Using Topic-Based Query Paraphrasing to improve document retrieval in Learning to Rank Model

Tushita Singh
University Of Massachusetts Amherst

Shreya Dubey
University Of Massachusetts Amherst

ABSTRACT

Query Reformulation is a popular information retrieval method to retrieve relevant documents from a huge corpus for a given search query. Query paraphrasing is one of the effective techniques which improves retrieval effectiveness. However, query expansion using existing techniques gives good results for general queries but performs poorly on queries that require capturing the domain of the query. For domain-specific queries, the query expansion techniques do not take into consideration the relevance of a document in a particular domain. Also, training a model for domain-specific query reformulation requires a manually annotated dataset for every domain. Dataset annotation is expensive, time-consuming, and not a scalable option. Our project is aimed at using Topic-modeling for domain-specific query expansion to improve document retrieval for user-given queries. We have developed two models, one where the query paraphrasing is done by performing topic-modeling for query reformulation and the other uses a pre-trained Sentence BERT model. Both the models are learning to rank models where few documents are retrieved from a large collection using the BM25 retrieval model and then the retrieved documents are passed to a BERT model which is trained using the topic-based and paraphrased queries respectively. We evaluate and compare both models on different metrics including MAP and NDCG.

ACM Reference Format:
Tushita Singh and Shreya Dubey. 2022. Using Topic-Based Query Paraphrasing to improve document retrieval in Learning to Rank Model. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Query Expansion for effective Document Retrieval is a well-known task in the Information Retrieval Community. In recent times, user needs have been diverse and thus there is a need to formulate domain-specific models. For example, during the covid times, we saw a sharp increase in new scientific knowledge being shared in the form of papers, thus the need for a robust document retrieval system was observed. Consequently, it was observed that the words or vocabulary the users used for the search is diverse from the commonly known vocabulary, leading to improper retrieval of documents. In this paper, we try to tackle the above-mentioned problem by query paraphrasing. A good paraphrase should be adequate and fluent while being as different as possible on the surface lexical form so as to capture any outliers that were previously not possible with the user's query. We compare two models in the terms of query paraphrasing. In the first model, we made use of the general method of lexical paraphrases for improving document retrieval performance, which replaces the content words of the queries with their synonyms. In the second model, we formulate paraphrase queries by making use of the paraphrasing model and compare its document retrieval performance. The paraphrased queries generated are used to get the top few documents by BM25 retrieval and then the documents are re-ranked using a BERT model. We believe this would help bring out the previously lower-ranked relevant documents. Our evaluation metrics assessed the effect of query paraphrasing on document retrieval performance

In the next section we describe related research. In section 2, we discuss the observed patterns for query paraphrasing and document retrieval by mentioning the researched work till date. In Section 3, we discuss the resources used by our mechanism. The paraphrase generation and document retrieval processes are described. Section 4 presents sample paraphrases, followed by our evaluation and concluding remarks.

2 RELATED RESEARCH

The main objective of IR systems is retrieving documents from a large space of Information. These systems majorly depend on calculating the similarity between a search query and documents, and retrieve a list of documents that are arranged in descending order of similarity. But since queries are usually short sentences and the generic similarity scoring between a sentence and document seems to fall short of the full intent of the user. In such cases using a differently worded query might work. Addressing the problem of vocabulary mismatch is essential for such short queries and important for effective information retrieval[5]. The main reasons behind the vocabulary mismatch problem are word synonymy and polysemy. When synonym words and word inflections are used in the documents, the system fails in retrieving related documents, resulting in many false negatives and a decline in the overall recall. When the user query contains polysemous words, the exact keyword matching retrieves many unrelated documents (false positives) and implies a decrease in precision [2].[2] mentions their are usually two techniques that can be used to alter query depending on their length : Query refinement techniques for long queries and Query expansion methods for short queries, which insert a few related terms and expand the original query. Query Expansion (QE) bridges the vocabulary gap between the relevant documents

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

and the user's query, by adding relevant terms to the user query and increasing the likelihood of a correct match. From the early works on query expansion, the source from which expansion terms are selected defines the different approaches:local context, global context and Knowledge-based methods. However, query expansion methods can suffer from introducing non-relevant information when expanding the query.Query paraphrasing is a well-researched domain and has been a point of research for quite some time now. In Query Paraphrasing we want reformulate the query such that it makes sense rather adding new words to the query. [8] have used WordNet to propose synonyms for the words in a query. However, they apply heuristics to select which words to paraphrase.

