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. [13] 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 [6], [8]. 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 [9], GPT-2 [10]). 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. [12] 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. [7], 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. 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 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. Elaborate more?
- We release an end-to-end pipeline, Birch[1], that applies BERT to document retrieval over large document collections via integration with the open-source Anserini information retrieval toolkit. An accompanying Docker image is also included to ensure that anyone can easily deploy and test our system. 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.

[1] https://github.com/castorini/birch

Chapter 2

Background and Related Work

2.1 Document Retrieval

2.1.1 Non-Neural Methods

2.1.2 Okapi BM25

Binary and its extension to BM25... shortcomings semantic

Okapi BM25 (commonty dubbed BM25) is a bag-of-words ranking function that was developed to accommodate documents of variable lengths. BM25 ranks documents based on the occurrence of query terms in each document, paying more attention to the rarer terms in the query. The goal of this approach is to take into account term frequency and document frequency while estimating the relevance of a document for a given query without introducing too many additional parameters. (Sparck) To achieve this goal, BM25 implementations define two parameters for term frequency saturation and field-length normalization, respectively. Tuned for most common datasets?

2.1.3 RM3

RM3 is a pseudo-relevance feedback mechanism where the original query is expanded by adding terms found in the contents of relevant BM25 documents. Why is it effective, why does it improve performance, how common is it

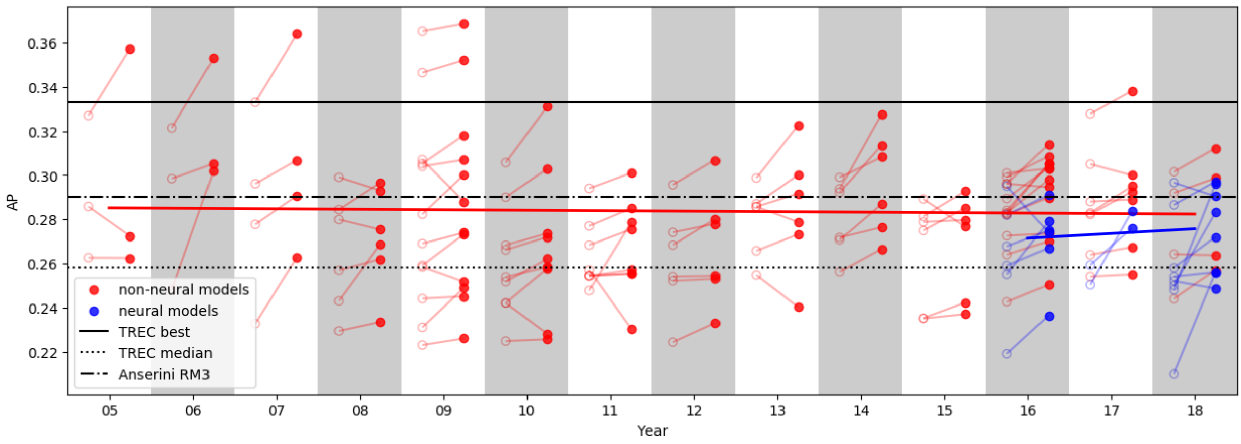

Figure 2.1: .l.

Previous approaches and results on these datasets

To be clear, our focus is on neural ranking models for *ad hoc* document retrieval, over corpora comprising news articles. Formally, in response to a user query Q, the system's task is to produce a ranking of documents from a corpus that maximizes some ranking metric—in our case, average precision (AP). We emphasize that this problem is quite different from web search, where there is no doubt that large amounts of behavioral log data, along with other signals such as the webgraph, have led to large improvements in search quality [**?**]. Instead, we are interested in limited data conditions—what can be achieved with modest resources outside web search engine companies such as Google and Microsoft who have large annotation teams—and hence we only consider "classic" TREC newswire test collections.

2.1.4 Neural Methods

The dominant approach to *ad hoc* document retrieval using neural networks today is to deploy the neural model as a reranker over an initial list of candidate documents retrieved using a standard bag-of-words term-matching technique. Despite the plethora of neural models that have been proposed for document ranking, there has recently been some skepticism about whether they have truly advanced the state of the art [**?**], at least in the absence of large amounts of behavioral log data only available to search engine companies.

In a meta-analysis of over 100 papers that report results on the dataset from the Robust Track at TREC 2004 (Robust04), [11] found that most neural approaches do not

compare against competitive baselines. To provide two recent examples, McDonald report a best AP score of 0.272 and Li 0.290, compared to a simple bag-of-words query expansion baseline that achieves 0.299. Further experiments by [11] achieve 0.315 under more rigorous experimental conditions with a neural ranking model, but this is still pretty far from the best-known score of 0.3686 on this dataset [1].

