RADBOUD UNIVERSITY

MSC THESIS

Re-Ranking BERT; Revisiting Passage Re-Ranking with BERT on MS MARCO

Author:
Tom JANSSEN GROESBEEK

Supervisor:
Prof. Dr. Ir. A.P. DE VRIES

A thesis submitted in fulfillment of the requirements for the degree of Master of Data Science

in the

Research Group Data Science
Department of Computer Science

August 27, 2020

Declaration of Authorship

I, Tom JANSSEN GROESBEEK, declare that this thesis titled, "Re-Ranking BERT; Revisiting Passage Re-Ranking with BERT on MS MARCO" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Tom Janssen Groesbeek

Date: August 27, 2020

RADBOUD UNIVERSITY

Abstract

Faculty of Science
Department of Computer Science

Master of Data Science

**Re-Ranking BERT;
Revisiting Passage Re-Ranking with BERT on MS MARCO**

by Tom JANSSEN GROESBEEK

In this thesis, the task of passage ranking using the MS MARCO passage ranking dataset is examined. Given an input query and candidate passages retrieved by the baseline ranker BM25 model, the task is to re-rank these passages by relevance to the query. Currently, the relevance labels provided by the dataset are assumed to be incomplete as not every query-passage pair is assessed on relevancy. The hypothesis of this work is that there are more relevant passages per query and an online assessment is organized in order to gather these additional labels. With these new relevance labels, the performances of the BM25 ranker and the state-of-the-art BERT model on the passage ranking task are re-examined. Both models show increased performances on the passage ranking task when they are evaluated with the additional relevance labels. While originally BERT outperforms the BM25 ranker, utilizing the new relevance labels shows that BM25 achieves equal and in some settings even better performance. Additionally, both models are evaluated using multigraded relevance labels and the results of this evaluation show that they perform equally well in ranking multigraded passages, but that BERT does not improve the rankings of BM25.

Contents

List of Figures

List of Tables

List of Abbreviations

AP	**A**verage **P**recision
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
BM25	**B**est **M**atch 25
CG	**C**umulative **G**ain
DCG	**D**iscounted **C**umulative **G**ain
DNN	**D**eep **N**eural **N**etworks
ERR	**E**xpected **R**eciprocal **R**ank
IR	**I**nformation **R**etrieval
LSTM	**L**ong **S**hort-**T**erm **M**emory
LTR	**L**earning-**T**o-**R**ank
MFR	**M**ean **F**irst **R**elevant
ML	**M**achine **L**earning
MRC	**M**achine **R**eading **C**omprehension
MRR	**M**ean **R**eciprocal **R**ank
MS MARCO	**M**icro**S**oft **MA**chine **R**eading **CO**mprehension
NDCG	**N**ormalized **D**iscounted **C**umulative **G**ain
P	**P**recision
QA	**Q**uestion **A**nswering
RBP	**R**ank **B**iased **P**recision
RR	**R**eciprocal **R**ank

Chapter 1

Introduction

1.1 Asking Questions

Search engines like Google and Bing have made it easier to search the internet in the pursuit of fulfilling our information need. Google receives over 63,000 searchers per second on any given day, which is roughly 2 trillion searches per year. 15% of all searches have never been searched before on Google and an average person conducts 3 to 4 searches every single day[1]. The monthly search volume of Bing (worldwide) is 12 billion searches [2]. We like to search the internet.

There are also many different search engines to choose from nowadays. Well known engines like Google, Bing and Yahoo!, but also ones like DuckDuckGo and Ecosia which differentiate themselves from the "big" ones by promoting privacy or by planting trees. Yet, with so many competitors, Google dominates the search engine market with a market share of over 90% in 2020[3]. A core believe of the company is that *there is always more information to be found*[4], which is why their researchers are continuously busy improving their products and services.

One of the latest innovations by Google is a technique for natural language processing pre-training, based on neural networks, called Bidirectional Encoder Representations from Transformers (BERT). It is open-source with the purpose to allow anyone to train question answering systems[5]. The company claims it helps to improve their own question answering system by applying the BERT model to both ranking and relevant passage extraction in their search system. Doing so makes their system understand the language used in queries better.

A recent article titled *How cutting-edge AI is helping scientists tackle COVID-19* emphasizes what important role good functioning question answer systems fulfill in these times[6]. Currently numerous research papers on the COVID-19 virus are published on a daily basis. All this prior work is important for researchers working on a cure or vaccine. This is where question answer systems with good language comprehension come into play. By training machines to comprehend user questions and make them able to filter through the entire corpus of COVID-19 publications. Articles could be ranked and answer snippets with summaries relevant to the user question could be retrieved. Eventually, answering questions on the spot and accelerating work on a cure or vaccine.

[1] https://seotribunal.com/blog/google-stats-and-facts/ (Last accessed: 2/7/2020)

[2] https://www.statista.com/topics/4294/bing/ (Last accessed: 2/7/2020)

[3] https://gs.statcounter.com/search-engine-market-share#monthly-201907-202007-bar (Last accessed 16/7/2020)

[4] https://www.google.com/about/philosophy.html (Last accessed: 16/7/2020)

[5] https://www.blog.google/products/search/search-language-understanding-bert/ (Last accessed 3/7/2020)

[6] https://www.weforum.org/agenda/2020/06/this-is-how-ai-can-help-us-fight-covid-19/ (Last accessed: 3/7/2020)

1.2 Task Definition

This study re-evaluates the performance of BERT on the task of passage ranking formulated by MS MARCO[7]. The task was based on the passages and questions from the Question Answering Dataset[8]. Passages marked as having the answer in the dataset helped to derive relevance labels for the passage ranking task, which makes it one of the largest relevance datasets. It was constructed in order to facilitate the benchmarking of Machine Learning (ML) based retrieval models. In specific those models that benefit from supervised training. It has also been the focus of the 2019 and 2020 TREC Deep Learning Track[9].

Four different tasks were proposed alongside the passage ranking dataset:

1. Passage Re-Ranking: Given a candidate top 1000 passages as retrieved by BM25, re-rank passages by relevance.

2. Passage Full Ranking: Given a corpus of 8.8m passages generate a candidate top 1000 passages sorted by relevance.

3. Document Re-Ranking: Given a candidate top 1000 documents as retrieved by BM25, re-rank documents by relevance.

4. Document Full Ranking: Given a corpus of 3.2m documents generate a candidate top 1000 documents sorted by relevance.

The focus of this study is the *Passage Re-Ranking* task, but a slightly modified version of the task. Because of computational reasons the current study works with a candidate top 100 of passages as retrieved by BM25. More details on the research approach are provided in Section 1.4.

1.3 Background and Related Work

This section provides the theoretical background related to the current study. The domain under which this study falls is Machine Reading Comprehension (MRC) and in specific Information Retrieval (IR). Part of this domain is the task of ranking candidate answer sources from a large and diverse set of documents. Recent years, several of such datasets emerged providing a solid base for the development and testing of complex passage and document ranking models. Some of which were developed with the focus on progressing the field of Question Answering (QA). The current study takes upon the task to extend the IR domain by enhancing the MS MARCO passage ranking dataset and in the progress of doing so re-evaluating the performance of the BERT model on the passage ranking task.

1.3.1 Question Answering

The domain of Machine Reading Comprehension focuses on learning machines to read and understand a text like we humans do (Zhang et al., 2019). Part of which is to teach a machine to read and understand questions and let it provide the correct answer, an important facet of information retrieval often termed Question Answering. QA systems combine natural language processing and information retrieval

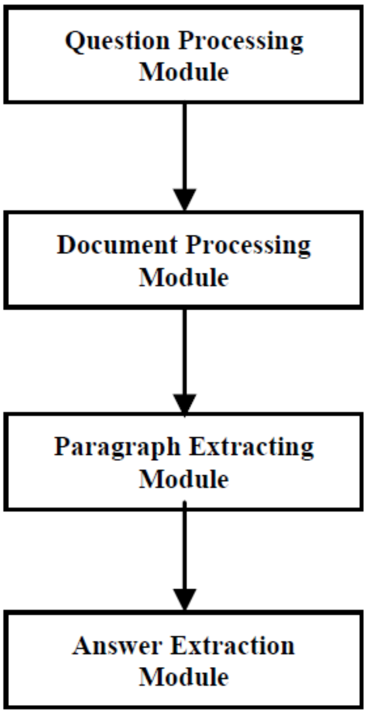
[7] https://microsoft.github.io/msmarco/ (Last Acccessed: 30/6/2020)

[8] https://microsoft.github.io/msmarco/#qna (Last Acccessed: 30/6/2020)

[9] https://trec.nist.gov/ (Last Acccessed: 30/6/2020)

techniques and often solve different subtasks to go from understanding the query to providing a well-formed answer. Figure 1.3.1, from the work by Pundge, Khillare, and Mahender, 2016, depicts a common framework for QA systems.

Figure 1.1: Framework of a QA System. Source: Pundge, Khillare, and Mahender, 2016.


Question Processing Module
Document Processing Module
Paragraph Extracting Module
Answer Extraction Module

Such systems start off by processing the question in the *Question Processing Module*. This could be as simple as directly relaying the user's input or more complex in systems were more preprocessing is done. Think of query expansion(Gauch, Wang, and Rachakonda, 1999), answer type detection (Prager et al., 2008, Li and Roth, 2002), or converting the query to word embeddings (Ye et al., 2016). Aim of this module is to classify and analyse the question in order to restrict the candidate information sources in subsequent modules and to provide support in narrowing down which answer extraction method to apply in the final module (Buscaldi et al., 2010).

Next, the *Document Processing Module* is tasked to generate a list of potential informative documents. The system now tries to scan the used corpus for documents that might contain an answer to the question. Part of this module is a text search engine that can handle large datasets and computes some sort of relevance score between the question and the content of the documents in order to rank the returned documents on relevance. Like Lucene[10], a text search engine library that uses a standard *TF-IDF* model (Term Frequency - Inverse Document Frequency) to perform ranked retrieval (Tellex et al., 2003).

The returned documents often contain more information than necessary to be able to construct an answer. The *Paragraph Extraction Module* helps to extract the part or parts of the document that are relevant to the question. First, smaller parts of the documents are scored on relevancy after which a ranked list of passages is

[10] `jakarta.apache.org/lucene/docs/index.html` (Last Accessed: 2/7/2020)

returned. The goal is to search for relevant but compact text snippets in relation to the input query, instead of returning entire documents (Cui et al., 2005), which in turn enables more efficient answer extraction. Document and passage retrieval face the challenge of determining what content is relevant to the query. Terms used in the query could be unrelated to the final answer and thus applying any term-based search method could cause the system to miss truly relevant documents or passages. Additionally, a passage extraction system has to determine the optimal size of the passages. Small passages could lack the needed information for the final answer, while large passages might still contain irrelevant information. In the end, all pieces of information gathered by the first three modules determine the performance of the final module (Buscaldi et al., 2010).

The final part of any QA system, the *Answer Extraction Module*, gets as input the retrieved candidate texts or text-snippets and is tasked with retrieving terms or exact phrases to form an answer to the question. This module may use different Natural Language Processing (NLP) techniques like Named Entity Recognition (NER) in order to accomplish this (Lee et al., 2006). Even this final module could perform some sort of classification and ranking as, e.g., entities are extracted from candidate passages, classified on their type and then ranked on how well their type matches the query (Abney, Collins, and Singhal, 2000).

1.3.2 Document and Passage retrieval

The overall performance of a QA system depends heavily on the effectiveness of its intermediate modules. Even though the process of document retrieval is very different from extracting and composing an answer, it plays a crucial role. If this module fails to retrieve any relevant documents to the question, the entire system fails to return an answer fulfilling the information need of the user (Hu, 2006). Over the years, since the first question answering systems such as Baseball (Green Jr et al., 1961), numerous studies have been dedicated to tackling QA and the related challenge of candidate passage retrieval (i.e., MacAvaney, Yates, and Hui, 2017, Xiong et al., 2017, Wang et al., 2017, Frermann, 2019).

One of the driving forces behind these studies has been the Text REtrieval Conference (TREC) that was started in 1992 as part of the TIPSTER Text program. Co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defence, the purpose of the conference was to support Information Retrieval (IR) research by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies[11]. During each new TREC[12] a test set of documents and questions is provided by NIST and participants run their retrieval systems on the data (Voorhees and Tice, 2000). They then submit a list of retrieved top-ranked documents. Individual results are then pooled by NIST, retrieved documents are judged on correctness and results are evaluated. The TREC ends with workshops were participants can share experiences.

