Rethinking Query Expansion for BERT Reranking

Ramith Padaki(✉), Zhuyun Dai, and Jamie Callan

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
{rpadaki,zhuyund,callan}@cs.cmu.edu

Abstract. Recent studies have shown promising results of using BERT for Information Retrieval with its advantages in understanding the text content of documents and queries. Compared to short, keywords queries, higher accuracy of BERT were observed on long, natural language queries, demonstrating BERT's ability in extracting rich information from complex queries. These results show the potential of using query expansion to generate *better* queries for BERT-based rankers. In this work, we explore BERT's sensitivity to the addition of structure and concepts. We find that traditional word-based query expansion is not entirely applicable, and provide insight into methods that produce better experimental results.

Keywords: Neural IR · BERT · Query expansion

1 Introduction

Prevalence of term-matching in several popular search engines has led users to be conditioned to produce short keyword-based queries when attempting to express their information need. A recent study from [2] shows that while applying BERT [4] for reranking, queries written in natural language enable better search results than keywords. The added structure between keywords, which is often ignored in classic retrieval systems, helps BERT better understand the query and leads to higher retrieval accuracy. These allow users to more effectively express their information need and help systems to better disambiguate documents with similar-looking content.

While this is true, there are several reasons why natural language queries may not be available from a user: due to lack of clarity in the information need, due to an inability to phrase the information need as a natural language statement, or quite simply because users are not aware of the capability of natural language understanding in search engines. In such situations, automatically expanding queries is a popularly accepted approach to add new terms and add additional context for the search engine [1]. We hypothesize that while BERT would benefit from expansion, traditional techniques may not be suitable as they do not contain inherent word order, and further lack natural language structure relating them.

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 297–304, 2020.
https://doi.org/10.1007/978-3-030-45442-5_37

This paper explores methods of expanding on an original query in a BERT-based reranker. We distinguish between two means of expanding the original query: first by adding structural words that help create a coherent natural language sentence, and second, by adding additional terms to add new concepts to the original query. We show that neither of the two are individually sufficient, and in-fact a combination of the two benefits reranking with BERT the most.

2 Related Work

Most retrieval models, including many previous state-of-the-art neural ranking models such as [3,8], assumed independence among words in a sentence or only consider short-term dependencies. This caused issues when attempting to disambiguate word meanings and to understand word relations. Recently, large progress has been made on learning contextual word representations using deep neural language models [4,14]. In the domain of neural IR, BERT has been shown to be effective for passage retrieval [12] and document retrieval, specifically when provided long natural language queries for training and evaluation [2].

The addition of contextualization of query-text allows deep neural ranking models to capture several latent traits of languages that were previously difficult to capture. A recent study [2] shows that BERT-based neural rerankers achieved *better* performance on longer queries than on short keywords queries. This makes automatic query expansion desirable. Classic query expansion techniques expand the original query with terms selected from related documents; terms are usually added as a bag-of-words [1,9]. There is prior research that shows classic query expansion to be effective for a few neural ranking models where query terms and document terms are matched softly [10]. However, there is no prior work in exploration of new ways to add pseudo-relevance feedback to BERT-based rankers that rely on free-flowing natural language-text.

3 Discerning the Effect of Structure and Concepts

When people write a short query in the search engine, they often have a longer, more complex question in mind. An example is shown in Table 1: the query-title is the keyword query commonly used during search, and query-description is the real information need. It has been previously shown that applying a BERT reranker using descriptive queries written in natural language provide impressive performance gains over using classic keywords queries [2]. We hypothesize that this is due to a combination of two factors. First, the sentence *structure* in natural language that draws relations between different concepts, and second, the introduction of new *concepts* that are closely tied with the underlying information need. This work aims to understand the effects of these factors through various forms of expanding the original query, as follows.

Expansion with Structure. [2] has shown that non-concept words, such as stop-words, also contribute to BERT's effectiveness by building sentence structure. We therefore utilize two methods to gauge BERT's sensitivity to structure:

Table 1. Original title query, original description query, and variants of expansion for Query 607 in Robust04 dataset.