Although Sculley remind us that the goal of science is not *wins*, but knowledge, the latter requires first establishing strong baselines that accurately quantify proposed contributions. Comparisons to weak baselines that inflate the merits of an approach are not new problems in information retrieval. ...

Having placed evaluation on more solid footing with respect to well-tuned baselines by building on previous work, this paper examines how we might make neural approaches "work" for document retrieval. One promising recent innovation is models that exploit massive pre-training ..., leading to BERT [3] as the most popular example today. Researchers have applied BERT to a broad range of NLP tasks with impressive gains: most relevant to our document ranking task, these include BERTserini for question answering and [7] for passage reranking.

2.2 Semantic Matching

Latent semantic models

Unsupervised language models

BERT

2.3 BERT

With the increasing availability of large corpora, pretrained deep language models have been rapidly gaining traction among NLP researchers. Recent work in NLP has demonstrated that language model pretraining has proven extremely effective for many natural language processing tasks ranging from machine translation to reading comprehension. One of the latest and most sophisticated pretrained deep language models is undoubtedly the Bidirectional Encoder Representations from Transformers (BERT), which has already enjoyed widespread popularity across the NLP community. [3] Unlike previous language models, such as OpenAI's Generative Pretrained Transformer (GPT) [10], BERT produces

deep bidirectional representations by conditioning on both left and right context in all layers by employing a new pretraining objective called "masked language model" (MLM). Conceptually, MLM randomly masks some of the tokens from the input with the goal of predicting the original token based only on its left and right context.

As expected, optimizing this objective requires a very complex model: for example, the larger BERT model requires around 340 million parameters be optimized. In fact, training this model end-to-end takes four days to complete even on 16 high-end tensor processing units (TPUs) [3]. Fortunately, there exists a technique to benefit from these models without having to train an entire model from scratch. The most versatile and widely adopted approach to applying these neural models to downstream NLP tasks is based on "freezing" their last layer, and "fine-tuning" on external data for the specific task. Not only does this approach introduce only a few task-specific parameters to optimize, but it also greatly boosts the performance of many NLP tasks given the rich semantic

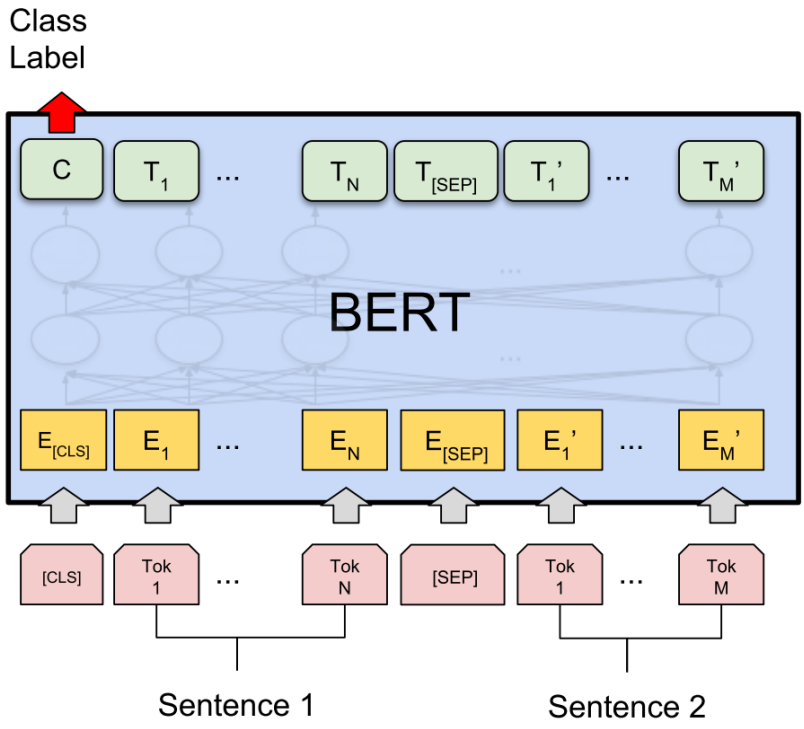

Figure 2.2: BERT Sentence Pair Classification Model.

expressiveness introduced by the pretrained language model. Figure 2.2 visualizes the input and output for fine-tuning BERT for a "sentence pair classification" model. To form an input, two sequences of tokens are concatenated with the meta-token $[SEP]$, i.e: separator, and prepended with $[CLS]$, corresponding to the "class" meta-token. A single-layer neural network is added to the end of this network with the class label as the input, and subsequently trained for the specific downstream task.

Success in other tasks

BERT document retrieval

BERTserini [**?**] combines the open-source information retrieval toolkit Anserini[1] with a BERT-based reader to identify answers from a large corpus of Wikipedia articles in an end-to-end fashion. With this pipeline, the authors are able to achieve large improvements over previous question answering systems. Although this work aims to solve an entirely different task, we incorporate a similar architecture in this project with success.

