GMR-229 Semantic Search using Sentence Transformers

Abstract

Traditional keyword-based search engines often miss the contextual meaning in user queries. This research develops a semantic search engine leveraging Sentence Transformers to improve retrieval by understanding query and document context. By generating sentence embeddings, we enable retrieval based on semantic similarity, not mere keyword matching. Our process includes data preprocessing, feature extraction, and ranking documents by cosine similarity to query embeddings. Initial testing shows enhanced relevance compared to the BM25 baseline, indicating Sentence Transformers' effectiveness. This work highlights how semantic search engines can deliver contextually aligned results, paving the way for further improvements in natural language-based retrieval systems.

Introduction

The rapid growth in digital content has revolutionized information retrieval, making search engines vital tools for accessing relevant data. However, traditional keyword-based search engines are limited by their reliance on exact word matching, often failing to understand the context and intent behind user queries. This results in inaccurate or irrelevant search outcomes, as these engines lack the capability to capture semantic meaning. With the increasing volume of online content, there is a pressing need for retrieval systems that can interpret the nuances of user queries and deliver more context-aware results.

To address these limitations, we propose a semantic search engine using Sentence Transformers, a deep learning model that generates rich, contextual embeddings for both queries and documents. Our approach enhances search relevance through three key contributions: (1) a preprocessing pipeline for standardized text data, (2) Sentence Transformer embeddings for semantic similarity matching, and (3) a ranking function based on cosine similarity. Preliminary results, compared against the BM25 baseline, highlight the effectiveness of our method in improving search relevance and user experience. This work opens new possibilities for more sophisticated, intent-driven retrieval systems.

Research Question(s)

1. How effectively does a semantic search engine, using Sentence Transformers, improve retrieval relevance compared to traditional BM25-based methods on the 20 Newsgroups dataset?
2. How does contextual understanding of queries and documents influence retrieval quality in a semantic search engine compared to a lexical search engine?
3. What impact does document length and content diversity in the 20 Newsgroups dataset have on the performance of Sentence Transformers in semantic search?
4. To what extent does fine-tuning Sentence Transformers on the 20 Newsgroups dataset improve retrieval performance, and is it necessary for high-quality results?

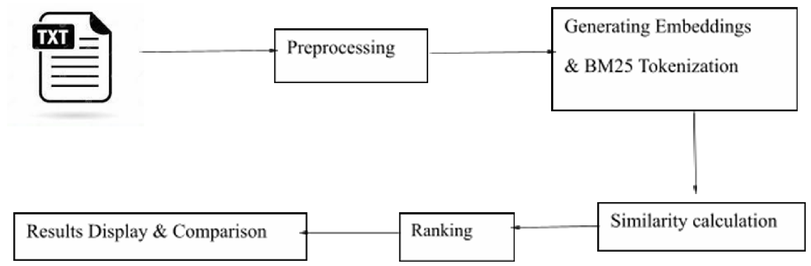
Materials

Dataset: The 20 Newsgroups dataset, containing articles across 20 topics, was used to train and test our semantic search engine.
Tools: Implemented in Python with libraries including PyTerrier, scikit-learn, pandas, and Sentence Transformers for embeddings. BM25 was used as a baseline model, while Sentence Transformers handled contextual embeddings and ranking.
Evaluation Metrics: Compared retrieval performance of the semantic search engine against the BM25 baseline to assess improvements in retrieval relevance.

Methods

Data Preprocessing: Text data was normalized by tokenizing, removing stop words, and lowercasing.
Embedding Generation: Created vector embeddings for documents and queries using Sentence Transformers for semantic relevance, while BM25 ranked based on keyword frequency.
Retrieval and Ranking: Cosine similarity was used for ranking semantic embeddings, compared against BM25 scores to assess improvements in retrieval relevance.

Pipeline:



Results

```
query = "Artificial intelligence and machine learning"
```

```
Semantic Search Results:
From: ipj@unix.brighton.ac.uk ((( Fleg Software )))
Subject: Artificial Intelligence in Medicine
Org
From: kempmp@phoenix.oulu.fi (Petri Pihko)
Subject: Re: Consciousness part II - Kev Strikes Back!
Or
From: shavlik@cs.wisc.edu (Jude Shavlik)
Subject: Program & Reg Forms: 1st Int Conf on Intell Sys fo
From: clarke@acme.ucf.edu (Thomas Clarke)
Subject: Re: How do you build neural networks?
Organizatio
From: gary@ke4zv.uucp (Gary Coffman)
Subject: Re: Math?? (Was US govt & Technolgy Investment
Keyword
```

```
BM25 Search Results:
From: shavlik@cs.wisc.edu (Jude Shavlik)
Subject: Program & Reg Forms: 1st Int Conf on Intell Sys fo
From: pedwards@csd.abdn.ac.uk (Pete Edwards x 2270)
X-Priority: 1 (Highest)
X-Sender: pedwards@139.1
From: oaf@zurich.ai.mit.edu (Oded Feingold)
Subject: where is
Organization: M.I.T. Artificial Intell
From: ipj@unix.brighton.ac.uk ((( Fleg Software )))
Subject: Artificial Intelligence in Medicine
Org
From: oaf@zurich.ai.mit.edu (Oded Feingold)
Subject: Re: UVA
Organization: M.I.T. Artificial Intelli
```

Conclusions

The implementation of the semantic search engine using Sentence Transformers demonstrated the potential to enhance information retrieval from the 20 Newsgroups dataset. Our research questions focused on improving the relevance of search results through the use of contextual embeddings compared to traditional keyword-based approaches like BM25.

The results indicated that the semantic search engine effectively captures the nuances of natural language, offering more relevant document recommendations for user queries. Overall, this project highlights the advantages of leveraging advanced embedding techniques in search applications, paving the way for further developments and optimizations in future research.

Acknowledgments

- Dr. Md Abdullah Al Hafiz Khan – Faculty Advisor
- Kennesaw State University College of Computing and Software Engineering (CCSE)
- Computing Showcase Day (C-Day) sponsors, judges and audience.
- Online Research Resources

Contact Information

- Arpana Challa
  email: achalla3@students.kennesaw.edu
- Roshni Satish
  email: rsatish@students.kennesaw.edu

References

- beeFormer: Bridging the Gap Between Semantic and Interaction Similarity in Recommender Systems | Vojtěch Vančura, Pavel Kordík, Milan Straka
- A Survey of Pre-trained Language Models for Processing Scientific Text | Xanh Ho, Anh Khoa, An Tuan dao, Junfeng Jiang and Yuki Chida

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Authors: Roshni Satish (LinkedIn: https://www.linkedin.com/in/roshni-satish/)
Arpana Challa (LinkedIn: https://www.linkedin.com/in/arpana-challa-6626a9196/)