Title query	Human genetic code
Description query	What progress is being made in the effort to map and sequence the human genetic code
GeneratedStructure	What is the human genetic code
TemplateStructure	What is the human genetic code?
ClassicQEConcepts	Gene dna genome genes research
GoogleQuestions	How long is the human genetic code?
	How many genes are in the human genome 2018?
	Who broke the DNA genetic code?
	What is human code?

GeneratedStructure and TemplateStructure. GeneratedStructure uses a neural machine translation approach to generate new, synthetic questions from the original keyword question. This could have good practical use if proved an effective technique. We adopt the approach proposed in [5]. As shown in Table 1, GeneratedStructure tends to copy the keywords and adds a few question words, adding structure without new concepts.

TemplateStructure tests the maximum possible range of benefit from adding structure to queries. It uses a templating process by hand, manually converting the keyword queries to a question using one of several templates. Queries can be reformulated into "who, what, when, why, how" questions or a request to "describe xyz". These templated questions were generated by the authors, with care being taken to restrict addition of new words other than to relate keyword query words with each other. All original keyword terms are ensured to be included in the reformulation as well. The templated questions provides an upper bound of the effectiveness of expanding with structure.

Expansion with Concepts. This method expands the query with a set of related concepts, while grammar structures are not considered. Our method ClassicQEConcepts leverages RM3 [9], a classic pseudo-relevance feedback based query expansion model, to find related concepts. The expansion terms are concatenated to the original query, ordered by their scores estimated by RM3 as in Table 1.

Combining Structure and Concept Expansion. The last method expands the query with both sentence structure and new concepts. To do so, we rely on scraping Google's suggestions for reformulated queries to acquire additional related questions. If found during scraping, it is used in conjunction with the original title query, else just the title query is used. We refer to this approach as GoogleQuestions.

Qualitative analysis reveals that the suggested questions do not always align with the original query description. To fully verify the power of combined

addition of structure and concepts, we again resort to an oracle to filter the suggested questions manually to ensure that there is description match. We achieve this by manually eliminating questions that do not align with the query description. The annotation was conducted by the authors, with an guideline that allows questions that re-formulate the original question in some other logical structure, thereby being redundant in semantics but not structure. Additionally, questions that were a direct consequence of the original question in concepts were allowed. The manual filtering leads to the FilteredGoogleQuestions.

4 Experimental Setup

Dataset. We use Robust04, a widely-used ad-hoc retrieval benchmark. It contains 249 queries and 0.5M documents. We use two types of queries: *Descriptions*, containing long natural language text describing the information need and *Titles*, the short keyword query text commonly used by search engine users.

Baselines and Experimental Methods. Baselines include three standard bag-of-words retrieval models using the Indri search engine: Indri-LM uses the query language model, Indri-SDM uses the sequential dependency model, and Indri-QE uses the classic query expansion algorithm RM3 [9]. Baselines also include two BERT rerankers from [2]. BERT-Title-Title was trained and evaluated on the query titles, and BERT-Desc-Desc was trained and evaluated on the query descriptions. The authors provided the rankings of all baselines except Indri-QE. For Indri-QE, the parameters were selected through a parameter sweep, including number of feedback documents, number of feedback terms, and weight of the original query. Our experimental methods replace the query titles in the BERT reranker with various types of expanded queries as described in Sect. 3, by concatenating the expansion to the original query title. Document-text remains unchanged.

BERT Reranker and Hyper Parameters. We adopt the model and data setup of the passage-based BERT reranker BERT-MaxP [2]. It splits documents into passages, estimates the relevance between the query and a passage using BERT's two-sentence classification model, and ranks documents using their max passage scores. We fine-tune the model as used in BERT-MaxP [2], using a batch-size of 16 and a learning rate of $1e^{-5}$ for 1000 iterations with 5-fold cross-validation. We train up to a depth of 1000 in the initial retrieved documents and only rerank the top 100 documents from the initial ranking at test time. We also sample 10% of all passages with overlap in addition to the first passage of every document during training to prevent over-fitting, as original proposed [2]. Other existing BERT rerankers use a similar architecture as BERT-MaxP, and apply more complex techniques to improve accuracy, e.g., domain adaptation [2,16], fusion with BM25 [16], and customized pooling layers [11]. This work focuses on using the simpler BERT-MaxP so that the effectiveness of queries is more clear.

