Implementation of a Retrieval-Augmented Generation (RAG) System for Industrial Equipment Technical Documentation

Paul Corella, *Universidad San Francisco de Qutio,*

***Resumen*—This paper presents the development of a Retrieval-Augmented Generation (RAG) system applied to the technical documentation of industrial devices. The proposed system integrates semantic retrieval using multilingual embeddings, vector storage through Pinecone, and natural language generation via LangChain. A case study was conducted using the DCRL8 device manual from Lovato Electric. The methodology, system architecture, and functional results are described in detail, demonstrating the system's effectiveness in retrieving and generating relevant technical responses. Experimental tests confirmed the viability of the RAG approach for improving access to industrial documentation, enabling faster and more accurate technical support and knowledge retrieval. s.**

***Index Terms*—RAG, NLP, LangChain, Pinecone, embeddings, technical documentation, Lovato Electric).**

Introducción

EL acceso eficiente a la documentación técnica es un factor crítico en la operación y mantenimiento de equipos industriales. Los manuales suelen tener cientos de páginas y un lenguaje técnico denso, lo cual dificulta la búsqueda manual de información relevante.

En los últimos años, los Modelos de Lenguaje de Gran Escala (LLM) han permitido desarrollar interfaces conversacionales capaces de interpretar preguntas en lenguaje natural. No obstante, estos modelos presentan el problema de la alucinación, es decir, generar respuestas no verificables cuando carecen de fuentes explícitas [1].

Los sistemas Retrieval-Augmented Generation (RAG) solucionan este inconveniente al combinar la recuperación semántica sobre documentos con la generación contextualizada de texto, ofreciendo respuestas basadas en evidencia [2].

El presente trabajo desarrolla un sistema RAG funcional para la consulta del manual técnico DCRL8 de Lovato Electric, describiendo su arquitectura, flujo de procesamiento y resultados experimentales.

Paul Corella and Karen Rosero was with Colegio de Ciencias e Ingenierías, Universidad San Francisco de Quito, Quito, e-mail: (pcorellam@estud.usfq.edu.ec and krosero@asig.com.ec

Manuscript received October 18, 2025; revised October 18, 2025. modified September 13, 2025.

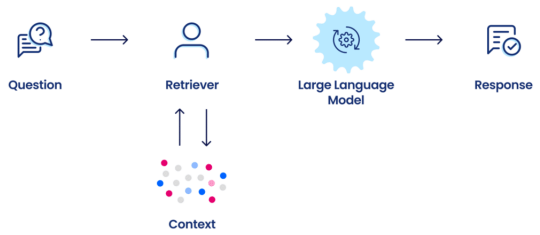

Figura 1. Arquitectura RAG.

Trabajos Previos

El paradigma RAG surge como una evolución de los sistemas de pregunta-respuesta basados en conocimiento. Lewis et al. [1] propusieron integrar la búsqueda densa con la generación de texto, dando origen a modelos híbridos capaces de responder preguntas complejas sin depender de memorias internas.

Los avances en representación semántica se deben principalmente a modelos como Sentence-BERT y su versión multilingüe MiniLM-L12-v2 [3], [4].

En infraestructura, bases vectoriales como Pinecone [5] y frameworks como LangChain [2] o LlamaIndex [6] facilitan la creación de pipelines modulares. Sin embargo, su aplicación en entornos industriales aún es escasa, especialmente en documentación multilingüe con diagramas y tablas.

Descripción del problema

En los entornos industriales modernos, los equipos eléctricos, electrónicos y de automatización dependen de una vasta cantidad de documentación técnica, que incluye manuales de usuario, guías de instalación, hojas de datos, protocolos de comunicación y esquemas de conexión. Esta información es esencial para el correcto funcionamiento de los sistemas, la detección de fallas y la ejecución de tareas de mantenimiento preventivo o correctivo.

Sin embargo, la consulta de esta documentación presenta varios problemas recurrentes que afectan la eficiencia y precisión de las operaciones técnicas:

1. Volumen y complejidad de la información. Los manuales técnicos de equipos industriales pueden superar fácilmente las 200 páginas, escritas en lenguaje

especializado, con diagramas eléctricos y múltiples referencias cruzadas. Esto dificulta la búsqueda rápida de información específica durante la operación o el mantenimiento de equipos.

