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

...

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 **viii**

List of Figures **ix**

1 Introduction **1**

1.1 Contributions 3

1.2 Thesis Organization 3

2 Related Work **4**

3 Cross-Domain Relevance Transfer with BERT **5**

4 Experimental Results **6**

5 Conclusion **7**

References **8**

APPENDICES **9**

List of Tables

List of Figures

Chapter 1

Introduction

Document retrieval refers to the task of generating a ranking of typically unstructured text documents from a potentially large corpus D in response to a user query Q. The query representation is matched against a previously produced inverted index of the collection to determine candidate documents and compute a similarity score for each document. The retrieved documents are returned in order of relevance based on 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, these techniques favor documents that share the most common terms with the query. As a result, document retrieval systems may fail to detect documents that do not contain the exact query terms, but are nonetheless directly relevant. For example, ... Add figure?

The classic approach to document retrieval clearly neglects to exploit rich semantic information embedded in the document texts. This shortcoming has been the subject of NLP researchers for some time now: Examples...

With the advent of deep learning, numerous neural models that extract semantic matching signals to aid in document retrieval have also been proposed to by-pass the need to manually engineer natural language features. These 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...

Despite growing interest in neural networks, some researchers have recently voiced concern as to whether their use has truly contributed to progress in the field of information retrieval citation, at least in the absence of large amounts of behavioral log data only available to search companies. ... Take stuff from their neural hype paper

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) [1]. 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. Why?

This thesis presents a novel way to successfully apply BERT to yield large improvements in "ad hoc" document retrieval on newswire articles. Applying BERT to document retrieval on newswire collections requires solving two fundamental challenges. First, relevance judgements are provided only at the document level in most collections. That is, given a query, we only know what documents are relevant, not which specific spans within the documents are. ... Second, most newswire documents exceed the length that BERT was designed to handle.

To address these challenges, we propose aggregating sentence-level evidence to rank documents instead of relying solely on document-level relevance judgements. We overcome the lack of sentence level judgements by leveraging sentence- or passage-level judgements available in document collections in other domains to fine-tune BERT models. Surprisingly, we demonstrate that models of relevance can be successfully transferred across domains.

Reproducibility It is important to note that the representational power of neural networks come at the cost of challenges in interpretability. For this reason, ...

Integration challenges This system also highlights the need to bridge the worlds of natural language processing and information retrieval from a software engineering perspective in order to benefit from both. On one hand, most deep learning toolkits today, including TensorFlow and PyTorch, are written in Python with a C++ backend. On the other hand, the open-source search library that facilitates document retrieval in our system as well as many others, Lucene, is implemented in Java, and hence runs on the Java Virtual Machine (JVM). Therefore, the integration between Python and the JVM presents a technical challenge to be addressed.

TREC DL?

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] 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.

[2] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts, 1994.

[3] Donald Knuth. *The TeXbook*. Addison-Wesley, Reading, Massachusetts, 1986.

[4] Leslie Lamport. *LaTeX — A Document Preparation System*. Addison-Wesley, Reading, Massachusetts, second edition, 1994.

APPENDICES