Research in this field has caused an surge of QA datasets to emerge, such as NewsQA (Trischler et al., 2016), DuReader (He et al., 2017), NarrativeQA (Kočiský et al., 2018), SearchQA (Dunn et al., 2017), QUSAR (Dhingra, Mazaitis, and Cohen, 2017) and SQuAD (Rajpurkar et al., 2016). While there are numerous datasets, they differ a lot. Datasets can be open-domain or closed-domain datasets. If a dataset is closed-domain, it means that the content of that dataset is domain-specific and can help to train QA models to answer questions related to that domain but not beyond

[11] https://trec.nist.gov/overview.html (Last Accessed: 2/7/2020)
[12] https://trec.nist.gov/data/qamain.html (Last Accessed: 2/7/2020)

that. The Baseball system was highly domain-specific as it could only answer questions about US baseball players[13]. In open-domain datasets, on the other hand, no predefined domain knowledge is expected and a system should be capable to scan different sources of text documents in order to generate an answer (Qi et al., 2019).

Another distinction is made by Ingale and Singh, 2019, who state that there are two types of datasets for machine reading comprehension and question answering. The first type are *Datasets with Extractive Answers* which originate from Cloze style queries (Taylor, 1953). Such style queries are short texts with a blank part that needs to be filled. What must be filled is an appropriate token that is based on the reading and understanding of a relevant document (Ghaeini et al., 2018). Extractive datasets contain a large amount of documents or passages, and questions for which the answers are direct segments of corresponding passages. A system tested on those datasets should select a correct text span from the given context, which in itself is objectively gradable as systems are not intended to generate answers themselves. The SQuAD and NewsQA datasets are examples of extractive datasets.

The other type of datasets are *Descriptive* or *Narative Answer Datasets*. In contrast to extractive datasets, answers are not exact text spans from candidate documents or passages. The answers in these kind of datasets are more fluent and stand-alone sentences. Here the task is often to produce more answer like responses instead of simply retrieving specific extracts from documents. Example datasets are NarrativeQA and MS MARCO.

1.3.3 MS MARCO

MS MARCO, which stands for MAchine Reading COmprehension, is a large scale real-world dataset focused on machine reading comprehension, question answering, passage ranking, keyphrase extraction and conversational search studies (Bajaj et al., 2016). The dataset was created by sampling and anonymizing Bing and Cortana usage logs. It contains search queries and a set of extracted passages from documents retrieved by Bing in response to the question. Human editors were asked to construct answers based on the contents of the retrieved passages. In addition, the editors were asked to mark the passages that they used to construct these answers. But editors did not have to ensure that all relevant passages were annotated. The passages and documents could very well lack information necessary to answer the questions. If editors could not answer a question with the provided passages, they could annotate the question as unanswerable. These questions were kept in the dataset as it was believed that it was important for the development of an MRC model to recognize unanswerable questions because of insufficient available information. Further details on the contents of the dataset are provided in Chapter 2.

The reason behind the creation and publication of the MS MARCO dataset was that the authors (Bajaj et al., 2016) wanted to address shortcomings of existing MRC and QA datasets. According to them, previous datasets are not large enough in order to train deep neural models with large numbers of parameters. Even if there are large MRC datasets available, these are often synthetic or the questions are constructed by crowd workers based on passages or documents provided to them. Which would mean that the questions in those datasets do not represent a "natural" distribution of the information need that users may want to satisfy. The authors argue that the MS MARCO questions are more representative of natural information needs as these are gathered from actual search queries submitted by users to Bing.

[13] http://ai.stanford.edu/blog/answering-complex-questions/ (Last Accessed: 2/7/2020)

Moreover, text submitted by users in the real-world are often messy. Input can include typos, abbreviations, just a couple of terms or full of spelling and grammar mistakes. Retrieved passages or documents can contain conflicting information. In contrast, datasets preceding MS MARCO often contain high-quality stories or text spans. Instead, MRC systems should be tested on realistic datasets to ensure that they are robust to noisy and problematic inputs.

Aside from the dataset alone, the authors also propose three different machine learning tasks that each differ in level of difficulty. The first task, the novice task, formulates that a system should predict if an question is answerable and if so then generate the correct answer. Otherwise it should clarify that no answer is present. The second task, the intermediate task, is an extension of the first task as it expects that the system generates a well-formed answer. This means that if the answer is read-aloud it should make sense without the context of the question and retrieved passages. The final task is the *passage re-ranking* task in which a system is provided with a question and a set of 1000 retrieved passages using the BM25 model (Robertson et al., 1995). It is now the task for the system to re-rank these passages based on how relevant the content is in order to answer the question.

1.3.4 BERT

As was the intention of the MS MARCO team, the dataset is large enough to be able to train deep neural networks (DNN). Many researchers have taken this opportunity to fine-tune and test established DNNs on the corpus and achieve improved performances on the passage ranking task. For example the kernel based neural model for document ranking by Xiong et al., 2017 (rank 81[14]) or the bi-LSTM network with co-attention mechanism between query and passage representation by Alaparthi, 2019 (rank 61[15]), both performing better than the baseline which was settled by the BM25 ranking function (rank 86[16]). This function is based on the classical probabilistic model of information retrieval (Robertson and Zaragoza, 2009). This model assumes that probabilities of relevance for query-document pairs is estimated and that documents are ranked in descending order of these probabilities. BM25 contains components of the TF-IDF weighting scheme, in which query terms found in a document are scored according to their frequency in that document, reducing the impact of terms that are frequent across all the documents in the corpus. This way, documents with unique terms score better (Zhai and Massung, 2016).

Lately, another state-of-the-art DNN has been performing well on the passage ranking task. The model named BERT, which stands for Bidirectional Encoder Representations from Transformers, was initially developed to pre-train word representations (Devlin et al., 2018). The authors behind the BERT model argued that existing techniques to pre-training language representations at that time had the major limitation of being unidirectional which would limit the choice of architectures to be used during pre-training. In the unidirectional case, an architecture could either be left-to-right or right-to-left, which means that tokens are processed in either one of these directions and the current token can only attend to the previous tokens processed. This is sub-optimal for tasks, such as question answering, where context from both directions can come in handy. They proposed the bidirectional BERT model in order to improve fine-tuning based approaches. Already in the first

[14] Last checked: 2/7/2020
[15] Last checked: 2/7/2020
[16] Last checked: 2/7/2020

publication on the model, it achieved outstanding results on 11 natural language processing tasks evaluated on different QA datasets such as SQuAD.

Since then many researchers have fine-tuned and extended the BERT model on the MS MARCO dataset and one by one have achieved improved performances. Currently, BERT dominates the MS MARCO passage ranking leaderboard, as many BERT based submissions outperform the BM25 baseline by a large increase in performance. These include the work by Nogueira and Cho, 2019 who re-implemented BERT to be able to re-rank query-based passages (rank 31[17]) and by Han et al., 2020 who encode queries and passages using BERT and combine it with a learning-to-rank (LTR) model constructed with TF-Ranking in order to further improve ranking performances (rank 19[18]).

1.4 Research Approach

In the original paper (Bajaj et al., 2016), introducing the MS MARCO dataset, the authors clearly explain how the questions, passages and query-passage pair relevance labels are gathered. As stated in Section 1.3.3, the questions in the dataset are a set of user question queries sampled from Bing's search logs. The passages are extracted from web documents retrieved by the Bing retrieval system. Both seemingly sound methods.

However, the chosen method for gathering relevancy labels is open to criticism. As described by the authors, human editors were tasked to annotate relevant passages. For every question they were only shown 10 passages retrieved from relevant web documents by the Bing retrieval system. They were asked to mark the passages they used to construct an answer to the question with *is_selected = 1*. If a passage was not used to construct an answer (as the answer or relevant information was not present), they should mark it with *is_selected = 0*. For every question the authors then decided on the relevant passages by filtering on this annotation.

The fact that the editors only got to go over 10 passages retrieved by Bing and not the entire collection of passages or documents in the dataset is a significant shortcoming of this dataset. Also, the annotations only specify which passages were used by the editors in constructing an answer, but editors were not obliged to mark all relevant passages. While we can assume that the marked passages are indeed relevant, we cannot assume that unmarked passages are irrelevant. A fact noted by the authors themselves: "As the editors were not required to annotate every passage that were retrieved for the question, this annotation should be considered as incomplete—*i.e.*, there are likely passages in the collection that contain the answer to a question but have not been annotated as is_selected: 1." This begs the question whether the MS MARCO relevance labels might be incomplete in such a way that it affects evaluation. Therefore, the main research question of this thesis is formulated as follows:

RQ1 Does the MS MARCO passage ranking dataset contain more relevant passages per query than currently labelled, such that evaluation is affected?

For some time now, models based on the BERT architecture have dominated the leaderboard of the MS MARCO passage re-ranking task[19]. From the now 87 submissions, 62% utilizes the BERT architecture and 5 out of 10 submissions making

[17] Last checked: 15/7/2020

[18] Last checked: 15/7/2020

[19] https://microsoft.github.io/msmarco/ (Last Accessed: 30/6/2020)

it to the top 10 have used BERT. BERT is currently even topping the leaderboard. In conjunction with the research on the completeness of the MS MARCO passage ranking dataset, this thesis will re-evaluate the performance of BERT on this natural language challenge to re-assess the apparent dominance of the model.

A similar approach is taken as that by Padigela, Zamani, and Croft, 2019 and Crijns, 2019, as the performance of BERT will be compared to the baseline BM25. This work discriminates itself from previous work in the fact that it is hypothesized that the current MS MARCO dataset is incomplete and thus an online assessment will be performed in order to gather more relevancy labels. Both models will then be compared on their performance on the original dataset and the enhanced version. The following sub-questions are addressed:

RQ1.1 How does BM25 perform on the MS MARCO passage ranking task when multiple relevant passages per query are provided?

RQ1.2 How does BERT perform on the MS MARCO passage ranking task when multiple relevant passages per query are provided?

RQ1.3 Does re-ranking with BERT improve initial rankings by BM25 on the MS MARCO passage ranking task?

Furthermore, the current MS MARCO dataset, like many other QA datasets, contains binary relevance labels (either relevant or irrelevant). In the process of gathering new relevancy labels for query-passage pairs it was decided to gather graded labels, with the motivation to enable research on the performance of BERT in comparison to the BM25 baseline in case of multi-label data. Which is why the following final research question is addressed in this thesis:

RQ2 What is the effect of graded relevance judgements on the relative performance of re-ranking MS MARCO passages with BERT?

1.5 Report Structure

This report is subdivided into 6 chapters. The current section is part of Chapter 1 which features the introduction of the study. Here the problem domain and motivation are explained and related previous research is summarized to put current work into perspective. This chapter also introduces the research questions and the research approach taken to answer those questions. The next two chapters provide the necessary information to understand how the study was executed. Chapter 2 will elaborate on the dataset that is used during this study and the approach taken to enhance it, while Chapter 3 provides in-depth details on the experiments run during this study. The chapter starts off with an explanation of the information retrieval task that forms the core of the experiments. Followed by background information and implementation details on the different models that will be compared on their performance on the aforementioned task. The chapter ends with an explanation of the different metrics used to evaluate the performances of both models. In Chapter 4 the results of this study will be presented accompanied by an interpretation of those results. A more elaborate discussion of these results and the methods used can be found in Chapter 5. Any shortcomings of the this study as well as possible future work will also be included in this chapter. Finally, the conclusion of this study can be found in Chapter 6, which entails the answers to the research questions.

Chapter 2

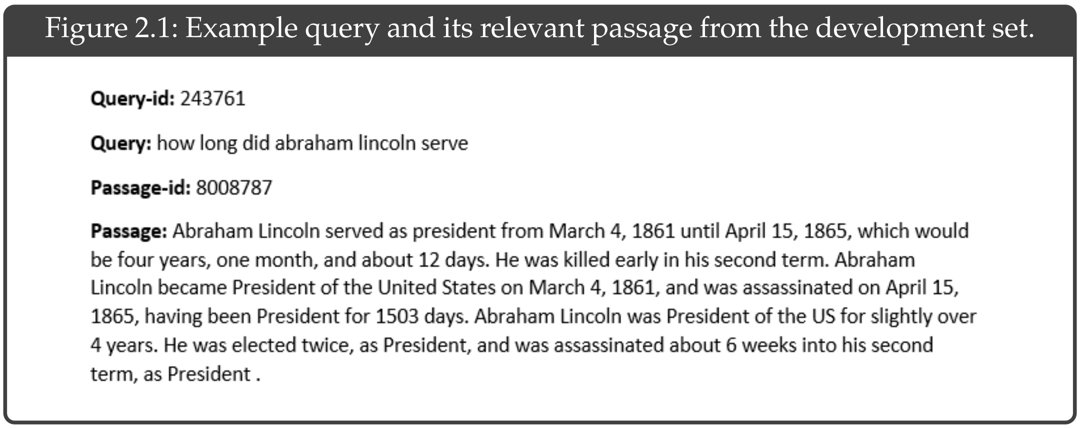
Data

This chapter will elaborate on the data used for the experiments of this study. The following section details the MS MARCO passage ranking dataset and its limitations. In order to enhance the original passage ranking dataset, an online assessment was held. The final section will explain what was the purpose of this assessment, how it was constructed and what assessments were collected. Limitations of the online assessment are discussed in Chapter 5.