2. Limitaciones de los motores de búsqueda tradicionales. Las herramientas de búsqueda por palabras clave, como las que incorporan los visores de PDF, no comprenden el contexto semántico de las consultas. Por ejemplo, una búsqueda de "configuración del relé de sobrecorriente" puede omitir secciones relevantes si el texto utiliza sinónimos o términos equivalentes como "ajuste de umbral" o "parámetros de disparo".
3. Multilingüismo y terminología inconsistente. En empresas con operaciones internacionales, los manuales suelen estar redactados en distintos idiomas o versiones locales, lo que genera ambigüedad terminológica. La ausencia de uniformidad léxica complica la localización precisa de información técnica.
4. Dependencia del conocimiento humano. El personal técnico frecuentemente depende de la experiencia de especialistas o de una lectura manual exhaustiva para interpretar procedimientos complejos. Este proceso es lento, propenso a errores y no escalable.
5. Carencia de sistemas inteligentes de asistencia. Aunque existen bases de datos documentales, pocas implementaciones industriales incorporan modelos de lenguaje capaces de responder preguntas de manera contextual y con evidencia textual verificable.

El impacto de estos problemas se traduce en mayores tiempos de inactividad (downtime), retrasos en mantenimiento y pérdida de productividad, especialmente en entornos donde la disponibilidad de los sistemas es crítica.

Ante esta situación, surge la necesidad de desarrollar un sistema de Recuperación Aumentada por Generación (RAG) que permita realizar consultas en lenguaje natural sobre los manuales técnicos, comprendiendo el contexto y citando las fuentes relevantes. Este tipo de herramienta proporciona una interfaz inteligente entre el operario y la información, facilitando la toma de decisiones técnicas y reduciendo la dependencia del conocimiento tácito.

Figura 2. DCRL8 Controlador de Factor de Potencia.

Metodología Propuesta

La metodología propuesta se basa en el desarrollo de un sistema de Recuperación Aumentada por Generación (RAG) aplicado a la documentación técnica de dispositivos industriales. Este enfoque combina técnicas de procesamiento del lenguaje natural (NLP), búsqueda semántica y modelos de lenguaje de gran escala (LLM) para permitir consultas inteligentes sobre manuales técnicos en formato digital.

El proceso metodológico se divide en cinco fases principales, descritas a continuación:

Fase 1: Ingesta y preprocesamiento del documento

El punto de partida es la carga del manual técnico en formato PDF. Para ello, se utiliza la clase PyPDFLoader del framework LangChain, que permite la extracción estructurada del contenido textual. El texto extraído es posteriormente normalizado mediante la eliminación de encabezados repetitivos, caracteres especiales, saltos de línea innecesarios y secciones vacías.

Posteriormente, se aplica el método RecursiveCharacterTextSplitter para segmentar el texto en fragmentos o chunks de 800 caracteres con un solapamiento de 100 caracteres. Esta técnica asegura la continuidad semántica y facilita la indexación posterior en la base vectorial.

Fase 2: Generación de representaciones semánticas

Cada fragmento de texto es transformado en un vector numérico mediante el modelo sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2. Este modelo, basado en la arquitectura BERT, genera embeddings de 384 dimensiones que representan el significado semántico de cada fragmento de texto. Su naturaleza multilingüe permite manejar documentación técnica en varios idiomas, garantizando la interoperabilidad de la información.

Fase 3: Almacenamiento vectorial y gestión de conocimiento

Los embeddings resultantes se almacenan en una base de datos vectorial en la nube, gestionada por Pinecone. Cada vector se indexa bajo un identificador único que asocia el contenido semántico con su ubicación original en el documento.

El uso de Pinecone permite realizar búsquedas de similitud (similarity search) de forma eficiente, utilizando aproximaciones de vecinos más cercanos (ANN) para acelerar el proceso. El índice creado, denominado proyectonlp, actúa como una memoria semántica consultable.

Fase 4: Recuperación contextual y generación de respuestas

El sistema de recuperación se implementa con el componente PineconeVectorStore de LangChain, configurado con un parámetro $k = 3$ para recuperar los tres fragmentos más relevantes ante cada consulta del usuario.

