

LEGAL DOCUMENT RETRIEVAL USING SEMANTIC EMBEDDINGS

Tran Vinh Khanh

¹ Trường ĐH.....

² University of Science
HCMC, Vietnam

³ National Institute of Informatics

What ?

- We propose a semantic embedding-based framework for legal document retrieval.
- We represent user queries and legal texts in a shared semantic vector space.
- We perform semantic retrieval and ranking of legal text segments that best match natural language queries.

Why ?

- The number of legal documents is rapidly increasing and becoming more complex, making accurate retrieval more difficult.
- Keyword-based search systems often fail to capture the true semantic intent of user queries.
- Legal queries are frequently expressed in natural language that differs from formal legal terminology used in statutes.
- This work has high practical value for legal information retrieval and automated legal assistance systems.

Overview

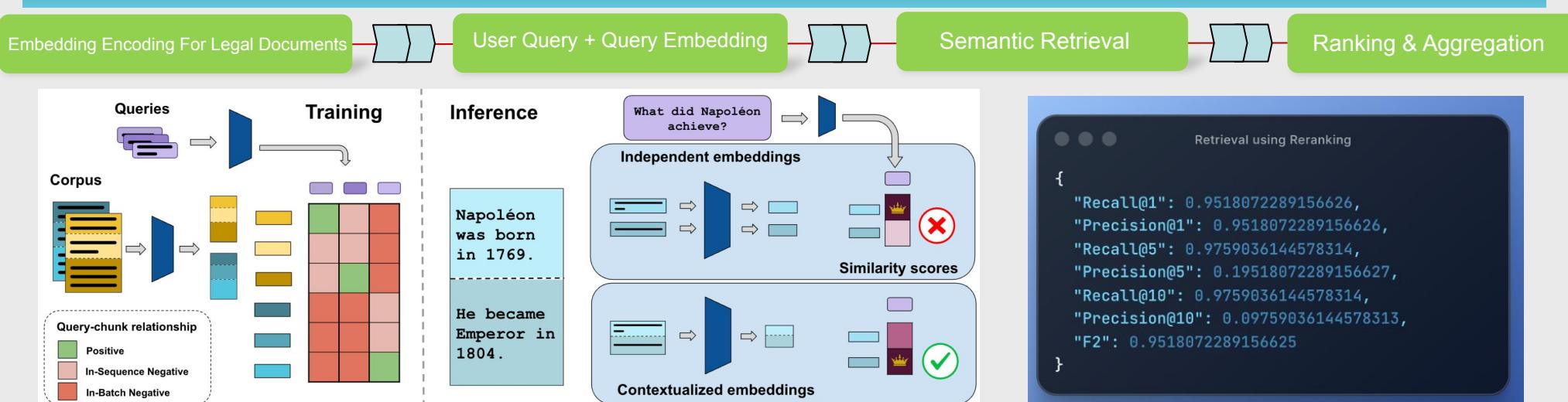


Figure 1. Overview of Semantic Embedding-Based Legal Document Retrieval

Figure 2. Results on Private Test

Description

1. Embedding Encoding For Legal Documents

- Legal documents are collected from public and official sources and normalized into a unified format.
- Each text segment is encoded into a semantic embedding vector using a pre-trained language model.
- All document embeddings are stored in a vector space database, enabling efficient similarity-based retrieval.

3. Face-Track Matching

- The system performs similarity search in the vector space to retrieve the top-k legal text segments that are most semantically relevant to the query.
- Semantic similarity is computed using standard metrics such as cosine similarity.
- This stage efficiently reduces the search space from the entire corpus to a small set of relevant candidates.

2. User Query & Query Embedding

- Users submit queries in natural language, reflecting real-world legal information needs.
- The query is preprocessed and encoded into a semantic embedding in the same vector space as legal documents.

4. Ranking & Aggregation

- Retrieved text segments are ranked based on semantic similarity scores.
- Segments belonging to the same legal document are aggregated to compute a document-level relevance score.
- Final results are sorted in descending order of relevance and returned to the user.

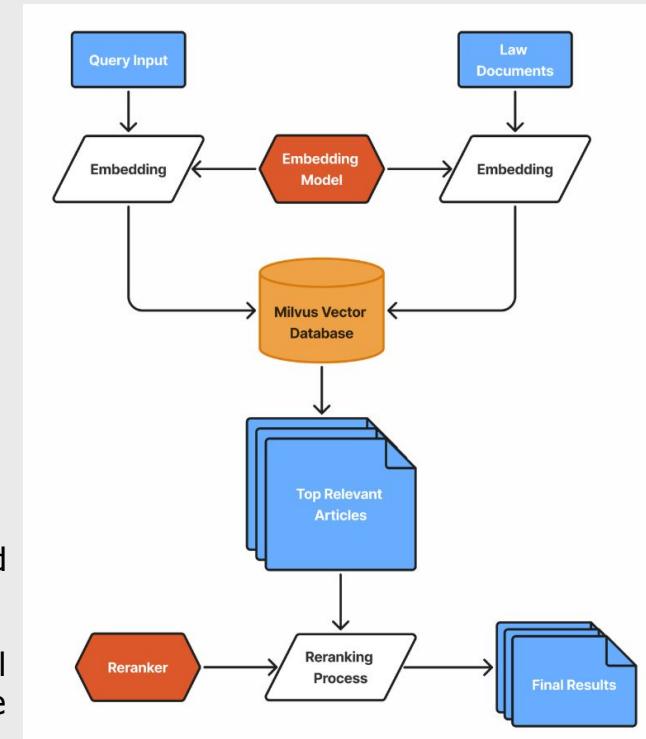


Figure 3. Overview of System