2.1 MS MARCO

MS MARCO is a collection of datasets focused on deep learning in search[1]. The dataset used in this study is the MS MARCO passage ranking dataset released on the 26th of October 2018. It contains 1,010,916 queries and 8,841,823 million passages extracted from over 3,563,535 million documents. The task to be evaluated with this dataset is to re-rank a list of candidate passages by relevance to a given query. The dataset thus also contains a set of relevance labels in the form of query and passage id pairs. These pairs specify that the passage is relevant to the paired query.

Figure 2.1: Example query and its relevant passage from the development set.



Query-id: 243761

Query: how long did abraham lincoln serve

Passage-id: 8008787

Passage: Abraham Lincoln served as president from March 4, 1861 until April 15, 1865, which would be four years, one month, and about 12 days. He was killed early in his second term. Abraham Lincoln became President of the United States on March 4, 1861, and was assassinated on April 15, 1865, having been President for 1503 days. Abraham Lincoln was President of the US for slightly over 4 years. He was elected twice, as President, and was assassinated about 6 weeks into his second term, as President .

The dataset is divided into a training, a development and an evaluation set. However, the evaluation set does not contain relevance labels. Therefore the focus of this section will be on the training and development set. Table 2.1 depicts the number of unique query ids in both query subsets and Table 2.2 depicts the number of unique query ids in the relevance label subsets. If a query id is present in the relevance label set, it means that there is at least one relevant passage linked to it. Because the total number of unique training query ids in table 2.2 is less than in table 2.1, it is clear that not all queries in the training set have a corresponding relevant passage. All the queries in the development set have at least one relevant passage.

[1] https://microsoft.github.io/msmarco/ (Last accessed: 25/6/2020)

TABLE 2.1: The number of unique query ids in the training and development query datasets.

Query Subset	# Unique Query ids
training queries	808,731
development queries	6,980

TABLE 2.2: The number of unique query ids in the training and development relevance label datasets.

Relevance Label Subset	# Unique Query ids
training relevance labels	502,939
development relevance labels	6,980

In the current MS MARCO dataset a query can also have multiple relevant passages, but these cases are not that common. Table 2.3 and Table 2.4 show how many queries in, respectively, the training and development set have an **X** number of relevant passages. In the training set, 59% of the queries have only one relevant passage and approximately 38% have zero relevant passages. In the development set, approximately 94% of the queries have only one relevant passage.

TABLE 2.3: This table depicts the number of relevant passages per query in the training subset.

# Queries	% Queries	# Rel Passages per Query	Total Rel Passages
305,792	37.81%	0	0
477,580	59.05%	1	477,580
21,868	2.70%	2	43,736
2,718	0.34%	3	8,154
612	0.08%	4	2,448
131	0.02%	5	655
22	0.00%	6	132
8	0.00%	7	56
808,731	100.00%		532,761

TABLE 2.4: This table depicts the number of relevant passages per query in the development subset.

# Queries	% Queries	# Rel Passages per Query	Total Rel Passages
0	0.00%	0	0
6590	94.41%	1	6590
331	4.74%	2	662
51	0.73%	3	153
8	0.12%	4	32
6980	100.00%		7437

According to Bajaj et al., 2016 the way the set of relevant passages was constructed was by human editors who annotated the passages they used to compose an answer with to the query. A set of on average 10 passages was included with each query. These passages were taken from relevant web documents returned by the passage retrieval system of Bing. If the editors used one of the passages to construct an answer to the query they annotated the passage by setting the *is_selected* parameter to 1. If no answer was constructed from the set of passages for a given query, the entire set of passages should be annotated by setting *is_selected* to 0. Next, to create the list of relevant query and passage identifier pairs, the *is_selected* annotation was used to identify all relevant passages for a given query. However, the editors were not obliged to provide every passage retrieved for a given query with an annotation. The annotations should thus be viewed as incomplete when considering the ranking problem as it is possible that other passages in the collection are relevant to a query but for which the annotation for *is_selected* is not equal to 1.

2.2 Online Assessment

To produce a different dataset, focused on passage ranking evaluation, an online assessment in context of this thesis project was held to collect more query-passage pair relevancy labels. One of the reasons being that Bajaj et al., 2016 state that the original relevancy annotations should be viewed as incomplete. Another reason was to test the hypothesis that, in contrast to the numbers depicted in Table 2.3 and Table 2.4, many queries have more than 1 relevant passage.

In order to verify this hypothesis, it was necessary to gather relevancy assessments from many different assessors. An online assessment tool was created by making use of Google's mobile app development platform named Firebase[2]. Among the different tools offered by the platform, the database and authentication services were used to create an assessment interface. By incorporating authentication via email, assessors could pause their work and continue on a later moment in time, making it possible to gather as many assessments as possible. At the same time the authentication process helped to create user identifiers. These identifiers enabled exact monitoring of the assessment process on individual basis, preventing duplicate assessments by an assessor. Any user input was stored on Google Cloud via Firebase's Firestore. After the current study was finished, the data was collected and removed from the Cloud and any personal information was replaced by anonymized identifiers, to be able to store the data for future work and also to protect the privacy of the assessors.

Not all MS MARCO queries were used during the online assessment. Instead it was decided to only use queries:

1. From the MS MARCO development set.
2. For which the MS MARCO relevant passage was already ranked high enough by the BM25 model.

The reason for the first decision of solely using the development set is that, as part of this study, a BERT model fine-tuned on the passage ranking training set of MS MARCO was used. Excluding the training set would prevent data leakage from the fine-tuning process into the any experiments run for this study. The second decision was taken because for the experiments of this thesis we worked with an initial top 100 ranking by BM25. The current study was only interested in those queries for which the MS MARCO relevant passage was already ranked high enough by the BM25 model for there to be any significant improvement by the BERT model in an additional re-ranking. It would not be of any interest for this thesis to study queries

[2] https://firebase.google.com/ (Last accessed: 25/6/2020)

for which the relevant passage was ranked very low (e.g. around 100) and for BERT to re-rank it just a few ranks higher. Therefore, given this initial top 100 ranking by BM25 the queries were labelled according to the ranking of the MS MARCO relevant passage. The following labels were used[3]:

- **high** if the relevant passage was ranked 1-20
- **medium** if the relevant passage was ranked 21-80
- **low** if the relevant passage was ranked 81-100
- **outside scope** if the relevant passage was ranked >100

The queries either labeled low or outside scope were discarded. There remained 2329 queries labeled high and 897 queries labeled medium from the initial 6980. These numbers were still too large for the scope of this study and it was decided to randomly sample a stratified subset of 600 queries in total (540 high queries and 60 medium queries). These 600 queries were then distributed among the assessors by the online assessment tool. We setup the data collection process to ensure that every query-passage pair would be assessed by three different assessors. So it would be wise to distribute the same query to at least three assessors before distributing another query. But at the same time, as 600 was still a large number, exploration of new queries was important to ensure a diverse enough number of queries would be assessed. This is why queries were distributed in sets of three and initially two of these sets were passed along to a new assessor. One set would always consist out of queries that were already assessed but by less than 3 assessors. The other set would randomly (with a probability of 50%) consist out of either assessed queries or unassessed queries to promote exploration. There was no limit on the number of queries that one could assess. If an assessor would decide to carry on with the work after processing the first six queries, two new sets would be picked containing queries not yet assessed.

Each query was accompanied by the top 20 passages retrieved by BM25. These passages, as well as the queries, were shown to each assessor in randomized order. The assessor could then assess each passage on a scale from 1 (totally irrelevant) to 5 (perfectly relevant). Which was different from the MS MARCO relevant labels as these labels were binary (irrelevant or relevant). Figure 2.2 shows an example of what an user was presented when assessing.

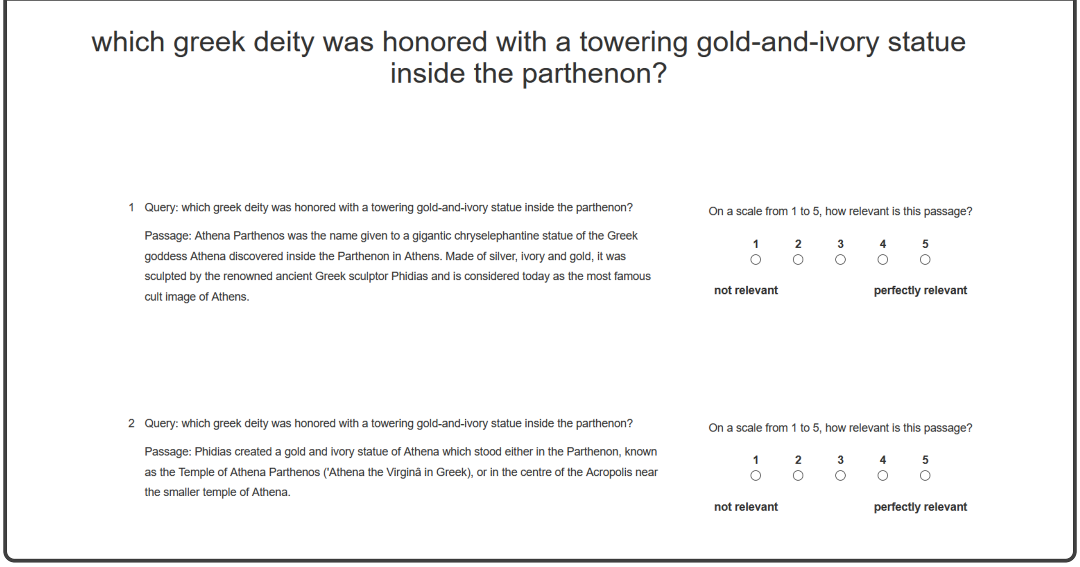
Assessments were gathered for the duration of one month, during which 37 different assessors helped to assess 125 unique queries with corresponding passages. Not all of these queries were used for this study. Two criteria were devised in order to select queries to be used during the experiments. For any query:

- The number of assessors that processed it should be at least three.
- The MS MARCO relevant passage should be judged relevant by the assessors as well.

Any assessed query that did not meet this criteria was discarded from the dataset. The final dataset contained 42 queries which were used for experiments. Detailed statistics on these queries and the new relevance labels can be found in Chapter 4. In order to check if either of these two criteria was met by any of the queries, a filtering system was designed. Algorithm 1 depicts the pseudo-code of this filtering system.

[3] If a query had more than 1 relevant passage, the highest ranked relevant passage was used to decide on the label.

Figure 2.2: Snapshot from the online assessment tool.



which greek deity was honored with a towering gold-and-ivory statue inside the parthenon?

1 Query: which greek deity was honored with a towering gold-and-ivory statue inside the parthenon?

Passage: Athena Parthenos was the name given to a gigantic chryselephantine statue of the Greek goddess Athena discovered inside the Parthenon in Athens. Made of silver, ivory and gold, it was sculpted by the renowned ancient Greek sculptor Phidias and is considered today as the most famous cult image of Athens.

On a scale from 1 to 5, how relevant is this passage?

1 ○ **2** ○ **3** ○ **4** ○ **5** ○

not relevant **perfectly relevant**

2 Query: which greek deity was honored with a towering gold-and-ivory statue inside the parthenon?

Passage: Phidias created a gold and ivory statue of Athena which stood either in the Parthenon, known as the Temple of Athena Parthenos ('Athena the Virginâ in Greek), or in the centre of the Acropolis near the smaller temple of Athena.

On a scale from 1 to 5, how relevant is this passage?