Query Expansion Models and Hyperparameters. GeneratedStructure follows prior work [5] that uses CopyNet [7] to translate keywords to questions. We

used the AllenNLP [6] implementation of CopyNet using an embedding dimension of 100 (initialized to GloVe [13] vectors) and an encoder/decoder of size 400 units. The model was trained over a processed Wiki-Answers dataset, containing 3M pairs of questions paired with a synthetically generated keyword question provided by authors of [5]. ClassicQEConcepts used the Indri implementation of RM3 [9]. The query expansions parameters were the same as used in Indri-QE.

5 Experimental Results

This section first studies whether the addition of new concepts has greater benefits during training or evaluation. Then experiments are conducted to find the contributions of different sources of expansion of structure and concepts.

5.1 Concepts and Structure During Training/Evaluation

It is not immediately apparent as to whether the usage of structure and concepts contribute more heavily at train or evaluation. This experiment compares

Table 2. Performance of the BERT reranker [2] on varying training and evaluation sources. Format for reference is BERT-<train>-<eval>.

Method	P@10	P@20	NDCG@10	NDCG@20	MAP
BERT-Desc-Desc	0.552	0.456	0.559	0.524	0.257
BERT-Title-Title	0.486	0.407	0.492	0.467	0.232
BERT-Title-Desc	0.474	0.386	0.479	0.439	0.196
BERT-Desc-Title	0.490	0.408	0.498	0.469	0.233

Table 3. Performance of the BERT reranker [2] when tested on various types of expansions of the original query title. The first 3 methods used Indri's bag-of-words retrieval. The other models used the BERT reranker trained on query descriptions.

Method	P@10	P@20	NDCG@10	NDCG@20	MAP
Indri-LM	0.425	0.358	0.437	0.417	0.211
Indri-SDM	0.432	0.367	0.448	0.427	0.222
Indri-QE	0.439	0.372	0.442	0.427	0.239
Query Title	0.490	0.408	0.498	0.469	0.233
ClassicQEConcepts	0.440	0.376	0.448	0.429	0.216
GeneratedSructure	0.472	0.399	0.486	0.463	0.232
TemplateStrucutre	0.484	0.402	0.489	0.460	0.224
GoogleQuestions	0.488	0.404	0.502	0.471	0.234
FilteredGoogleQuestions	**0.508**	**0.413**	**0.526**	**0.486**	**0.239**
Query description	0.552	0.456	0.559	0.524	0.257

performance when training on query titles and query descriptions, when evaluated on these criteria. Results are presented in Table 2. Format for reference is BERT-<train>-<eval>. For example, BERT-Desc-Title implies that the model was trained on query-descriptions and evaluated on query-titles.

The results confirm findings from [2] that using query-descriptions for training and evaluation benefit the BERT reranker over just using the query-title. More importantly, the gain of using query-descriptions when training (BERT-Desc-Desc) is better than when only using it for evaluation (BERT-Title-Desc). We adopt this approach going forward, training on descriptions while evaluating on a concatenation of the original query with extensions of the original query. Our experiments indicate that this performs better than training on expanded queries.

5.2 Query Expansion with Concepts and Structures for BERT

The next experiment tests the effectiveness of query expansion using structures/concepts in BERT. We expand the query titles with various approaches as described in Sect. 3. Results are in Table 3. Indri-LM/SDM/QE are classic bag-of-words retrieval baselines using Indri; Query Title/Description applies the BERT reranker with query titles/descriptions during evaluation. A good query expansion should be able to out-perform Query Title, and be close to Query Description.

Adding Structure. Results from GeneratedStructure and TemplateStructure reveal that neither of the two strategies out-perform the original query title. This leads us to conclude that structure alone does not provide much evidence for short queries, as short queries do not contain too many complex relations between them; structure is more important with addition of additional keywords, where they help to build the complex relations between the many concepts.

