Query-by-Example for Scientific Article Retrieval

Midterm Evaluation Presentation

Group - 4

Ashwani Kumar Kamal - 20CS10011
Hardik Pravin Soni - 20CS30023
Sourabh Soumyakanta Das - 20CS30051
Shiladitya De - 20CS30061

Table of Contents

1. Problem Description
2. Progress Made
3. Experiments Done
4. Results and Analysis
5. Future Plans
6. Work Division
7. References

Table of Contents

1. Problem Description
2. Progress Made
3. Experiments Done
4. Results and Analysis
5. Future Plans
6. Work Division
7. References

Description Of the Problem Statement

Domain: Scientific Publications/En
Background: Its very common for users to note to come up with perfect queries (following proper syntax or method). Hence, the idea of query by example (QBE) came into being.
In this project, documents are retrieved based on the rhetorical structure of a paper like background, method, results etc.
Task: In this project the task is to come up with methods to improve the different metrics associated with extracting relevant documents per query.

Table of Contents

1 Problem Description

2 Progress Made

3 Experiments Done

4 Results and Analysis

5 Future Plans

6 Work Division

7 References

Progress Made

- **Finalising the dataset**
  We have finalised two datasets relevant to the project description. These are as follows:
  - **CSFCube Dataset**
    Dataset containing annotatated test set for validation along with training
  - **Semantic Scholar Open Research Corpus (S2ORC) Dataset**
    Large Corpus of 800,000 papers related to computer science. This is to be considered for covering wide range of queries.

Progress Made

- **Base Models Implemented:**
  - **tf-idf** - This is the most basic model (vector space ranking model).
  - **BM25Okapi** - Standard Bm25Okapi Model from the BM25Okapi Library.
  - **Bert_NLI** - The sentence bert nli (from `'nli-roberta-base-v2'` of SentenceTransformer) is used.
  - **Bert_PP** - The sentence bert pp (from `'paraphrase-TinyBERT-L6-v2'` of SentenceTransformer) is used.
- **Paper Readings:**
  As such a large no. of papers have been covered. They have been listed in the References section.

Table of Contents

1 Problem Description

2 Progress Made

3 Experiments Done

4 Results and Analysis

5 Future Plans

6 Work Division

7 References

Experiments Carried Out

- **Training Phase:**
  In the Training Phase we trained the different models on our data. The relevant files are stored in `Results` directory under appropriate model names.
  The training is done for `all`, `background`, `method`, `results` sections of the paper as given in the dataset.
- **Testing Phase:**
  Cosine Similarity has been used as mentioned in the paper and corresponding results has been ranked as per the cosine similarity scores and different metrics like Ranked Precision, Precision@20 etc. are calculated.

Table of Contents

1 Problem Description

2 Progress Made

3 Experiments Done

4 Results and Analysis

5 Future Plans

6 Work Division

7 References

Results and Analysis

Metrics Implemented

To Measure, Express and Compare the Performances of our Models, we have carefully Implemented the Following Metrics :-

- Ranked Precision (RP)
- Precision at K (P@k)
- Recall at K (R@K)
- $NDCG_{\%20}$

The **results** obtained are presented in the following two slides.

Results and Analysis

1. Background

Model	Ranked Precision	P@20	R@20	NDCG%20
TF-IDF	0.1777	0.2266	0.3789	0.4795
SENTBERT-NLI	0.2004	0.2750	0.4328	0.5781
SENTBERT-PP	0.2332	0.3109	0.5024	0.5974

2. Method

Model	Ranked Precision	P@20	R@20	NDCG%20
TF-IDF	0.0892	0.0748	0.2434	0.2440
SENTBERT-NLI	0.1656	0.1028	0.3265	0.3393
SENTBERT-PP	0.1826	0.0998	0.3388	0.3865