1 ○ **2** ○ **3** ○ **4** ○ **5** ○

not relevant **perfectly relevant**

Algorithm 1: Pseudo code on how the experiment query subset was created.

```
Result: Experiment Query Subset
experiment_query_ids = [];
for query_id in assessment_dataset do
    nr_assessors = get_nr_assessors(query_id);
    if nr_assessors $\geq$ 3 then
        ms_marco_rel_passage_id =
          get_rel_passage(query_id,ms_marco_dataset);
        query_assessments = assessment_dataset[query_id];
        binary_assessments =
          make_binary(query_assessments,binary_threshold);
        assessed_rel_passage_ids = majority_voting(binary_assessments);
        if ms_marco_rel_passage_id in assessed_rel_passage_ids then
            experiment_query_ids.append(query_id);
        end
    end
end
```

First it was checked if a query was assessed by at least three different assessors. If this criterion was met, the original multigraded assessments were transformed to binary. This could be done by making use of a binary threshold. Every passage with a grade below this threshold is labeled irrelevant and every passage with a grade equal or above this threshold is labeled relevant. Taking a low threshold will result in many relevant passages, while a high threshold will result in very few relevant passages. This is why this study takes two different thresholds to explore if the choice of binary threshold affects performances. The binary threshold was set at either 2 (<2 irrelevant, ≥2 relevant; **The Liberal Dataset**) or 3 (<3 irrelevant, ≥3 relevant; **The Strict Dataset**). Higher thresholds were also explored, but resulted in too few relevant passages for queries to meet the criteria of this study.

After the multigraded assessments were transformed to binary, majority voting was applied to decide on the new binary relevance labels. Multigraded relevance labels were also created for the passages, but majority voting was not the optimal method to decide on these labels. Instead, for any query-passage pair, the ceiled median of the assessor grades was taken to decide on the graded relevance label (resulting in **The Graded Dataset**).

In some cases the assessors failed to provide any input leading to missing data. It was often the case that only one assessor forgot to provide an assessment for only one passage. In those cases, the missing data was ignored and only the assessments from the remaining assessors were taken into consideration. If they did not agree on the relevancy, the original MS MARCO label was taken in the binary case or a grade of 1 (irrelevant) was given to that specific query-passage pair in the graded case. There was one query that had seven different assessors process it from which one assessor failed to assess 19 from the 20 passages. Because of the large number of assessors, it was quickly decided to only consider the assessments by the other six assessors.

Chapter 3

Experimental Setup

This chapter explains the setup used to perform the experiments of this thesis. What exact data is used and how this data was selected. How BM25 and BERT have been implemented and what settings were used for ranking passages. Finally, this section will specify which evaluation metrics were used and how they were formulated.

3.1 Passage Ranking

The passage ranking experiments were performed on a subset of the development set of the MS MARCO passage ranking dataset[1]. The development set is divided into two files, namely one query file containing the query ids and query texts and one relevance file containing query id and passage id pairs. There are no specific development passages and so all passages are stored in one big collection file containing passage ids and passage texts. Only those queries in the development set that have a query id in the relevance file are used, resulting in 6980 queries.

The original passage re-ranking task set out by MS MARCO states that systems should re-rank a set of 1000 retrieved passages. This study works with a initial set of 100 retrieved passages which are then re-ranked. One of the reasons is that this is computationally more suitable for the scope of this study. Working with 1000 retrieved passages per query takes more time and because Amazon Web Services was used to run the ranking systems this would become a very costly undertaking. Moreover, this study was only interested in those queries for which the MS MARCO relevant passage was already ranked in the top 100 by BM25. Retrieving a top 1000 passages per query would have been redundant.

The following experiments were performed:

Experiment 1 BM25 old relevance dataset vs. new relevance dataset: The performance of BM25 is measured on the passage ranking task using two different binary relevance label datasets. The original MS MARCO relevance labels (Dataset MS MARCO) are always used and the other labels are constructed using the online assessment input and varying the binary threshold between the values 2 (Dataset Liberal) and 3 (Dataset Strict).

Experiment 2 BERT old relevance dataset vs. new relevance dataset: The performance of BERT is measured on the passage ranking task using two different binary relevance label datasets. The original MS MARCO relevance labels (Dataset MS MARCO) are always used and the other labels are constructed using the online assessment input and varying the binary threshold between the values 2 (Dataset Liberal) and 3 (Dataset Strict).

[1] `https://msmarco.blob.core.windows.net/msmarcoranking/collectionandqueries.tar.gz`

Experiment 3 BM25 vs. BERT (MS MARCO): The performances of BM25 and BERT are measured on the passage ranking task using the original binary MS MARCO relevance labels.

Experiment 4 BM25 vs. BERT (Liberal): The performances of BM25 and BERT are measured on the passage ranking task using newly constructed binary relevance labels. These labels are constructed using the online assessment input and setting the binary threshold at 2.

Experiment 5 BM25 vs. BERT (Strict): The performances of BM25 and BERT are measured on the passage ranking task using newly constructed binary relevance labels. These labels are constructed using the online assessment input and setting the binary threshold at 3.

Experiment 6 BM25 vs. BERT (Graded): The performances of BM25 and BERT are measured on the passage ranking task using newly constructed multi-label relevance labels. These labels are constructed using the online assessment input.

3.2 Implementation Details

This study uses BM25 to retrieve an initial set of passages per query and then lets BERT re-rank these passages. For both models an existing open source implementation[2] is used. The following two subsections will explain what sources are used and with which settings the models were used during experiments.

3.2.1 BM25

For this study the Anserini toolkit was used to retrieve an initial ranking of passages with BM25[3]. Built on Lucene, this open-source information retrieval toolkit aims to narrow down the gap between academic IR research and the practice of building real-world search applications. The main goal of the toolkit is to offer reproducible ranking baselines with clear documentation such that they are easy to use (Yang, Fang, and Lin, 2018).

As initialization of the toolkit for using BM25 with the MS MARCO dataset, the entire passage collection is indexed in Anserini. Then the toolkit is used to retrieve the top 100 passages ranked by their BM25 score. BM25 is used with the default Anserini settings of k1=0.82 and b=0.68, optimized on recall@10 because the purpose of the Anserini BM25 model is to serve as input to re-rank models such as BERT. It should therefore maximize the number of relevant documents retrieved[4].

Before testing the BM25 implementation on the experiments of this study, the model was tested on the development set with 1000 hits. Next, by making use of msmarco_eval and trec_eval the MRR@10, MAP and recall@1000 were computed to evaluate the output from the model. Identical results were achieved as stated by Anserini on their Github page, validating our experimental setup[5].

[2] Exact details on how to get these models up and running can be found on my github: http://tomjanssengroesbeek.nl/Master_Thesis_CoAs_BM25_BERT/.

[3] http://tomjanssengroesbeek.nl/Master_Thesis_CoAs_BM25_BERT/instructions/rurevm_setup/anserini_bm25 (Last accessed: 3/7/2020)

[4] https://github.com/castorini/anserini/blob/master/docs/experiments-msmarco-passage.md (Last accessed: 3/7/2020)

[5] https://github.com/castorini/anserini/blob/master/docs/experiments-msmarco-passage.md (Last accessed: 6/7/2020)

3.2.2 BERT

Previous work by Nogueira and Cho, 2019 has shown that re-implementing BERT for query-based passage re-ranking and then pre-training it on the MS MARCO passage ranking dataset ensures outstanding performance on the passage ranking task. Because the process of pre-training is not within the scope of this study, it was decided to utilize the adapted BERT model made available by Rodrigo Nogueira and Kyungyun Cho[6]. This study makes use of their pre-trained BERT$_{\text{BASE}}$ model.

As stated in the initial paper on BERT by Devlin et al., 2018 the difference between BERT$_{\text{BASE}}$ and BERT$_{\text{LARGE}}$ is their sizes measured in number of layers. Both models contain transformer block layers (L), hidden layers (H) and self-attention heads (A). The BERT$_{\text{LARGE}}$ model has 24 transformer block layers, 1024 hidden layers and 16 self-attention heads with 340 million total parameters. The BERT$_{\text{BASE}}$ model is smaller in size as it has 12 transformer block layers, 768 hidden layers and 12 self-attention heads with 110 million total parameters. Because the BERT$_{\text{BASE}}$ model is smaller in size, it is faster to train and evaluate. Nogueira and Cho pre-trained both models on the MS MARCO passage ranking training set. Fine-tuning of the models was performed with a batch size of 32 (32 sequences * 512 tokens = 16,384 tokens/batch) for 400,000 iterations. This corresponds to training on 12.8 million query-passage pairs, which is roughly less than 2% of the full training set. Further details on the pre-training process are to be found in Nogueira and Cho, 2019.

Their BERT$_{\text{LARGE}}$ is currently ranked 31th on the MS MARCO passage ranking leaderboard[7] with a MRR@10 score of 35.9 on the evaluation set and 36.5 on the development set. Their BERT$_{\text{BASE}}$ scores approximately 2 MRR@10 points lower on the development set with a score of 34.7. Before running their model on the experiments of this study, their model was tested on the development set with 1000 hits. An identical score of 34.7 was achieved[8], confirming that the implementation of their model was without errors.

3.3 Evaluation Metrics

Four different evaluation metrics are used to evaluate the performance of both BM25 and BERT on the passage ranking task. The evaluation metric used for the MS MARCO leaderboard is the Mean Reciprocal Rank (MRR), which is why this score is also computed for this thesis. Aside from the MRR, the Mean First Relevant (MFR) and the Average Precision (AP) are computed for both models. Finally, the Normalized Discounted Cumulative Gain (NDCG) is computed making use of the graded relevance labels. This metric is computed using only the labels obtained from the online assessment as the original dataset did not contain graded relevance labels.

3.3.1 Mean Reciprocal Rank

The Reciprocal Rank (RR) reflects the position or rank of the first relevant document in a ranked list. Equation 3.1 shows how the RR is computed. The lower the first relevant item to a specific query is ranked, thus the higher the denominator, the lower the RR score will be.

[6] https://github.com/nyu-dl/dl4marco-bert (Last accessed: 6/7/2020)

[7] https://microsoft.github.io/msmarco/#leaderboard (Last accessed: 6/7/2020)

[8] http://tomjanssengroesbeek.nl/Master_Thesis_CoAs_BM25_BERT/instructions/bert/ (Last accessed: 6/7/2020)

$$RR_i = \frac{1}{rank_i} \tag{3.1}$$

Where $rank_i$ is the rank of the first relevant item i in the ranked list of retrieved items. To measure a systems performance on retrieving relevant items for multiple queries, the Mean Reciprocal Rank (MRR) can be computed. This is the arithmetic mean of RR scores for each query. Equation 3.2 shows how the MRR is computed.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|} RR_i = \frac{1}{|Q|}\sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{3.2}$$

Where Q stands for the set of queries. This metric is useful when measuring a systems capability of retrieving one relevant item, like in a case where there is just one relevant item and it should be ranked as high as possible. An example of these so-called *known item searches* is when an user is looking for the homepage of a known company (Zhai and Massung, 2016).

The MS MARCO passage ranking challenge uses this metric to rank systems on its leaderboard. In order to be able to compare results, this metric was also computed for the experiments performed during the current study. However, this metric is not completely flawless. Fuhr, 2018 mentions several common mistakes in IR evaluation and states that the use of MRR is one of them. The reason for this is that in order to compute the MRR, one needs to compute the mean of the summed RR scores. However, the RR is an ordinal scale and for these kind of scales it is not valid to compute the mean or standard deviation. Like stated in the work by Stevens et al., 1946, computing the mean of an ordinal scale is an error because the successive intervals on the scale are unequal in size. Means and standard deviations should therefore not be used with these scales. An alternative to the MRR metric is proposed by Fuhr, which is the Mean First Relevant (MFR). This metric regards the rank numbers directly and then computes the arithmetic mean for a set of queries. Equation 3.3 shows how this metric is computed.

$$MFR = \frac{1}{|Q|}\sum_{i=1}^{|Q|} rank_i \tag{3.3}$$

Where Q stands for the set of queries and $rank_i$ is the rank of the first relevant item i. For completeness, this study will also compute the MFR scores for each experiment. Fuhr does not mention how to compute MFR when there are cases when there is no relevant item in the ranked list. For example, if one is only interested in computing the MFR for the top 10 items but none of these items is relevant.

When computing the MRR one deals with these cases by appointing a zero score to them, as the RR is computed by taking the rank of the first relevant item as the denominator. The lower its rank, the more its RR score will go towards zero. But for the MFR it is not logical to take zero as score for cases when there are no relevant items present. The opposite seems more logical, where those cases are appointed a very high score. So for this study, it was decided to assign queries for which there was no relevant item present in the top N ranked items a score of $N + 1$.

