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, 2018

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Bruce Bruce Professor, Dept. of Philosophy of Zoology, University of Wallamaloo
Supervisor(s):	Doris Johnson Professor, Dept. of Zoology, University of Waterloo Andrea Anaconda Professor Emeritus, Dept. of Zoology, University of Waterloo
Internal Member:	Pamela Python Professor, Dept. of Zoology, University of Waterloo
Internal-External Member:	Deepa Thotta Professor, Dept. of Philosophy, University of Waterloo
Other Member(s):	Leeping Fang Professor, Dept. of Fine Art, University of Waterloo

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Standard bag-of-words term-matching techniques in document retrieval fail to exploit rich semantic information embedded in the document texts. We propose adapting BERT as a neural reranker for document retrieval with large improvements on news articles. We are faced with two fundamental challenges: 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, and can be directly used for ranking news articles. By leveraging semantic cues learned across various domains, we propose a model that achieves state-of-the-art results across three standard 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 technical challenges involved in the integration of NLP and IR capabilities.

Acknowledgements

I would like to thank all the little people who made this thesis possible.

Dedication

This is dedicated to the one I love.

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 the exact query terms, but are nonetheless relevant. For example, consider a document that expresses relevant information in a way that cannot be resolved without the use of external semantic tools.

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.

Figure 1 displays 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. [4] show that the average query terms may not appear in as many as 40% of relevant documents in TREC "ad hoc" retrieval datasets.

Clearly, the classic exact matching approach to document retrieval neglects to exploit rich semantic information embedded in the documents. To overcome this shortcoming, a number of models such as Latent Semantic Analysis [1] that maps both queries and documents into distributed representations has been proposed. This innovation has enabled semantic matching to aid in document retrieval by extracting useful semantic matching signals. With the advent of neural networks, neural language models have been quickly adopted to learn better distributed representations of text. How much better? Why? Moreover, deep neural models have eliminated the need to manually engineer natural language features. Therefore, deep neural networks have since largely replaced the earlier models based on manual decomposition of document matrices. Examples...

One recent innovation that has changed the tide in NLP research has been massively pre-trained language models with its most popular example today being Bidirectional Encoder Representation Transformers (BERT) [2]. BERT has achieved state-of-the-art results across a wide 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. [3] suggest that the lackluster success of deep neural networks in information retrieval may be owing to the fact that crucial characteristics of the "ad hoc" document retrieval task are not properly addressed. Specifically, they emphasize that 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, neural models are usually involved in multi-stage architectures where a list of candidate documents are retrieved with a standard bag-of-words term-matching technique as described above. The documents in this list are then rescored and reranked by the neural model. Some notable examples include...

In this thesis, we present a novel way to apply BERT to "ad hoc" document retrieval on long documents – particularly, newswire articles. A BERT reranker is deployed as part of an end-to-end document retrieval pipeline with significant improvements on standard TREC newswire collections. Specifically, we adapt BERT for binary relevance classification over text to capture notions of relevance. One more sentence to describe? We point out that applying BERT to document retrieval on newswire documents is not trivial due to two principal challenges. Firist of all, BERT has a maximum input length of 512 tokens, which is insufficient to accommodate the entirety 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 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 the challenges encountered in the implementation of such an end-to-end retrieval pipeline in an attempt to bridge 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 on 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. neural or otherwise

- 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, along with the design rationale behind our approach to tightly-coupled integration between Python to support neural networks and the Java Virtual Machine to support document retrieval using the open-source Lucene search library. something about demo, TREC DL...

1.2 Thesis Organization

The remainder of this thesis is organized in the following order: add link to actual chapters Chapter 2 reviews related work in neural document retrieval, 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. What about TREC DL? MS MARCO? 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.

Chapter 2

Related Work

Chapter 3

Cross-Domain Relevance Transfer with BERT

Chapter 4

Experimental Results

Chapter 5

Conclusion

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

APPENDICES