On the other hand, Nogueira et al. [7] propose to re-rank MS MARCO passages based on a simple re-implementation of BERT, outperforming the previous state of the art by 27% in MRR@10 and replacing the top entry in the leaderboard of the MS MARCO passage retrieval task at the time of publication. Our neural model is inspired by the BERT re-implementation described in this paper.

2.4 Evaluation Metrics

The standard approach to evaluation in information retrieval relies on the distinction between "relevant" and "irrelevant" documents with respect to an information need as expressed by a query. A number of automatic evaluation metrics has been formalized for ranking tasks such as document retrieval. These metrics rely on The size of most document collections makes it infeasible for humans to manually judge the relevance of all documents. All relevant documents need to be labelled to prevent false negatives, i.e: treating documents which are in fact relevant as irrelevant.

2.4.1 Mean Average Precision (MAP)

Precision specifies what fraction of a set of retrieved documents is in fact relevant for a given query q. Precision can easily be extended to evaluate ranked retrieval results by ...

[1] https://github.com/castorini/anserini

Average precision (AP) expresses the average of the precision values obtained for the set of top k documents for a query. Support that $D = \{d_1, ..., d_{m_j}\}$ is the set of all relevant documents for a query q_j, then AP can be formulated as:

$$AP = \frac{1}{m_j} \sum_{k=1}^{m_j} P(R_{jk}) \tag{2.1}$$

where R_{jk} represents the set of top k tanked retrieval results.

The respective AP for each query can be aggregated to obtain mean average precision (MAP) for the overall retrieval effectiveness in the form of a single-figure measure of quality across various recall levels:

$$MAP = \frac{\sum_{j=1}^{|Q|} AP}{Q} = \frac{1}{Q} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} P(R_{jk}) \tag{2.2}$$

It has been show to have especially good discrimination and stability compared to other metrics, which makes it the ideal choice for large text collections [5]. It is hence one of the standard metrics among the TREC community.

2.4.2 Precision at k (P@k)

Unlike MAP which factors in precision at all recall levels, certain applications have a distinctly different notion for ranking quality. Particularly in the case of web search, the user often only cares about the results on the first page or two, but not all of them. This restriction essentially leads to measuring precision at fixed low levels of retrieved results, i.e: top k documents – hence the name for metric "precision at k". On the one hand, it eliminates the need for any estimate of the size of the set of relevant documents. However, it also produces the least stable out of all measures. Moreover, precision at k does not average well because the total number of relevant documents for a query has a very strong influence on its value.

2.4.3 Normalized Discounted Cumulative Gain (NDCG@20)

NDCG is uniquely useful in applications with a non-binary notion of relevance, e.g: a spectrum of relevance. Like precision at k, it is evaluated as a weighted sum over the top k search results, and normalized so that a perfect ranking yields NDCG equals 1.

math stuff

NDCG is a popular choice for systems with machine learning approaches.

2.5 Datasets

Add statistics and examples

Elaborate on splits

2.5.1 Fine-Tuning

As discussed in Section 1, applying BERT to document retrieval requires leveraging passage- or sentence-level relevance judgements fortuitously available in large text collections. Since no such newswire collection currently exists, we train the BERT relevance classifier on three out-of-domain collections.

TREC Microblog datasets draw from the Microblog Tracks at TREC from 2011 to 2014, with topics (i.e., queries) and relevance judgments over tweets. We use the dataset prepared by Rao et al. (2019) MS MARCO features user queries sampled from Bing's search logs and passages extracted from web documents. Each query is associated with sparse relevance judgments by human editors. TREC CAR uses queries and paragraphs extracted from English Wikipedia: each query is formed by concatenating an article title and a section heading, and passages in that section are considered relevant. This makes CAR, essentially, a synthetic dataset.

2.5.2 Evaluation

We conduct end-to-end document ranking experiments on three TREC newswire collections: the Robust Track from 2004 (Robust04) and the Common Core Tracks from 2017 and 2018 (Core17 and Core18). Robust04 comprises 250 topics, with relevance judgments on a collection of 500K documents (TREC Disks 4 and 5). Core17 and Core18 have only 50 topics each; the former uses 1.8M articles from the New York Times Annotated Corpus while the latter uses around 600K articles from the TREC Washington Post Corpus.

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] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

[6] 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.

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

[8] 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.

[9] 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.

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

[11] Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. Critically examining the "neural hype": Weak baselines and the additivity of effectiveness gains from neural ranking models. In *Proceedings of the 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*, pages 1129–1132, Paris, France, 2019.

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

[13] 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.