3.3.2 Precision

This metric captures how accurate a retrieval system is by measuring how many of the retrieved items are relevant. If a system has 100% precision it means that all retrieved items are relevant (Zhai and Massung, 2016). Equation 3.4 shows how precision is computed.

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \tag{3.4}$$

Here the numerator contains the union of documents that are both retrieved and relevant. The number of retrieved relevant documents is divided by number of all retrieved documents. One limitation of precision is that it does not take into account the rank of the retrieved documents. Take for example two systems that both retrieve 10 items from which 5 are relevant items. System A ranks those 5 items in the top 5, while system B ranks those 5 items in the bottom 5 of the list. In this case both systems will obtain a precision of 0.5. This metric purely measures how good a system performs in retrieving relevant documents versus irrelevant documents.

Another metric often mentioned alongside precision is recall. This metric is used to measure how many of the entire set of relevant items is retrieved by a system. Ideally a system retrieves only all relevant items and has 100% precision and recall. But often high recall is accompanied by low precision. A system with high recall could simply return many items and thus retrieve all relevant but also many irrelevant items.

This study will only compute the precision and ignore recall, because the MS MARCO dataset does not contain all relevance labels. Computing recall will be of no use when not all query-passage pairs are assessed on relevancy. Precision can still provide useful insights as the datasets does contain some relevance labels. It is of interest for this study to research how accurate the systems are in retrieving those relevant items.

3.3.3 Normalized Discounted Cumulative Gain

In case of items with multi-label ratings, the Normalized Discounted Cumulative Gain (NDCG) metric can be used to evaluate systems. The NDCG is build up of several components. Equation 3.5 shows the formula for the Cumulative Gain (CG).

$$CG(L) = \sum_{i=1}^{n} r_i \tag{3.5}$$

Where r_i is the gain of result i and we define gain as the multi-label rating. Since the rating can be translated to how much information an user gains when viewing that item. Let i range from one to n, where n can be set to any specific cutoff. For example, one can compute the CG for a top 10 items with multi-label ratings. If one would examine more than 10 documents, the CG will increase.

The CG does not take into account the rank of the viewed items. This is where the Discounted Cumulative Gain (DCG) comes into play. The DCG weights (discounts) the contribution of gain from different items according to their rank. The general notion behind this weighting is that users will not always continue down a ranked list examining all retrieved items. It does not discount the document at position one, as it is assumed that users will always see this document. The next documents will be discounted as there is a possibility that users will not notice them. Discounting happens by dividing the gain of result i by a weight based on that position. This exact formula is depicted in Equation 3.6.

$$DCG(L) = r_1 + \sum_{i=2}^{n} \frac{r_i}{\log_2 i} \tag{3.6}$$

Where r_i is the gain of result i and r_1 is thus the gain for result one. It is clear from this equation that result one is not discounted while all the subsequent results are. Discounting is performed by dividing by the logarithm of the rank of the item. The DCG formula still needs to be normalized to be able to make this measure comparable across different queries (Zhai and Massung, 2016). Equation 3.7 shows the final formula which is used to compute the NDCG.

$$NDCG(L) = \frac{DCG(L)}{IDCG} \tag{3.7}$$

For any given list L we compute the NDCG by dividing the DCG of list L by its ideal DCG or IDCG. This is the DCG of the ideal form of the list, where the most relevant documents are ranked at the top and sorted in descending order of relevance. By dividing the DCG of a list by its IDCG, it is normalized to obtain a value between 0 and 1.

The current study measures the NDCG of the different models in order to deal with the graded relevance labels that were gathered. To be able to measure the previously mentioned metrics, these gathered relevance labels had to be transformed to binary. Which was performed by considering two different thresholds. For the NDCG, no such transformation had to be performed and the actual assessor labels could be utilized. However, the assessor were only presented the top 20 items retrieved by the BM25 model. While this formed no problems when computing the NDCG for the BM25 model with different cutoff values below 20, problems did occur when computing the NDCG under the same setting for the BERT model; as it was likely that BERT would retrieve a different top 20, containing unassessed items. To deal with this problem any item not assessed was considered irrelevant and provided with a label equal to 1. The exact gains associated with each multigrade label are shown in Table 3.1.

TABLE 3.1: This table depicts what gain is associated with each degree of relevance. If no assessment was made, the gain was assumed to be equal to 1.

Relevance Degree	no assessment	1	2	3	4	5
Gain r_i	1	1	2	3	4	5

Chapter 4

Results

This chapter contains the results of the online assessment that was organized in interest of this study and the subsequent experiments that were performed. The first section will provide details on the data that was collected with the online assessment as well as some details on the assessors that participated. The second section presents the results of all the different experiments performed on the passage ranking task.

4.1 Results Online Assessment

This chapter provide statistics and results based on the 42 queries who met the criteria set out by this study as explained in Chapter 2. This section in specific provides details on the datasets used during the experiments.

Figure 4.1 shows three different boxplots. Each belongs to a different relevance label dataset. Either the original dataset (MS MARCO), the dataset with binary threshold set at 2 (Liberal) or the dataset with binary threshold set at 3 (Strict). The MS MARCO boxplot shows that the original dataset contained only one relevant passage for every experiment query. In contrast, the online assessment helped to gather many more relevant passages which can be seen by the boxplots of the Liberal and Strict datasets.

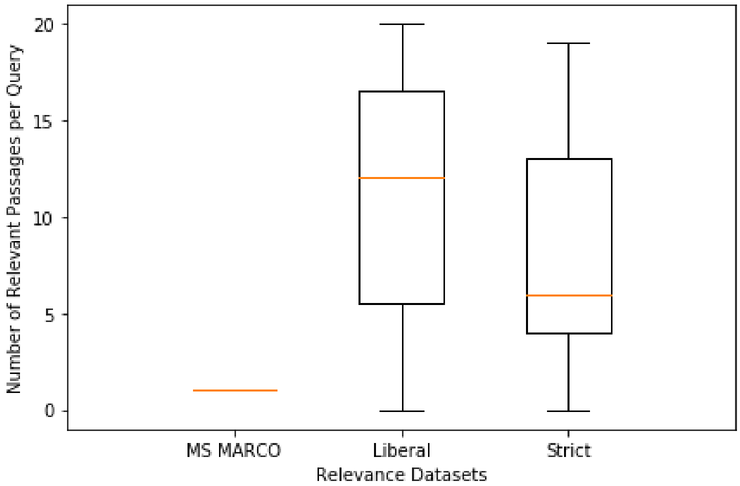


FIGURE 4.1: Distribution relevant passages for the experiment queries for each dataset.[1]

The original assessor input were graded relevance labels. Figure 4.2 shows the relevancy grade boxplots for each individual experiment query. This figure shows great variation across the different queries. For quite some queries, passages were often not graded highly relevant (visible by the fact that a grade of 5 is depicted as an outlier), while for a certain group of queries the opposite was true. For some queries it was even the case that no top 20 passage was graded irrelevant.

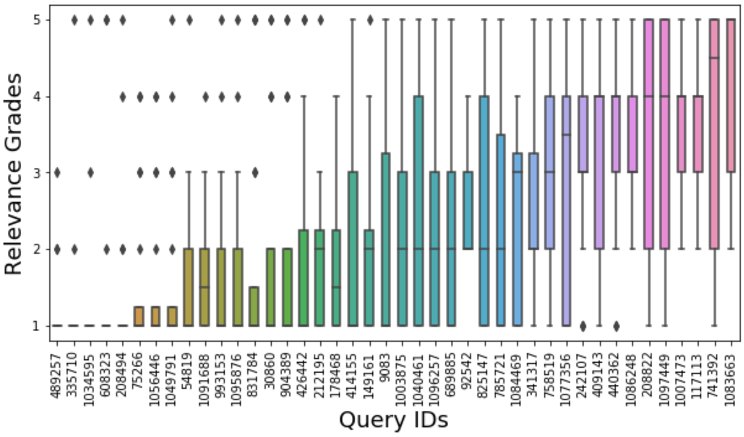


FIGURE 4.2: Distribution relevance grades among the experiment queries.

The next two figures provide a better image of the experiment queries themselves. Figure 4.3 shows the distribution of number of query terms across the experiment queries. The only preprocessing performed on the queries was punctuation removal. Other than that all query terms included were counted, even duplicate or stop words. Figure 4.4 shows the distribution of answer types across the experiment queries. In order to classify the answer types, the same rule-based answer type classifier[2] as in the work by Padigela, Zamani, and Croft, 2019 was used. This classifier is inspired on the work by Li and Roth, 2002 and is able to classify question based on six different expected answer types: abbreviation (ABBR), entity (ENTY), description (DESC), human (HUM), location (LOC) and numeric value (NUM). Padigela et al. used the classifier to better understand the performances of BM25 and BERT across different types of questions by measuring the MRR scores across the 6 different question types for both models. Their work shows that BERT performs the best on abbreviation type questions and achieves the lowest performance on numerical and entity type questions. In contrast, the BM25 performs the worst on abbreviation type questions and achieves its best performance on location and human type questions. This study performs a similar classification on the experiment queries in order to provide a more detailed analysis of the data used. The classifier failed to provide an answer type for 20 queries and so the answer types of these queries were manually constructed[3]. Figure 4.4 shows that there are zero abbreviation questions among

[1] Both boxplots for the Liberal and Strict dataset (depicted in Figure 4.1) show the minimum at 0. This is incorrect as these datasets did not contain any query with no relevant passage. There was no time left to investigate this error and so it was decided to keep it this way.

[2] https://github.com/superscriptjs/qtypes (Last accessed: 20/7/2020)

[3] Based on the descriptions found at: https://cogcomp.seas.upenn.edu/Data/QA/QC/definition.html (Last accessed: 9/7/2020)

the experiment queries and a large portion of queries are either of type description or numeric.

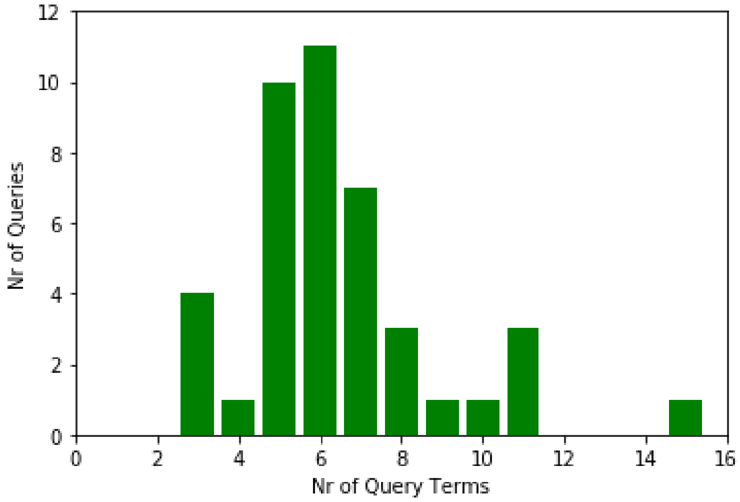


FIGURE 4.3: Distribution query lengths among the experiment queries.

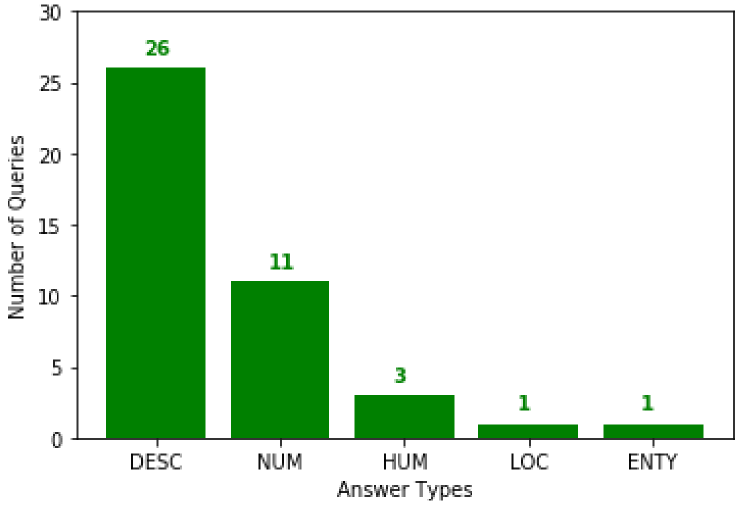


FIGURE 4.4: Distribution of answer types among the experiment queries.