3 MODEL

We have shown a comparison of the two models. In the first model, the query was reformulated by adding terms to the original query based on the topic models. And the second model made use of PARROT - a pre-trained SENTENCE BERT paraphrasing model to give reformulated queries. Both the reformulated and paraphrased queries were then used to retrieve a ranked list of documents via BM25. The retrieved documents were then re-ranked using a BERT model trained on the new query-document list. We evaluated our model on two datasets: MS-MARCO and TREC-COVID.

3.1 Dataset

We made use of two datasets to evaluate our model. MS MARCO and TREC-COVID are well-known query-document datasets. MS MARCO is an open-domain Bing question-answering dataset whereas TREC-COVID is the domain-specific dataset.

TREC-COVID followed the TREC model for building IR test collections through community evaluations of search systems. The document set used in the challenge is the COVID-19 Open Research Dataset (CORD-19). This is a collection of biomedical literature articles that is updated regularly. TREC-COVID uses the document set provided by CORD-19. CORD-19 consists of new publications and preprints on the subject of COVID-19, as well as relevant historical research on coronaviruses, including SARS and MERS. As of May 1, 2020, CORD-19 consists of 60K papers, of which full text is available for 48K. Out of which we used 20K documents.[6, 7]

MS-MARCO dataset contains queries, passages and relevance labels for query-passage pairs. We are using the The relevance labels are derived from what passages were marked as having the answer in the Question-Answering dataset. There are 8.8 million passages used for ranking the passages based on their relevance to a given query. MS MARCO Passage Ranking is a large dataset to train models for information retrieval. It consists of about 500k real search queries from the Bing search engine with the relevant text passage that answers the query. Our goal is to rank them based on their relevance.

We train the network as a binary label task, where every [query, passage] pair has a label 0 if it is irrelevant, or 1 if it is relevant. We use a positive-to-negative ratio: 4. That is, for 1 positive sample (label 1) we include 4 negative samples (label 0) in our training setup. For the negative samples, we use the triplets provided by MS Marco that specify (query, positive sample, negative sample).

[3]

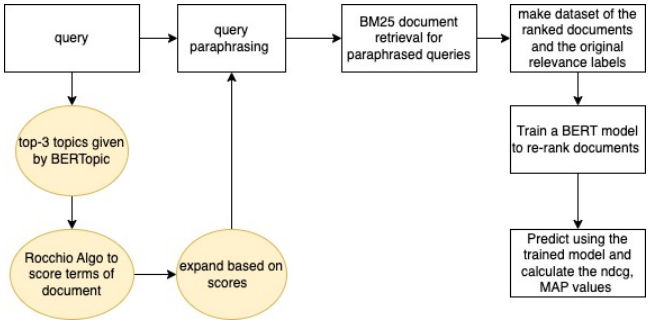

Figure 1: Paraphrasing using topic-modeling and re-ranking via trained BERT model

3.2 Model 1

In our first model, we investigate if topic models can further improve document retrieval performance. Semantic similarity detection is a fundamental task in natural language understanding, thus adding topic information has been useful for semantic similarity models. There is currently no standard way of using topic information for query reformulation.

We propose a novel topic-informed query reformulation on a BERT-based re-ranking architecture for better retrieval performance. The addition of topics to BERT helps particularly with resolving domain-specific cases. In the case of domain-specific queries, our model implicitly adds words closely related to the topic of the query which helps in giving preference to documents related to the domain of the query.

We used *BERTopic* which is a topic modeling technique that leverages transformers and c-TF-IDF, a class-based TF-IDF procedure, to create dense clusters allowing for easily interpretable topics. BERTopic has quite a number of functions which can be used to quickly access topi-related information. First, *BERTopic* was used for extracting topics from the MS-MARCO and TREC-COVID datasets. Due to limited resources, the topic modeling was done on a subset of 50, 000 passages of the MS-MARCO dataset and 28, 941 documents of the TREC-COVID dataset. The preprocessing was done on the documents including stopwords removal, lower-casing, stemming using Krovetz Stemmer, and removing whitespaces, numbers, and punctuations.