Adding New Concepts. Table 3 reveals that adding expansion terms to Indri produces results that are better than Indri-LM and Indri-SDM, but significantly worse than most BERT rerankers. On training ClassicQEConcepts, with the classical Indri-QE expansion terms, we find that the model under-performs all other BERT models. This indicates that classic query expansion using discrete terms is not suitable for BERT-based deep language models. From Table 1, classic QE expand the query with 'gene dna genome genes research', which is a bag of discrete words and lacks natural language structure that BERT is trained on [4]. This experiment helps establish that the new keywords and concepts must be related to each other in some coherent form, and that these relations actually benefit the ranking process.

Combining Structure and Concepts. FilteredGoogleQuestions provides significant lift in precision and NDCG when compared with the competing baseline Query Title. A win-loss analysis reveals that Google queries improve the original query by providing new concepts as well as several reformulations of the original query. Often, few but meaningful follow-up questions are more useful than several unrelated ones. Further, rephrasing the original sentence in multiple ways

benefits the reranker. For instance for the query "opening adoption records", the questions "Are adoption records public?", and "Should adoption records be open?" give a boost of +0.3 in NDCG@5. This behaviour has been previously studied in previous work [15] wherein BERT favors text sequence pairs that are close in semantic meaning. Often, synonyms and multiple paraphrased versions of the original intent benefit the reranker. On the other hand, without filtering, the GoogleQuestions model does not produce any improvement in performance. This is mainly due to the off-topic questions, which take up about 60% of all retrieved queries. Our results reveal promising direction of query expansion for BERT using related questions that people often searched together. We show that these natural questions, when on-topic, provide valuable information to the original keyword queries, and are more effective than class query expansions or solely adding structures.

6 Conclusion and Future Work

BERT has shown to be good at long descriptive queries in document reranking tasks. With a new paradigm in which deep contextual representations of text show promise in the field of text retrieval, we provide insight into means of emulating descriptive queries after experimental analysis of the behaviour of BERT.

Our results reveal traditional word-based query expansion are not sufficient. A good query for BERT-based rerankers requires both a rich set of concepts and grammar structures that build word relations. However, a critical aspect is identifying extensions of the original query that are in-domain to the corpus, and align with the original intent. Further work in this field would involve automatic identification of questions that are in-domain to the source corpus and alternate means of generating the same.

References

1. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Comput. Surv. (CSUR) **44**, 1–50 (2012)
2. Dai, Z., Callan, J.: Deeper text understanding for IR with contextual neural language modeling. In: The 42nd International ACM SIGIR Conference on Research & Development in Information Retrieval (2019)
3. Dehghani, M., Zamani, H., Severyn, A., Kamps, J., Croft, W.B.: Neural ranking models with weak supervision. In: The 40th International ACM SIGIR Conference on Research & Development in Information Retrieval (2017)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Ding, H., Balog, K.: Generating synthetic data for neural keyword-to-question models. In: Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval. ACM (2018)

6. Gardner, M., et al.: AllenNLP: a deep semantic natural language processing platform (2017)
7. Gu, J., Lu, Z., Li, H., Li, V.O.K.: Incorporating copying mechanism in sequence-to-sequence learning. arXiv preprint arXiv:1603.06393 (2016)
8. Guo, J., Fan, Y., Ai, Q., Croft, W.B.: A deep relevance matching model for ad-hoc retrieval. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (2016)
9. Lavrenko, V., Croft, W.B.: Relevance based language models. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research & Development in Information Retrieval (2001)
10. Li, C., et al.: NPRF: a neural pseudo relevance feedback framework for ad-hoc information retrieval. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2018)
11. MacAvaney, S., Yates, A., Cohan, A., Goharian, N.: CEDR: contextualized embeddings for document ranking. In: Proceedings of the 42nd International ACM SIGIR Conference on Research & Development in Information Retrieval (2019)
12. Nogueira, R., Cho, K.: Passage re-ranking with BERT. arXiv preprint arXiv:1901.04085 (2019)
13. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014)
14. Peters, M.E., et al.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
15. Qiao, Y., Xiong, C., Liu, Z., Liu, Z.: Understanding the behaviors of BERT in ranking. arXiv preprint arXiv:1904.07531 (2019)
16. Yilmaz, Z.A., Wang, S., Yang, W., Zhang, H., Lin, J.: Applying BERT to document retrieval with birch. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations (2019)