4.1.1 Assessor Statistics

In total 37 different assessors contributed to the online assessment. Figure 4.5 shows how many queries the assessors assessed. In most cases 6 queries were assessed, which is no surprise as participants were explicitly asked to assess at least 6 different queries. Figures 4.6 and 4.7 show the distribution of assessors among the entire set of assessed queries and the queries used during the experiments, respectively.

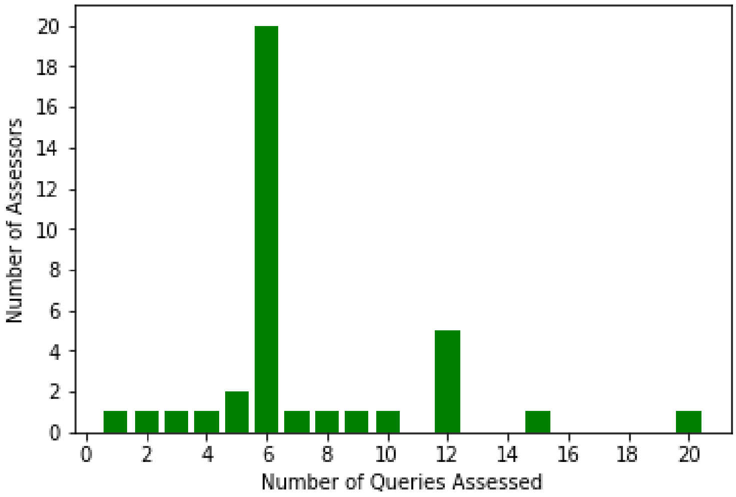


FIGURE 4.5: This plot depicts how many queries were assessed by how many assessors.

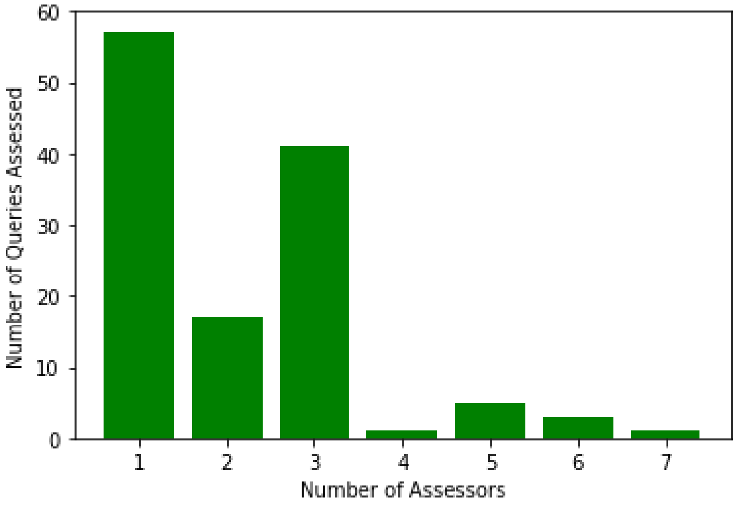


FIGURE 4.6: Distribution assessors among all assessed queries.

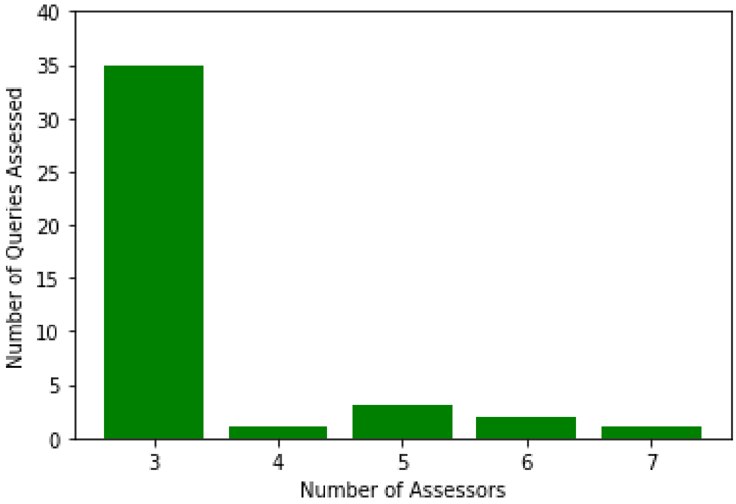


FIGURE 4.7: Distribution assessors among experiment queries.

The online assessment was made available to anyone who wanted to participate. The link was initially shared among fellow researchers at the Radboud University, but eventually also shared among friends and family who in turn shared it amongst acquaintances. No admission requirements were established up front and the only personal data collected from the participants was their e-mail address and their English level. The e-mail addresses were replaced by anonymized identifiers directly after the study was completed. The participants were asked to assess their own English level on a scale from 1 (beginner) to 9 (very advanced)[4]. Corresponding CEFR[5] levels were also provided to help the participant in assessing their English level. The idea behind collecting the English level of the participants was to eventually discard any data provided by those with little understanding of the English language. However, in the end no data was discarded as only a few participants had a English level below 7 (which was the threshold taken for this study) and no low English level participants assessed the same query, which meant that in all cases the query was assessed by a majority of high English level participants. If it was decided to discard the input of low English level participants less queries would meet the criteria of this study. Many queries would not be assessed by at least 3 different assessors anymore. Figure 4.8 shows the distribution of the assessors across the English levels. Most participants had a level above the threshold.

[4] https://www.londonschool.com/level-scale/ (Last accessed: 9/7/2020)

[5] https://www.londonschool.com/blog/all-about-cefr/ (Last accessed: 9/7/2020)

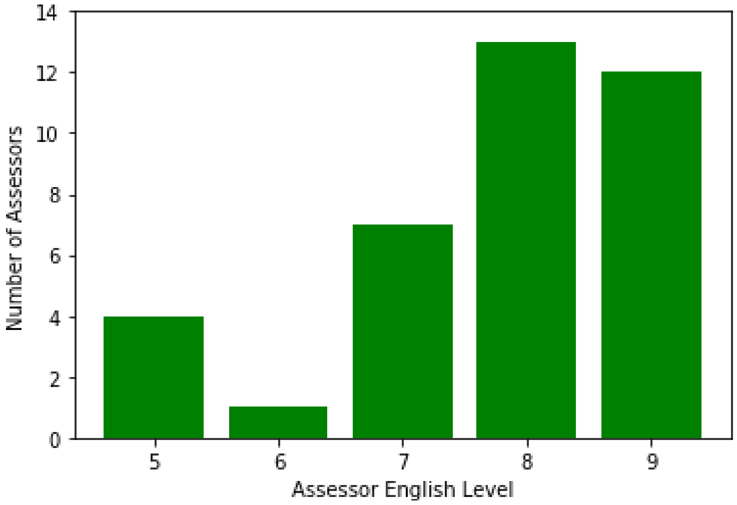


FIGURE 4.8: Distribution English level of assessors.

4.2 Results Experiments

Each of the following subsections is dedicated to the results of one specific experiment as mentioned in Section 3. Performance scores for the metrics MRR, P and NDCG are presented on a scale from 0 to 1, where 1 represents the perfect score. The same is true for the MFR metric as 1 also represents the perfect score. However, the scale used for this metric is not between 0 to 1 as it represents the mean rank. Instead MFR scores range between 1 and 11. The higher the MFR score the worse the performance of a system is, which is the opposite for the other metrics used. For each experiment, differences in score means are tested on significance using the Student Paired t-test. The choice for this statistical significance test was based on the work by Smucker, Allan, and Carterette, 2007. They show that there is little practical difference between the Randomization, Bootstrap and Student Paired t-test.

4.2.1 Experiment 1: BM25 old relevance labels vs. new relevance labels

Table 4.1 shows the results of evaluating BM25's rankings using either the MS MARCO relevance dataset or the Liberal relevance dataset. All three metrics show an improvement in the performance of BM25 in the passage ranking task when the new relevance labels are used for evaluation. The model even scores near perfect performances with the MRR (0.976 for both cutoffs) and MFR metric (1.05 for both cutoffs), which hints that the model was able to rank relevant items first in most of the cases. Taking a closer look at the rankings, BM25 was able to rank relevant items first for 90% of the queries. The precision scores for the two different cutoffs are also improved using the new labels. Where the highest increase in improvement is visible for P@5, clarifying that BM25 manages to rank more relevant compared to irrelevant passages in the top 5. The improvement in precision is less for P@10. One reason for this phenomenon could be that the top 5 ranks for a query already contain all relevant passages, which means that if we increase the cutoff from 5 to 10 we will introduce more irrelevant passages decreasing the precision score. Applying the paired student-t test on the differences in performance scores for using the MS

MARCO and Liberal datasets indicates statistically significant differences between the measured scores across all metrics.

TABLE 4.1: BM25's average performance across 42 queries on the MS MARCO passage ranking task using two different relevance datasets (MS MARCO and Liberal). Paired Student-t test significance is shown with asterisks: * significant at $p < .05$; ** significant at $p < .01$; *** significant at $p < .001$; **** significant at $p < .0001$.

Metric	MS MARCO	Liberal	Δ Diff
MRR@5	0.471	0.976	$+0.505^{****}$
MRR@10	0.491	0.976	$+0.485^{****}$
MFR@5	3.26	1.05	-2.21^{****}
MFR@10	4.14	1.05	-3.09^{****}
P@5	0.143	0.805	$+0.662^{****}$
P@10	0.086	0.671	$+0.585^{****}$

BM25's performance is also measured making use of the Strict relevance label dataset and these results are shown in Table 4.2. Compared to using the MS MARCO dataset, an improvement in performance across all metrics is measured. However, the increase in improvement is less than it was when utilizing the Liberal relevance label dataset. A reason for this is that by taking a more strict binary threshold of 3 instead of 2 causes more queries to be labeled as irrelevant. The retrieved rankings for any given query can contain a passage labeled relevant using the Liberal dataset and irrelevant using the Strict dataset. Still performances obtained with the Strict dataset show that BM25 is capable of retrieving many relevant passages (P@5 = 0.633; P@10 = 0.505). Using the Strict dataset shows that BM25 is able to rank a relevant passage first for 85% of the queries.

TABLE 4.2: BM25's average performance across 42 queries on the MS MARCO passage ranking task using two different relevance datasets (MS MARCO and Strict). Paired Student-t test significance is shown with asterisks: **** significant at $p < .0001$.

Metric	MS MARCO	Strict	Δ Diff
MRR@5	0.471	0.917	$+0.446^{****}$
MRR@10	0.491	0.917	$+0.426^{****}$
MFR@5	3.26	1.24	-2.02^{****}
MFR@10	4.14	1.24	-2.90^{****}
P@5	0.143	0.633	$+0.490^{****}$
P@10	0.086	0.505	$+0.419^{****}$

4.2.2 Experiment 2: BERT old relevance labels vs. new relevance labels

With experiment 2 we compute the performance of BERT on the passage ranking task using the original MS MARCO relevance labels and the new relevance label datasets. Table 4.3 shows the performance scores for BERT using the MS MARCO and the Liberal relevance label datasets. Across all three metrics used, improvement in performance is measured when using the Liberal relevance label dataset. Like the BM25 model, BERT obtains near perfect MRR and MFR scores and manages to rank a relevant passage first for 90% of the queries. Precision scores are also improved using the Liberal relevance labels.

TABLE 4.3: BERT's average performance across 42 queries on the MS MARCO passage ranking task using two different relevance datasets (MS MARCO and Liberal). Paired Student-t test significance is shown with asterisks: ** significant at $p < .01$; *** significant at $p < .001$; **** significant at $p < .0001$.

Metric	**MS MARCO**	**Liberal**	**Δ Diff**
MRR@5	0.718	0.948	$+0.230^{***}$
MRR@10	0.728	0.948	$+0.220^{****}$
MFR@5	2.02	1.12	-0.90^{**}
MFR@10	2.48	1.12	-1.36^{**}
P@5	0.171	0.671	$+0.500^{****}$
P@10	0.093	0.502	$+0.409^{****}$

Table 4.4 shows the performance of BERT using the MS MARCO and Strict relevance label datasets. BERT's performance across all metrics is improved using the new relevance labels.

TABLE 4.4: BERT's average performance across 42 queries on the MS MARCO passage ranking task using two different relevance datasets (MS MARCO and Strict). Paired Student-t test significance is shown with asterisks: ** significant at $p < .01$; *** significant at $p < .001$; **** significant at $p < .0001$.