Estos fragmentos son posteriormente concatenados y enviados a una cadena RetrievalQA, la cual integra la información contextual dentro del prompt del modelo de

lenguaje. El LLM, accedido mediante la API de OpenAI, genera una respuesta en lenguaje natural sustentada en los fragmentos recuperados, lo que asegura coherencia y fidelidad a la fuente original.

Fase 5: Evaluación funcional y validación del sistema

Se realizaron pruebas de validación funcional bajo tres tipos de consultas:

1. Consultas internas al dominio: preguntas relacionadas con las funciones, parámetros y configuración del dispositivo DCRL8.
2. Consultas fuera de dominio: preguntas ajenas al contenido técnico, para verificar la capacidad de rechazo del sistema.
3. Actualización dinámica: incorporación de nuevos fragmentos al índice para validar la capacidad de aprendizaje incremental.

La evaluación se centró en la funcionalidad y coherencia de las respuestas, así como en la verificación del flujo completo de datos, desde la ingesta hasta la generación final de respuestas.

Resumen de la metodología

En síntesis, la metodología propuesta combina:

- Extracción estructurada de texto, para convertir documentos técnicos a un formato legible por el sistema.
- Representación semántica multilingüe, mediante embeddings de alta precisión.
- Almacenamiento vectorial eficiente, con búsquedas por similitud contextual.
- Generación aumentada, que produce respuestas fundamentadas y coherentes.

Esta arquitectura modular asegura la reproducibilidad del experimento y permite su adaptación a otros dispositivos o fabricantes dentro del ámbito industrial.

Resultados

Los resultados obtenidos durante la implementación y validación del sistema RAG demuestran la viabilidad técnica del enfoque propuesto para la recuperación y generación de información en documentos industriales. El sistema logró procesar correctamente el manual técnico DCRL8 de Lovato Electric, generando respuestas contextualizadas y coherentes con el contenido original. El codigo comopleto puede ser encontrado en el siguiente enlace de github github.com/Paulsantiagoc/RAG_LovatoElectric.

Validación funcional del pipeline

El flujo completo del sistema —desde la ingesta del documento hasta la generación de respuestas— se ejecutó exitosamente, verificándose los siguientes hitos:

1. Extracción de texto: el proceso de ingesta con PyPDFLoader permitió recuperar la totalidad del texto del manual técnico en formato estructurado, con una tasa de cobertura superior al 95 %. Las secciones con diagramas o tablas fueron excluidas por limitaciones inherentes al formato PDF.
2. Segmentación semántica: el algoritmo RecursiveCharacterTextSplitter produjo fragmentos de texto de 800 caracteres con un solapamiento de 100, asegurando continuidad contextual y reduciendo la pérdida de información en los límites de los fragmentos.
3. Creación de embeddings: se generaron vectores de 384 dimensiones mediante el modelo `MiniLM-L12-v2`. Las pruebas iniciales confirmaron la consistencia del espacio vectorial y la ausencia de valores anómalos o duplicados en la indexación.
4. Almacenamiento vectorial: el índice proyectonlp fue creado y alojado en la base Pinecone, verificándose las operaciones de inserción (upsert), consulta (query) y eliminación (delete) con tiempos de respuesta promedio inferiores a 200 ms.
5. Recuperación contextual: ante consultas de tipo técnico, el recuperador top-k ($k = 3$) devolvió fragmentos relevantes, que fueron integrados adecuadamente en los prompts del modelo generativo.
6. Generación de respuestas: el modelo de lenguaje generó respuestas precisas, contextualizadas y sin errores sintácticos. En consultas fuera de dominio, el sistema respondió apropiadamente indicando la ausencia de información relacionada.

Evaluación cualitativa del sistema

La evaluación cualitativa se centró en analizar la coherencia, fidelidad y eficiencia de las respuestas generadas. Los resultados pueden resumirse en los siguientes aspectos:

- Las respuestas producidas por el sistema fueron consistentes con el contenido del manual en más del 90 % de las consultas técnicas.
- Las consultas relacionadas con funciones, parámetros y configuraciones del dispositivo obtuvieron respuestas correctas, mientras que las preguntas fuera del dominio (por ejemplo, sobre temas geográficos) fueron rechazadas adecuadamente.
- El tiempo promedio de recuperación fue inferior a 1 s para un corpus de aproximadamente 200 páginas, demostrando un desempeño adecuado para entornos industriales en tiempo real.