Results and Analysis

3. Result

Model	Ranked Precision	P@20	R@20	NDCG%20
TF-IDF	0.1083	0.1333	0.3067	0.3851
SENTBERT-NLI	0.1278	0.1826	0.4023	0.4072
SENTBERT-PP	0.1548	0.2273	0.5484	0.5183

4. Aggregated

Model	Ranked Precision	P@20	R@20	NDCG%20
TF-IDF	0.1247	0.1437	0.3084	0.3676
SENTBERT-NLI	0.1643	0.1859	0.3866	0.4404
SENTBERT-PP	0.1898	0.2119	0.4631	0.4995

Table of Contents

1 Problem Description

2 Progress Made

3 Experiments Done

4 Results and Analysis

5 Future Plans

6 Work Division

7 References

Future Plans

- **Scraping the H-Index:** and citations of the authors to provide an additional feature in the search engine to get the results sorted by H-Index or citations and also by year of publishing.
- **Expanding to Full S2ORC Corpus and Further:** Pre-calculate embeddings and efficiently make candidate pool to make a general purpose research paper query tool.
- **Clean and Augment Dataset:** The data contains special Unicode characters and latex snippets. These can be removed or modified into natural language.
- **Applying Semantic Similarity and Intent:** By incorporating additional parameters to comprehend the semantic context of a user's queries, we can gain a more profound understanding of their intent. Techniques such as word embedding or Latent Semantic Analysis (LSA) can be utilized to generate vector representations of the user's queries.

Future Plans

- **Applying Neural Networks and related losses to improve the classification models:** There are several other machine learning techniques that can be used for generating vector representations of text in the documents for measuring semantic similarity. Here are a Few:- Doc2Vec, Universal Sentence Encoder, Transformer-based models or Graph-Based models.
- **Improving Search Results through User Feedback Systems:** A user feedback system can be incorporated into Query By Example (QBE) using machine learning techniques to improve the relevance of search results. The system collects user feedback on search results and uses this feedback to continuously refine the search algorithm.

Table of Contents

1 Problem Description

2 Progress Made

3 Experiments Done

4 Results and Analysis

5 Future Plans

6 Work Division

7 References

Our Work Division

As such we all have put our heads into the discussions and each and every step that we took but we did a specific allocation of work to make things easier. The work division is as follows:

1. **Ashwani Kumar Kamal:** Writing the code to generate the embeddings of tf-idf, bert_pp and reading the papers.
2. **Hardik Pravin Soni:** Writing the code to generate the embeddings of bert_nli and bert_pp and reading about S2ORC dataset and reading the papers.
3. **Sourabh Soumyakanta Das:** Writing the code for computing the metrics for different models and analysing the result and preparing report.
4. **Shiladitya De:** Writing the code for computing the metrics and analysing the results as well as preparing presentation and reading the papers.

Table of Contents

1 Problem Description

2 Progress Made

3 Experiments Done

4 Results and Analysis

5 Future Plans

6 Work Division

7 References

References

- Sheshera Mysore and Tim O'Gorman and Andrew McCallum and Hamed Zamani. 2021. CSFCube - A Test Collection of Computer Science Research Articles for Faceted Query by Example. In Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). https://openreview.net/forum?id=8Y50dBbmGU
- Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 4969–4983. https://doi.org/10.18653/v1/2020.acl-main.447
- Chong Wang and David M. Blei. Collaborative Topic Modeling for Recommending Scientific Articles. https://dl.acm.org/doi/pdf/10.1145/2020408.2020480

References

- M.A. Angrosh, Stephen Cranefield and Nigel Stanger. Contextual Information Retrieval in Research Articles: Semantic Publishing Tools for the Research Community.
  `https://www.semantic-web-journal.net/system/files/swj169_1.pdf`
- Yuna Jeong and Eunhui Kim. SciDeBERTa: Learning DeBERTa for Science Technology Documents and Fine-Tuning Information Extraction Tasks.
  `https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9791256tag=1`

Thank You!