desarrollo y la aplicación de sistemas de recuperación aumentada (RAG) en el contexto de las bases de datos vectoriales para la búsqueda semántica y su integración con grandes modelos de lenguaje en el ámbito de la inteligencia artificial generativa. Esta línea de investigación aborda el desafío de mejorar la generación de respuestas contextualmente adecuadas y la toma de decisiones informadas, evitando problemas ya conocidos de los LLM como las alucinaciones.

Otro aspecto clave es la mejora de la calidad de los resultados de búsqueda mediante la búsqueda semántica, ya que las soluciones de búsqueda convencionales o de palabras clave producen resultados limitados para tareas intensivas en conocimiento, mientras que las tecnologías de búsqueda semántica pueden escanear grandes bases de datos de información dispar y multimodal y recuperar datos con mayor precisión.

El proyecto se llevará a cabo en varias etapas, que incluyen investigación, desarrollo, implementación y evaluación de técnicas de RAG en el contexto de la búsqueda semántica.

Dado que RAG consta de múltiples componentes que deben evaluarse por separado y en combinaciones, se necesitará un conjunto de métricas de evaluación. Se utilizarán medidas de rendimiento estándar para evaluar

la eficacia de las técnicas propuestas. Algunas de las métricas para tener en cuenta en la evaluación de los distintos aspectos del RAG son:

- Fidelidad (faithfulness): Mide la coherencia fáctica de la respuesta generada frente al contexto dado. Se refiere a la capacidad de una respuesta generada por un sistema para ser fiel a los hechos presentados en los contextos proporcionados. Es decir, ¿la respuesta generada refleja correctamente la información contenida en los contextos?
- Relevancia de la respuesta, se centra en evaluar qué tan pertinente-relevante es la respuesta generada para la pregunta dada.
- Precisión del contexto: es una métrica que evalúa si todos los contextos obtenidos de la búsqueda vectorial tienen una clasificación alta. Lo ideal es que todos los fragmentos relevantes aparezcan en los primeros puestos.
- Context-recall: La medida de recuperación de contexto evalúa si se recuperó toda la información relevante requerida para responder la pregunta.
- Aspectos críticos: está diseñado para evaluar que la respuesta no contenga contenido inapropiado, dañino o engañoso. Además, los usuarios tienen la flexibilidad de definir sus propios aspectos para evaluar las presentaciones según sus criterios específicos.

  Asimismo, se considerarán las implicaciones éticas relacionadas con el uso de sistemas de inteligencia artificial en la recuperación de información y la generación de texto, incluida la privacidad de los datos, la equidad y la transparencia en los algoritmos utilizados. Se seguirán pautas éticas establecidas para garantizar que el proyecto se lleve a cabo de manera responsable y ética.

3. RESULTADOS OBTENIDOS / ESPERADOS

Se espera que este proyecto genere avances en la mejora de la respuesta obtenida por los grandes modelos de lenguajes, a través de pasarle un contexto obtenido con los métodos de búsqueda semántica. Estos resultados tendrán implicaciones importantes en diversos campos, desde la investigación en recuperación de información hasta el desarrollo de tecnologías de procesamiento de lenguaje natural más avanzadas.

Atendiendo al objetivo general, el proyecto contribuye a:

1. Mejorar la respuesta de los grandes modelos de lenguaje: Se espera que la implementación de sistemas RAG conduzca a una mejora significativa en la respuesta basada en inteligencia artificial generativa. Esto se traducirá en una mayor relevancia de los resultados de búsqueda y una reducción en el tiempo requerido para encontrar la información deseada. Incluso si los orígenes de datos de entrenamiento originales para un LLM son adecuados para sus necesidades, es difícil mantener la relevancia.
2. Mayor Comprensión del Contenido de los Documentos: Los sistemas RAG permitirán una comprensión más profunda del contenido de los documentos al capturar el contexto y la intención del usuario. Esto facilitará la identificación de información relevante incluso en casos de ambigüedad o variabilidad léxica, lo que mejorará la calidad de los resultados de búsqueda y de respuesta.
3. Integración Fluida con Grandes Modelos de Lenguaje (LLM): La integración de los

resultados de la búsqueda semántica, como contexto para los LLM, permitirá una interacción más fluida con estos modelos. Esto facilitará tareas como la generación de respuestas contextualmente adecuadas y la realización de preguntas y respuestas sobre el contenido de interés.

4. Aplicaciones en la Inteligencia Artificial Generativa: Se espera que los avances logrados en este proyecto tengan aplicaciones significativas en el campo de la inteligencia artificial generativa. La capacidad de utilizar sistemas RAG para enviar contexto a los LLM abre nuevas posibilidades para la generación de texto y la toma de decisiones informadas.
5. Contribuciones a la Investigación en Búsqueda Semántica: Este proyecto contribuirá al avance en el campo de la búsqueda semántica al desarrollar y aplicar técnicas avanzadas de recuperación de información basadas en vectores. Los resultados obtenidos proporcionarán nuevas perspectivas y enfoques para mejorar la relevancia y la precisión de los resultados de búsqueda en este ámbito.

4. FORMACIÓN DE RECURSOS HUMANOS

El equipo de trabajo está conformado por un director, un Codirector, dos Docentes Investigadores y diez alumnos becarios de investigación. Cuenta con dos tesis de Maestría.

Se proporciona capacitación y orientación para garantizar el éxito del proyecto y el desarrollo profesional de los miembros del equipo.

5. BIBLIOGRAFÍA

[1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

[2] Petroni, F., Lewis, P., Piktus, A., Rocktäschel, T., Wu, Y., Miller, A. H., & Riedel, S. (2020). How context affects language models' factual predictions. *arXiv preprint arXiv:2005.04611*.

[3] Li, H., Su, Y., Cai, D., Wang, Y., & Liu, L. (2022). A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*.

[4] Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

[5] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. In EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (pp. 6769-6781). (EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference). Association for Computational Linguistics (ACL).

[6] Jiang, Z., Araki, J., Ding, H., & Neubig, G. (2021). How can we know when language models know? on the calibration of language models for question answering. *Transactions of*