University of Waterloo E-Thesis Template for LaTeX

by

Zeynep Akkalyoncu Yilmaz

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2019

Abstract

Standard bag-of-words term-matching techniques in document retrieval fail to exploit rich semantic information embedded in the document texts. One promising recent trend in facilitating context-aware semantic matching has been the development of massively pre-trained language models, culminating in BERT as its most popular example today. In this work, we propose adapting BERT as a neural reranker for document retrieval with large improvements on news articles. Two fundamental issues arise in applying BERT to "ad hoc" document retrieval on newswire collections: relevance judgements in existing test collections are provided only at the document level, and documents often exceed the length that BERT was designed to handle. To overcome these challenges, we compute and aggregate sentence-level relevance scores to rank documents. We solve the problem of lack of appropriate relevance judgements by leveraging sentence-level and passage-level relevance judgements available in collections from other domains to capture cross-domain notions of relevance. We demonstrate that models of relevance can be transferred across domains. By leveraging semantic cues learned across various domains, we propose a model that achieves state-of-the-art results across three standard TREC newswire collections. We explore the effects of cross-domain relevance transfer, and trade-offs between using document and sentence scores for document ranking. We also present an end-to-end document retrieval system that incorporates the open-source Anserini information retrieval toolkit, discussing the related technical challenges and design decisions.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Document retrieval refers to the task of generating a ranking of documents from a large corpus D in response to a query Q. In a typical document retrieval pipeline, an inverted index is constructed in advance from the collection, which often comprises unstructured text documents, for fast access during retrieval. When the user issues a query, the query representation is matched against the index, computing a similarity score for each document. The top most relevant documents based on their closeness to the query are returned to the user in order of relevance. This procedure may be followed by a subsequent re-ranking stage where the candidate documents outputted by the previous step are further re-ranked in a way that maximizes some retrieval metric such as average precision (AP).

Document retrieval systems traditionally rely on term-matching techniques, such as BM25, to judge the relevance of documents in a corpus. More specifically, the more common terms a document shares with the query, the more relevant it is considered. As a result, these systems may fail to detect documents that do not contain exact query terms, but are nonetheless relevant. For example, consider a document that expresses relevant information in a way that cannot be resolved without external semantic analysis. Figure 1 displays

---

Query: international art crime
Text: The thieves demand a ransom of $2.2 million for the works and return one of them.

Figure 1.1: An example of a query-text pair from the TREC Robust04 collection where a relevant piece of text does not contain direct query matches.

one such query-text pair where words semantically close to the query need to be identified to establish relevance. This "vocabulary mismatch" problem represents a long-standing challenge in information retrieval. To put its significance into context, Zhao et al. [11] show in their paper on term necessity prediction that, statistically, the average query terms do not appear in as many as 30% of relevant documents in TREC 3 to 8 "ad hoc" retrieval datasets.

Clearly, the classic exact matching approach to document retrieval neglects to exploit rich semantic information embedded in the document texts. To overcome this shortcoming, a number of models such as Latent Semantic Analysis [2], which map both queries and documents into high-dimensional vectors, and measure closeness between the two based on vector similarity, has been proposed. This innovation has enabled semantic matching to improve document retrieval by extracting useful semantic signals. With the advent of neural networks, it has become possible to learn better distributed representations of words that capture more fine-grained semantic and syntatic information [5], [7]. More recently, massively unsupervised language models that learn context-specific semantic information from copious amounts of data have changed the tide in NLP research (e.g: ELMo [8], GPT-2 [9]). These models can be applied to various downstream tasks with minimal task-specific fine-tuning, highlighting the power of transfer learning from large pre-trained models. Arguably the most popular example of these deep language representation models is the Bidirectional Encoder Representations from Transformers (BERT) [3]. BERT has achieved state-of-the-art results across a broad range of NLP tasks from question answering to machine translation.

While BERT has enjoyed widespread adoption across the NLP community, its application in information retrieval research has been limited in comparison. Guo et al. [4] suggest that the lackluster success of deep neural networks in information retrieval may be owing to the fact that they often do not properly address crucial characteristics of the "ad hoc" document retrieval task. Specifically, the relevance matching problem in information retrieval and semantic matching problem in natural language processing are fundamentally different in that the former depends heavily on exact matching signals, query term importance and diverse matching requirements. In other words, it is crucial to strike a good balance between exact and semantic matching in document retrieval. For this reason, we employ both document scores based on term-matching and semantic relevance scores to determine the relevance of documents.