Metric	**MS MARCO**	**Strict**	**Δ Diff**
MRR@5	0.718	0.925	$+0.207^{***}$
MRR@10	0.728	0.925	$+0.197^{***}$
MFR@5	2.02	1.17	-0.85^{**}
MFR@10	2.48	1.17	-1.31^{**}
P@5	0.171	0.600	$+0.429^{****}$
P@10	0.093	0.426	$+0.333^{****}$

4.2.3 Experiment 3: BM25 vs. BERT (MS MARCO relevance dataset)

With experiment 3 we compare the performance of BM25 and BERT on the passage ranking task using the original MS MARCO relevance labels. Table 4.5 shows the performance scores of both models. For all three metrics BERT shows increased performance on the passage ranking task when compared with BM25. The higher MRR and MFR scores of BERT compared to BM25 indicate that BERT is better in ranking the MS MARCO relevant passage higher than irrelevant passages. Applying the Student Paired t-test on the difference between metric score means of BM25 and BERT indicates statistically significant differences between the scores for all metrics except precision.

TABLE 4.5: BM25 and BERT average performances across 42 queries on the passage ranking task using the original MS MARCO relevance dataset. Paired Student-t test significance is shown with asterisks:** significant at $p < .01$; *** significant at $p < .001$.

Metric	BM25	BERT	Δ Diff
MRR@5	0.471	0.718	$+0.247^{**}$
MRR@10	0.491	0.728	$+0.237^{**}$
MFR@5	3.26	2.02	-1.24^{***}
MFR@10	4.14	2.48	-1.66^{**}
P@5	0.143	0.171	$+0.028$
P@10	0.086	0.093	$+0.007$

4.2.4 Experiment 4: BM25 vs. BERT (Liberal relevance dataset)

Experiment 4 and 5 compare BM25 and BERT on the passage ranking task using the new relevance labels. For experiment 4 we compared BM25 and BERT using the Liberal relevance label dataset, for which the results are presented in Table 4.6. These results indicate the opposite of the results obtained during experiment 3. All three metrics show that BM25 shows better performance on the passage ranking task when compared to BERT. However, differences in MRR and MFR scores are quite little and applying the Student Paired t-test shows no statistically significant differences in score means. The differences in precision means, on the other hand, are statistically significant different.

TABLE 4.6: BM25 and BERT average performances across 42 queries on the passage ranking task using the Liberal relevance dataset. Paired Student-t test significance is shown with asterisks: * significant at $p < .05$; ** significant at $p < .01$; *** significant at $p < .001$.

Metric	BM25	BERT	Δ Diff
MRR@5	0.976	0.948	-0.028
MRR@10	0.976	0.948	-0.028
MFR@5	1.05	1.12	$+0.07$
MFR@10	1.05	1.12	$+0.07$
P@5	0.805	0.671	-0.134^{***}
P@10	0.671	0.502	-0.169^{***}

4.2.5 Experiment 5: BM25 vs. BERT (Strict relevance dataset)

Table 4.7 shows the mean performance scores for both BM25 and BERT on the passage ranking task using the Strict relevant label dataset. The score means measured for both BM25 and BERT across all three metrics show very small differences. BERT seems to achieve slightly better MRR and MFR scores, while BM25 achieves better precision. No statistically significant differences in MRR, MFR and P (except for P@10) score means are measured. Which gives the impression that both models perform near equal on the passage ranking task when the Strict dataset is used for evaluation.

TABLE 4.7: BM25 and BERT average performances across 42 queries on the passage ranking task using the Strict relevance dataset. Paired Student-t test significance is shown with asterisks: * significant at $p < .05$.

Metric	BM25	BERT	Δ Diff
MRR@5	0.917	0.925	+0.008
MRR@10	0.917	0.925	+0.008
MFR@5	1.24	1.17	−0.07
MFR@10	1.36	1.17	−0.19
P@5	0.633	0.600	−0.033
P@10	0.505	0.426	-0.079^{*}

4.2.6 Experiment 6: BM25 vs. BERT (Graded relevance dataset)

The final experiment compares the mean performance scores of BM25 and BERT when graded relevance labels are used. Table 4.8 shows the mean NDCG scores across different cutoffs for both the BM25 and BERT model on the passage ranking task. Comparing NDCG@20 scores, BM25 scores almost 0.1 point higher than BERT with a near perfect score of 0.915. Also evaluating with other cutoffs shows that BM25 achieves higher NDCG scores than BERT. However, differences in scores are minimal when using lower cutoffs, which indicates that both models do not produce such different higher top rankings.

TABLE 4.8: BM25 and BERT average NDCG scores across 42 queries on the passage ranking task using the Graded relevance dataset. Paired Student-t test significance is shown with asterisks: *** significant at $p < .001$.

Metric	BM25	BERT	Δ Diff
NDCG@5	0.787	0.780	−0.007
NDCG@10	0.828	0.784	−0.044
NDCG@20	0.915	0.824	-0.091^{***}

Chapter 5

Discussion

This chapter contains a discussion of several aspects of this study. The first section will focus on the online assessment. Decisions and approaches taken when designing the assessment tool as well as shortcomings of the final design will be discussed. The subsequent section will interpret and discuss the results of the different experiments performed. In the final section other limitations of this study are discussed and possible future work is proposed.

5.1 The Online Assessment

Section 4.1.1 explained that no admission requirements were established in order to select assessors. Instead the participants were only asked to judge their own level of comprehension of the English language. These judgements were gathered to be able to discard input from low English level participants, but in the end all input was considered. The documents which the participants were tasked to assess varied in complexity of English language usage. Some feedback provided after the online assessment was finished, was that certain queries and passages were difficult to comprehend because of the language. In hindsight, participants should have been examined on their capabilities to understand the items they would be assessing. Moreover, a common complaint provided after the assessment was that queries presented were very domain specific and were therefore difficult to assess because of the lack of the specific domain knowledge. This is however not a problem easy to resolve as the MS MARCO passage ranking dataset is an open-domain dataset and contains many queries and passages related to varying domains. Participants could have been prepared to the varying types of queries and passages and trained to handle those cases which are difficult to assess because of expected domain knowledge. Another possible solution would have been to provide an extra option for the participants to select during assessment. This extra option would indicate that the participant did not understand the specific passage. In turn this solution would most likely have introduced a lot of missing data. The final design of the assessment tool ensured that every participant would at least pick one of five relevancy grades.

This introduces another aspect of the assessment procedure that is debatable. The original MS MARCO passage ranking dataset contained binary relevance labels, either a passage is labeled relevant or irrelevant. The participants of the online assessment were tasked to provide multigraded relevance labels to the passages. They could pick a grade between 1 (totally irrelevant) and 5 (perfectly relevant). This introduces variability in the assessments as other aspects are considered in order to decide on a relevancy label. Even more so because the original MS MARCO judges were provided different assessment guidelines than the participants of the online assessment of this study. The MS MARCO judges were tasked to annotate passages relevant if these passages were used to construct an answer to the corresponding query. Participants of the online assessment did not need to construct an answer to the corresponding query. Instead they needed to inspect 20 retrieved passages for any given query and decide on multigraded relevance labels for all of them. They were instructed to inspect each query passage pair independently, without letting their judgements be affected by previous assessments. These differences in assessment procedure will most likely have caused different views on relevancy and, in some cases, caused the participants of the online assessment to disagree with assessments of the MS MARCO judges.

The input labels were multigraded relevance labels which had to be transformed to binary labels. First of all the decision of the binary threshold introduces differences in relevance labels. For this study it was decided to experiment with two thresholds set at 2 and 3. The former of the two caused for a liberal judgement of relevancy as only those passages graded with a 1 would be labeled irrelevant. Setting the threshold at 3, meant that those passages graded with a 1 or 2 would be labeled irrelevant. This threshold could very well be set at 4 or 5, which would mean that only those passages assessed with high relevancy labels would be labeled relevant. The current work did not experiment with these thresholds because they resulted in very few relevant passages and in turn led to an insufficient amount of queries that met the study's criteria.

Aside from the binary threshold, assessors could disagree on the relevance label. If there was no mutual agreement among the assessors, majority voting was performed. For this study it was decided to pick the binary relevance label for which more than half the assessors agreed upon. Another approach would have been to only pick the relevance label with mutual agreement among the assessors and otherwise keep the original MS MARCO relevance label. Again, this approach was not taken because it would have resulted in too few relevant passages.

In the final design of the assessment tool, participants were presented with 20 passages per query. These passages were the top 20 passages retrieved by the BM25 model. In order to prevent order bias, the passages were presented to each participant in randomized order. The same procedure was taken for the presentation of the queries. It was decided to let participants assess 20 passages, because this amount was reasonable for this study. Processing this amount of passages would take between 5 to 10 minutes, after which the participant could pause the assessment and continue at a different time. This number of passages was also not large enough for the participant to get tired of reading, which could cause the participant to rush the assessment. At the same time, it would have been preferable if more passages were assessed. Both BM25 and BERT were used to rank 100 passages per query, but only the top 20 of the BM25 retrieved passages were assessed by the participants. The remaining 80 passages were not assessed and thus the dataset as used during the experiments still contains incomplete relevance labels.

In summary, the assessment procedure could have been improved on several aspects. A selection procedure for the participants could have been designed, which would resulted in the selection of qualified assessors. Furthermore, assessors could have been better instructed or an elaborate training could have been provided in order to ensure more truthful assessments. Finally, more assessments should have been gathered in order to work with a more complete set of relevance labels during the experiments of this study.

5.2 Experiments

In Chapter 4 the results of the different experiments are presented. Experiment 1 and 2 were performed in order to research if the relevance labels that were gathered via the online assessment would affect the performance of the BM25 and BERT model on the MS MARCO passage ranking task. In order to study this, the BM25 and BERT top 100 rankings were evaluated using the original MS MARCO relevance labels and the newly acquired relevance labels constructed using two different binary thresholds. For both BM25 and BERT increased performances are measured when the new relevance labels are used during evaluation. One reason for the improved performances of the BM25 model is that the new relevance labels are gathered by assessment of the initial top 20 passages retrieved by BM25. The online assessment resulted in more relevant passages across the different experiment queries (as presented in Table 4.1) and because of the assessment procedure these relevant passages are already ranked high. The increased performances when evaluating with the new relevance labels indicate that many of these relevant passages were located in the top 10, resulting in high @5 and @10 scores for the different metrics used. BERT then re-ranks the top 100 ranking by BM25 and also achieves improved performances, possibly because it manages to keep many relevant passages in the top 20 when re-ranking.

In the subsequent experiments, the performances of both BM25 and BERT on the passage ranking task are compared using different relevance label datasets. With experiment 3 it is checked if BERT outperforms BM25 using the original MS MARCO relevance dataset. This dataset only contains one relevant passage per query, which is initially ranked by the BM25 model and then re-ranked by the BERT model. BERT outperforms BM25 using this relevance dataset, which mirrors the performance by BERT on the actual passage ranking task as is visible on the current MS MARCO leaderboard.

While using new relevance labels helps to improve the performance of both the BM25 and BERT model on the passage ranking task, BERT no longer outperforms BM25 when performances are compared. A possible reason for this is that both work with a top 100 ranking, but only 20 of these 100 passages were assessed during the online assessment. This introduces incomplete relevance judgements in the new relevance dataset. When BERT then re-ranks the top 100 rankings by BM25, the model possibly introduces unjudged passages to the top 20 resulting in a lower performance as compared to BM25.

We see that BM25 outperforms BERT on precision when the Liberal relevance labels are used, but that both models perform equally well when the Strict relevance labels are considered. This is due to the fact that the Strict relevance labels are constructed using a binary threshold of 3, which means that less passages are considered relevant as compared to using a threshold of 2. When using the Strict relevance labels, the top 20 rankings by BM25 contain more irrelevant passages than when the Liberal relevance labels are used. If the performance of BERT is indeed affected by it ranking unassessed passages higher (who are labeled irrelevant as well), its performance is now less affected by the use of the Strict relevance labels.

Finally, BM25 and BERT are also compared on their NDCG performances. Using the Graded relevance dataset to evaluate both models, it can be concluded that BM25 already achieves high NDCG scores and re-ranking with BERT does not improve these scores. Again, one reason for this is that both BM25 and BERT are computing the ranks of 100 passages while only the multigrade relevance labels of 20 passages are known. To be able to measure the NDCG score, the multigrade relevance label of the unassessed 80 passages is set to 1 (totally irrelevant). As BERT gets to re-rank the initial top 100 ranking of BM25, it is most likely that BERT introduces more unassessed and thus irrelevant passages to the top 20, which results in a lower performance measured with NDCG. The high NDCG score of BM25 indicates that the baseline ranker already ranks many highly relevant item close to the top. Both models achieve high NDCG scores (>0.7), which indicates that they are quite capable of returning near ideal rankings.