BERTopic provides functionality to find topics most similar to a search term. It creates an embedding for the search terms and compares that with the topic embeddings. The most similar topics are returned along with their similarity values. In our model, the search terms are terms of the query. We used the top 2 topics returned based on their similarity scores. For each of the 2 topics, we extract representative documents per topic using another BERTopic functionality. For query expansion, we derive inspiration from Pseudo Relevance Feedback where query terms are added from the top relevant documents retrieved in the first search. In our model, we use the extracted representative documents in place of the retrieved relevant documents used in Pseudo Relevance Feedback. We use the Rocchio relevance feedback algorithm which is used in Pseudo Relevance Feedback for formulating a query that is closer to both

the query and the topic-based representative documents.

$$Q = ALPHA * Qo + BETA * 1/relDocMag \sum_{Dj \epsilon Dr} Dj$$

In the above formula, Q is the final expanded query, Qo represents the original query and Dr is the set of all representative documents used for query expansion. The closeness of a term to the query and documents is determined by the term frequency in both. The associated weights (ALPHA, BETA) are responsible for shaping the modified vector in a direction closer to the original query or related documents. For our model, we used ALPHA = 1, and BETA = 0.75 based on some standard values that were used in other models. We limited the number of expansion words to 5 as the performance of the system decreased after adding more terms. The Rocchio-based score is calculated for all the terms in the representative documents and then the top few terms are added to the original query.

For training, we first retrieve the top few documents from BM25 for each query including the expanded queries. Since our model has multiple queries for each original query and due to resource constraints we extracted top-10 documents for each query and these documents were then trained on BERT for re-ranking.

All the queries including the expanded queries and original queries are then passed to the second layer of our model. For these reformulated queries the relevance label is set based on the original query. This is because the expanded queries represent the original query and our model should be trained to consider the relevant documents of the original query to be relevant to this new query. For training, we split the dataset into a ratio of 80:20 training and test datasets respectively. Due to limited resources, we could only train a subset of the dataset. For the TREC-COVID dataset, we used - queries and - documents and for MS-MARCO we used 8000 passage-query pairs. In the case of query expansion, query expansion was performed on the queries resulting in 16,000 passage document pairs for our model training. We used the "google/bert_uncased_L-4_H-512_A-8" [4]pre-trained BERT model for training our model and set the max length of our encodings as 128.

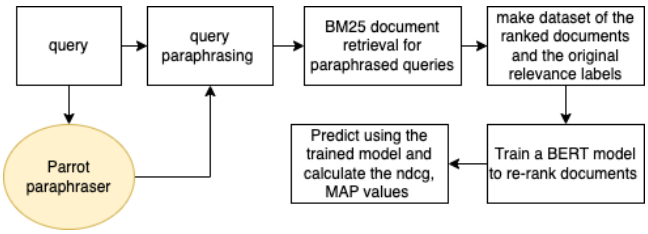

Figure 2: Paraphrasing using Parrot Paraphraser and re-ranking via trained BERT model

3.3 Model 2

In this model, we paraphrased our queries using the PARROT paraphrasing model.A Paraphrase-Generator is built using transformers which takes an English sentence as an input and produces a set of paraphrased sentences. This is an NLP task of conditional text-generation. Parrot is a paraphrase based utterance augmentation framework purpose built to accelerate training NLU models.The paraphrase generation model "parrot-paraphraser-on-T5"[1] has been fine-tuned on some of them MSRP Paraphrase, Google PAWS, ParaNMT, Quora question pairs, SNIPS commands, MSRP Frames. It had given good results on some of the question such as "what drugs have been active against sars-cov or sars-cov-2 in animal studies" was paraphrased as "what drugs are able to kill sars-cov in animal studies?" whereas in some cases the changes were very subtle such as "how does the coronavirus respond to changes in the weather" was paraphrased as "how does coronavirus respond to weather changes?".

Using these paraphrased queries we perform a document retrieval via the BM25 model. BM25 is a ranking function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document. Due to limited resources, we take 30,000 documents for TREC-COVID and 50, 000 passages of the MS-MARCO and produce index of the data for bm25 model to calculate the scores. We retrieve top-20 documents for re-ranking model training.