In this thesis, we extend the work of Yang et al. [10] by presenting a novel way to apply BERT to "ad hoc" document retrieval on long documents – particularly, newswire articles – with significant improvements. Following Nogueira et al. [6], we adapt BERT for binary relevance classification over text to capture notions of relevance. We then deploy

the BERT-based re-ranker as part of a multi-stage architecture where an initial list of candidate documents is retrieved with a standard bag-of-words term matching technique. The BERT model is used to compute a relevance score for each constituent sentence, and the candidate documents are re-ranked by combining sentence scores with the original document score.

We emphasize that applying BERT to document retrieval on newswire documents is not trivial due to two main challenges: Firist of all, BERT has a maximum input length of 512 tokens, which is insufficient to accommodate the overall length of most news articles. To put this into perspective, a typical TREC Robust04 document has a median length of 679 tokens, and in fact, 66% of all documents are longer than 512 tokens. Secondly, most collections provide relevance judgements only at the document level. Therefore, we only know what documents are relevant for a given query, but not the specific spans within the document. To further aggravate this issue, a document is considered relevant as long as some part of it is relevant, and most of the document often has nothing to do with the query.

We address the abovementioned challenges by proposing two effective innovations: First, instead of relying solely on document-level relevance judgements, we aggregate sentence-level evidence to rank documents. As mentioned before, since standard newswire collections lack sentence level judgements to facilitate this approach, we instead explore leveraging sentence-level or passage-level judgements already available in collections in other domains, such as tweets and reading comprehension. To this end, we fine-tune BERT models on these out-of-domain collections to learn models of relevance. Surprisingly, we demonstrate that models of relevance can indeed be successfully transferred across domains. It is important to note that the representational power of neural networks come at the cost of challenges in interpretability. For this reason, we dedicate a portion of this thesis to error analysis experiments in an attempt to qualify and better understand the cross-domain transfer effects. We also elaborate on our engineering efforts to ensure reproducibility and replicability, and the technical challenges involved in bridging the worlds of natural language processing and information retrieval from a software engineering perspective.

1.1 Contributions

The main contributions of this thesis can be summarized as follows:

- We present two innovations to successfully apply BERT to *ad hoc* document retrieval with large improvements: integrating sentence-level evidence to address the fact that BERT cannot process long spans posed by newswire documents, and exploiting cross-domain models of relevance for collections without sentence- or passage-level annotations.

- We explore through various error analysis experiments the effects of cross-domain relevance transfer with BERT as well as the contributions of BM25 and sentence scores to the final document ranking.

- With the proposed model, we establish state-of-the-art effectiveness on three standard TREC newswire collections at the time of writing. Our results on Robust04 exceed the previous highest known score of 0.3686 [1] with a non-neural method based on ensembles, which has stood unchallenged for ten years.

- We release an end-to-end pipeline that applies BERT to document retrieval over large document collections via integration with the open-source Anserini information retrieval toolkit. We elaborate on the technical challenges in the integration of NLP and IR capabilities, and the rationale behind design decisions.

1.2 Thesis Organization

Add link to actual chapters

The remainder of this thesis is organized in the following order: Chapter 2 reviews related work in neural document retrieval and transfer learning, particularly applications of BERT to document retrieval. Chapter 3 motivates the approach with some background information on the task, and introduces the datasets used for both training and evaluation as well as metrics. Chapter 4 proposes an end-to-end pipeline for document retrieval with BERT by elaborating on the design decisions and challenges. Chapter 5 describes the experimental setup, and presents the results on three newswire collections – Robust04, Core17 and Core18. Chapter 6 concludes the thesis by summarizing the contributions and discussing future work.

References

[1] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 758–759, New York, NY, USA, 2009. ACM.

[2] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019.

[4] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 55–64. ACM, 2016.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[6] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *arXiv:1901.04085*, 2019.

[7] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[8] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[9] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners.

[10] Wei Yang, Haotian Zhang, and Jimmy Lin. Simple applications of bert for ad hoc document retrieval. *arXiv preprint arXiv:1903.10972*, 2019.

[11] Le Zhao and Jamie Callan. Term necessity prediction. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 259–268. ACM, 2010.