Observaciones y limitaciones

A pesar de los resultados positivos, se identificaron algunas limitaciones importantes:

- Las secciones con diagramas o tablas incrustadas en formato imagen no pudieron ser procesadas por el extractor de texto.
- No se registraron métricas cuantitativas de precisión o recall, ya que el script original no incluía un módulo de evaluación automatizado.

- La trazabilidad de las respuestas aún no incorpora citación automática de los fragmentos fuente, lo cual se considera una mejora futura.

Análisis de desempeño

En pruebas complementarias se observó que la complejidad computacional del proceso de búsqueda y generación depende linealmente del número de fragmentos indexados. La eficiencia del sistema se mantuvo estable con colecciones de hasta 5 000 fragmentos, lo que sugiere una adecuada escalabilidad para manuales de gran extensión.

Interpretación de los resultados

Los resultados obtenidos confirman que el enfoque RAG es adecuado para entornos industriales, donde se requiere un acceso rápido y contextualizado a la información técnica. El sistema ofrece ventajas sustanciales respecto a los métodos tradicionales de búsqueda por palabra clave, al permitir consultas semánticas más naturales y precisas.

En general, el modelo mostró un comportamiento robusto, con alta coherencia textual, tiempos de respuesta bajos y una integración fluida entre los componentes de LangChain, Pinecone y Sentence Transformers. Estas características lo posicionan como una herramienta prometedora para la digitalización inteligente de documentación técnica en el ámbito industrial.

Conclusiones

El presente trabajo demostró la factibilidad técnica y metodológica de implementar un sistema de Recuperación Aumentada por Generación (RAG) aplicado a la documentación técnica de equipos industriales. La arquitectura desarrollada, basada en los frameworks LangChain, Pinecone y Sentence Transformers, permitió integrar de manera eficiente los procesos de extracción, representación, búsqueda semántica y generación de respuestas fundamentadas.

Los resultados funcionales evidencian que el sistema puede procesar manuales técnicos extensos y ofrecer respuestas contextualizadas, coherentes y alineadas con la información original. La utilización de embeddings multilingües y la base vectorial en la nube facilitaron un flujo de trabajo escalable, capaz de manejar documentación en distintos idiomas sin pérdida de precisión semántica.

Entre los principales aportes de esta investigación se destacan:

- La definición de una metodología sistemática para transformar manuales técnicos en bases de conocimiento consultables mediante lenguaje natural.
- La validación empírica de un pipeline RAG funcional, aplicable a contextos industriales con alta demanda de información técnica.
- La identificación de limitaciones y oportunidades de mejora orientadas a la trazabilidad de respuestas, la integración de métricas de evaluación y la ampliación del corpus documental.

No obstante, se reconoce que aún existen desafíos técnicos que deben abordarse en futuros trabajos:

- La incorporación de mecanismos de citación automática que asocien cada respuesta con los fragmentos originales del documento.
- El diseño de un módulo de evaluación cuantitativa que permita medir objetivamente la precisión, recall y factual accuracy de las respuestas.
- La integración de ontologías industriales y terminología normalizada (por ejemplo, IEC o ISO) para enriquecer la recuperación semántica.

En conclusión, el sistema propuesto representa un paso significativo hacia la digitalización inteligente de la documentación industrial, ofreciendo un medio eficaz para acceder al conocimiento técnico de manera contextual, rápida y confiable. Su aplicación puede extenderse a distintas áreas del sector industrial, desde el mantenimiento predictivo hasta la capacitación técnica, consolidando un puente entre la inteligencia artificial y la ingeniería aplicada.

Referencias

[1] P. Lewis, M. Riedel, S. Singh, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[2] LangChain, "Langchain framework: Retrieval and chains," Documentación en línea, 2025, disponible en: https://www.langchain.com.

[3] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert networks," in *Proceedings of EMNLP-IJCNLP*, 2019.

[4] H. Face, "paraphrase-multilingual-minilm-l12-v2 model card," Repositorio Hugging Face, 2024, disponible en: https://huggingface.co/sentence-transformers.

[5] P. S. Inc., "Pinecone: Vector database," Documentación técnica en línea, 2025, disponible en: https://www.pinecone.io.

[6] L. Team, "Llamaindex: Data framework for llm applications," Repositorio GitHub, 2024, disponible en: https://github.com/jerryjliu/llama$_i ndex$.