We combine the document retrieved by BM25 with the original relevance labels for the document for the following query. We do this as the paraphrased queries are similar to original query hence the relevance label will not differ, also we are training the model such that the retrieved document still be relevant to the original query. Similar to model 1 for training, we split the dataset into a ratio of 80:20 training and test datasets respectively. Due to limited resources, we could only train a subset of the dataset. We used the "google/bert_uncased_L-4_H-512_A-8"[4] pre-trained BERT model for training our model and set the max length of our encodings as 128. Transformer-based language models have shown significant improvement in re-ranking tasks.

4 EVALUATION

We scored our paraphrased queries based on the bigrams term matching. We calculated all the bigrams that were present in the documents. Each query's bigrams are collected and compared to the bigrams collected from the document collection. We do this to understand if paraphrasing queries might help us to match more documents. There is a significant improvement in the scoring of query after it is paraphrased.

For evaluation, we used the NDCG metrics and MAP metrics for both models. For baseline, we ran the BM25 for first-stage retrieval and then re-ranked the documents using our existing pre-trained model. We compared these values that we got from baseline, Model 1 and Model 1.

For evalutions metric we made use of MAP, NDCG@10 over the test set that we created while training for BERT. The TREC Eval metric combines a number of information retrieval metrics such as precision and normalized Discounted Cumulative Gain (nDCG). It is used to score rankings of retrieved documents with reference values.

We see from our runs that both the MAP and NDCG@10 scores are highest for Model 1 and lowest for Model 2 while baseline performs in between. Since our dataset is small we cannot be very sure. From what we have got, we can infer from this that paraphrasing

Queries	Model 1	Model 2
what is the current rate for a business loan	what is the current rate for a business loan mortgage refine credit	what's the rate for business loans?
How does the coronavirus differ from seasonal flu?	How does the coronavirus differ from seasonal flu? patients influenza cases	tell me the difference between the coronavirus and seasonal flu?
what part of maine is portland in	what part of maine is portland in county massachusetts hampshire population	which part of maine is portland?

Table 1: Paraphrased Queries with topic-modeling and paraphrase model

	Queries	bi-gram
original	what is the current rate for a business loan	48
Model 1	what is the current rate for a business loan mortgage fine credit	64

Table 2: bi-gram count for queries

Metric	baseline	Model 1	Model 2
MAP	0.434222	0.869092	0.25
NDCG@10	0.434220	0.675278	0.26

Table 3: Model 1

Metric	baseline	Model 1	Model 2
MAP	0.25	0.5255	0.23
NDCG@10	0.26	0.4312	0.53

Table 4: Model 2

using PARROT Sentence BERT is unable to capture the query information hence performs poorly while Model 1 is able to improve the metrics by retrieving documents based on the topic of the query.

5 CONCLUSION

We understood that topic modelling gives better results as it expands the queries with words that are in high density present in the documents. Paraphrasing isn't enough to retrieve new documents and we require more knowledge to extract low ranked documents. We plan to explore more on topic-based modelling in future.

REFERENCES

[1] Prithiviraj Damodaran. 2021. Parrot: Paraphrase generation for NLU.

[2] Jamal Abdul Nasir, Iraklis Varlamis, and Samreen Ishfaq. 2019. A knowledge-based semantic framework for query expansion. *Information Processing Management* 56, 5 (2019), 1605–1617. https://doi.org/10.1016/j.ipm.2019.04.007

[3] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[4] Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7047–7055. https://doi.org/10.18653/v1/2020.acl-main.630

[5] Muhammad Ahsan Raza, Rahmah Mokhtar, Noraziah Ahmad, Maruf Pasha, and Urooj Pasha. 2019. A taxonomy and survey of semantic approaches for query expansion. *IEEE Access* 7 (2019), 17823–17833.

[6] E. Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, W. Hersh, Kyle Lo, Kirk Roberts, I. Soboroff, and Lucy Lu Wang. 2020. TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection. *ArXiv* abs/2005.04474 (2020).

[7] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, K. Funk, Rodney Michael Kinney, Ziyang Liu, W. Merrill, P. Mooney, D. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, B. Stilson, A. Wade, K. Wang, Christopher Wilhelm, Boya Xie, D. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The Covid-19 Open Research Dataset. *ArXiv* (2020).

[8] Ingrid Zukerman and Bhavani Raskutti. 2002. Lexical Query Paraphrasing for Document Retrieval. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1* (Taipei, Taiwan) *(COLING '02)*. Association for Computational Linguistics, USA, 1–7. https://doi.org/10.3115/1072228.1072389