5.3 Future Work

Only a small subset of the development queries was used for the online assessment and in the end an even smaller set of queries was processed by the assessors. For each of these queries only the top 20 passages retrieved by BM25 were assessed on relevancy. The assessments gathered during this study prove that the relevance labels in the MS MARCO passage ranking dataset are indeed incomplete and the experiments run with the new labels yield contradicting results as compared to what was previously known on the performance of BM25 and BERT on the passage ranking task. It is suggested that future work repeats the assessment procedure on more query-passage pairs to see if similar results are achieved when a larger and more complete revised (sub)set is used.

The assessment procedure as used in this study was not clearly documented and tested before the start of the online assessment. Participants of the assessment did not undergo any preparation or training in which specific assessment guidelines were explained. Instead, they were only explained what type of items they were going to assess and what type of assessment they could give. It was then left to the participants to decide what degree of relevance was suitable for any given query-passage pair. Future work could follow a more directed approach as taken by Sormunen, 2002 were subjects are provided with more detailed judgement guidelines and the assessment procedure is discussed with assessors. Clearly documenting the assessment guidelines will prevent too much subjectivity in the relevance assessments and decrease disagreement among assessors.

Furthermore, participants of the online assessment were tasked with providing multigrade relevance labels on a scale from 1 to 5. The choice of this scale was based on the traditional 5-point Likert scale (Joshi et al., 2015). Other scales were not considered, but could be experimented with by future work. Sormunen, 2002 make use of a 4-point scale but 7- and even 11-point scales are also used (Borlund, 2003). The MS MARCO passage ranking dataset currently only contains binary relevance labels and so future work could focus on gathering multigrade labels for the entire dataset. This way models can be exhaustively tested on their capabilities of ranking items on their relative relevance. However, to be able to compare results with the current MS MARCO leaderboard, it is suggested that future assessments directly gather binary relevance labels or that more experiments with different binary thresholds are performed.

Lastly, the main interest of this study was to examine if the MS MARCO passage ranking dataset did indeed contain more relevant passages per query than currently labeled. Experiments were therefore only performed using the baseline BM25 ranker by Yang, Fang, and Lin, 2018 and the pre-trained BERT model by Nogueira and Cho, 2019. The performances of other models present in the MS MARCO leaderboard should also be revised in order to construct a more truthful ranking of models and their capabilities on the passage ranking task.

Chapter 6

Conclusion

The current work conducted an online assessment in order to gather more relevant query-passage pairs for the MS MARCO passage ranking dataset. The enhanced dataset was then used to conduct experiments in order to test the performance of both the BM25 and BERT models on the MS MARCO passage ranking task. For both models, it was researched if performances were affected by using the new relevance labels as compared to using the old labels. Moreover, the performances of both models were also compared using the old and the new relevance labels. Finally, graded relevance labels were gathered during the online assessment. These labels were used to compare the performance of BM25 and BERT on the passage ranking task in case of multi-label relevancy. During this work five research questions were addressed, for which the answers are given below:

RQ1 - *Does the MS MARCO passage ranking dataset contain more relevant passages per query than currently labeled, such that evaluation is affected?*

The online assessment that was conducted during this study, resulted in more relevant passages per query than currently labeled. Only a small subset of the entire collection of queries was taken and for each of these queries only 20 passages were assessed on relevancy. Nonetheless, more relevant passages were found and so it can be concluded that indeed that the MS MARCO passage ranking dataset contains more relevant passages per query than currently labeled. In turn, the use of these additional relevancy labels affects the evaluation of two distinct models on the MS MARCO passage re-ranking task, namely BM25 and BERT.

RQ1.1 - *How does BM25 perform on the MS MARCO passage ranking task when multiple relevant passages per query are provided?*

BM25 achieves increased performances on the MS MARCO passage ranking task when the newly acquired relevance labels are used during evaluation instead of the original MS MARCO relevance labels.

RQ1.2 - *How does BERT perform on the MS MARCO passage ranking task when multiple relevant passages per query are provided?*

Bert achieves increased performances on the MS MARCO passage ranking task when the newly acquired relevance labels are used during evaluation instead of the original MS MARCO relevance labels.

RQ1.3 - *Does re-ranking with BERT improve initial rankings by BM25 on the MS MARCO passage ranking task?*

For the subset of queries used by this work, BERT achieves higher performance on the passage ranking task compared to BM25 when the original MS MARCO relevant labels are used. This is not the case when the newly acquired relevance labels are used during evaluation. In case the Liberal relevance label dataset is used, BM25 outperforms BERT in mean precision and MAP scores and both models achieve similar MRR and MFR scores. Both models perform equally well on the passage ranking task when the Strict relevance label dataset is used. In both cases, BERT did not improve the initial rankings by BM25.

RQ2 - *What is the effect of graded relevance judgements on the relative performance of reranking MS MARCO passages with BERT?*

BERT does not improve initial rankings by BM25 on the passage ranking task when graded relevance labels are considered during evaluation. Both models do show high NDCG scores, which indicates that they are capable of ranking higher relevant passages higher than marginal relevant passages.

Bibliography

Abney, Steven, Michael Collins, and Amit Singhal (2000). "Answer Extraction". In: *Sixth Applied Natural Language Processing Conference*, pp. 296–301.

Alaparthi, Chaitanya Sai (2019). "Microsoft AI Challenge India 2018: Learning to Rank Passages for Web Question Answering with Deep Attention Networks". In: *CoRR* abs/1906.06056. arXiv: 1906.06056. URL: http://arxiv.org/abs/1906.06056.

Bajaj, Payal, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. (2016). "MS MARCO: A Human Generated Machine Reading Comprehension Dataset". In: *arXiv preprint arXiv:1611.09268*.

Borlund, Pia (2003). "The Concept of Relevance in IR". In: *Journal of the American Society for information Science and Technology* 54.10, pp. 913–925.

Buscaldi, Davide, Paolo Rosso, José Manuel Gómez-Soriano, and Emilio Sanchis (2010). "Answering Questions with an N-gram Based Passage Retrieval Engine". In: *Journal of Intelligent Information Systems* 34.2, pp. 113–134.

Crijns, Tanja (2019). "Have a Chat with BERT; Passage Re-Ranking using Conversational Context". MA thesis. Institute of Computing and Information Sciences, Radboud University.

Cui, Hang, Renxu Sun, Keya Li, Min-Yen Kan, and Tat-Seng Chua (2005). "Question Answering Passage Retrieval using Dependency Relations". In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 400–407.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

Dhingra, Bhuwan, Kathryn Mazaitis, and William W Cohen (2017). "Quasar: Datasets for Question Answering by Search and Reading". In: *arXiv preprint arXiv:1707.03904*.

Dunn, Matthew, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho (2017). "SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine". In: *arXiv preprint arXiv:1704.05179*.

Fremann, Lea (2019). “Extractive NarrativeQA with Heuristic Pre-Training”. In: *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 172–182.

Fuhr, Norbert (2018). “Some Common Mistakes in IR Evaluation, and How They Can Be Avoided”. In: *ACM SIGIR Forum*. Vol. 51. 3. ACM New York, NY, USA, pp. 32–41.

Gauch, Susan, Jianying Wang, and Satya Mahesh Rachakonda (1999). “A Corpus Analysis Approach for Automatic Query Expansion and its Extension to Multiple Databases”. In: *ACM Transactions on Information Systems (TOIS)* 17.3, pp. 250–269.

Ghaeini, Reza, Xiaoli Z Fern, Hamed Shahbazi, and Prasad Tadepalli (2018). “Dependent Gated Reading for Cloze-Style Question Answering”. In: *arXiv preprint arXiv:1805.10528*.

Green Jr, Bert F, Alice K Wolf, Carol Chomsky, and Kenneth Laughery (1961). “Baseball: An Automatic Question-Answerer”. In: *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, pp. 219–224.

Han, Shuguang, Xuanhui Wang, Mike Bendersky, and Marc Najork (2020). *Learning-to-Rank with BERT in TF-Ranking*. arXiv: 2004.08476 [cs.IR].

He, Wei, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. (2017). “DuReader: A Chinese Machine Reading Comprehension Dataset from Real-World Applications”. In: *arXiv preprint arXiv:1711.05073*.

Hu, Haiqing (2006). “A Study on Question Answering System using Integrated Retrieval Method”. In: *Unpublished Ph. D. Thesis, The University of Tokushima, Tokushima*.

Ingale, Vaishali and Pushpender Singh (2019). “Datasets for Machine Reading Comprehension: A Literature Review”. In: *Available at SSRN 3454037*.

Joshi, Ankur, Saket Kale, Satish Chandel, and D Kumar Pal (2015). “Likert Scale: Explored and Explained”. In: *Current Journal of Applied Science and Technology*, pp. 396–403.

Kočiskỳ, Tomáš, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette (2018). “The NarrativeQA Reading Comprehension Challenge”. In: *Transactions of the Association for Computational Linguistics* 6, pp. 317–328.

Lee, Changki, Yi-Gyu Hwang, Hyo-Jung Oh, Soojong Lim, Jeong Heo, Chung-Hee Lee, Hyeon-Jin Kim, Ji-Hyun Wang, and Myung-Gil Jang (2006). “Fine-Grained Named Entity Recognition using Conditional Random Fields for Question Answering”. In: *Asia Information Retrieval Symposium*. Springer, pp. 581–587.

Li, Xin and Dan Roth (2002). “Learning Question Classifiers”. In: *COLING 2002: The 19th International Conference on Computational Linguistics*.

MacAvaney, Sean, Andrew Yates, and Kai Hui (2017). "Contextualized PACRR for Complex Answer Retrieval." In: *TREC*.

Nogueira, Rodrigo and Kyunghyun Cho (2019). "Passage Re-Ranking with BERT". In: *CoRR* abs/1901.04085. arXiv: 1901.04085. URL: http://arxiv.org/abs/1901.04085.

Padigela, Harshith, Hamed Zamani, and W Bruce Croft (2019). "Investigating the Successes and Failures of BERT for Passage Re-Ranking". In: *arXiv preprint arXiv:1905.01758*.

Prager, John, Jennifer Chu-Carroll, Eric W Brown, and Krzysztof Czuba (2008). "Question Answering by Predictive Annotation". In: *Advances in Open Domain Question Answering*. Springer, pp. 307–347.

Pundge, Ajitkumar M, SA Khillare, and C Namrata Mahender (2016). "Question Answering System, Approaches and Techniques: A Review". In: *International Journal of Computer Applications* 141.3, pp. 0975–8887.

Qi, Peng, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D Manning (2019). "Answering Complex Open-Domain Questions Through Iterative Query Generation". In: *arXiv preprint arXiv:1910.07000*.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang (2016). "SQuAD: 100,000+ Questions for Machine Comprehension of Text". In: *arXiv preprint arXiv:1606.05250*.

Robertson, Stephen and Hugo Zaragoza (2009). *The Probabilistic Relevance Framework: BM25 and Beyond*. Now Publishers Inc.

Robertson, Stephen E, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. (1995). "Okapi at TREC-3". In: *Nist Special Publication Sp* 109, p. 109.

Smucker, Mark D, James Allan, and Ben Carterette (2007). "A Comparison of Statistical Significance Tests for Information Retrieval Evaluation". In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 623–632.

Sormunen, Eero (2002). "Liberal Relevance Criteria of Trec- Counting on Negligible Documents?" In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 324–330.

Stevens, Stanley Smith et al. (1946). "On the Theory of Scales of Measurement". In:

Taylor, Wilson L (1953). ""Cloze Procedure": A New Tool for Measuring Readability". In: *Journalism quarterly* 30.4, pp. 415–433.

Tellex, Stefanie, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton (2003). "Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering". In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 41–47.

Trischler, Adam, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman (2016). “NewsQA: A Machine Comprehension Dataset”. In: *arXiv preprint arXiv:1611.09830*.

Voorhees, Ellen M and Dawn M Tice (2000). “Building a Question Answering Test Collection”. In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 200–207.

Wang, Shuohang, Mo Yu, Jing Jiang, Wei Zhang, Xiaoxiao Guo, Shiyu Chang, Zhiguo Wang, Tim Klinger, Gerald Tesauro, and Murray Campbell (2017). “Evidence Aggregation for Answer Re-ranking in Open-Domain Question Answering”. In: *arXiv preprint arXiv:1711.05116*.

Xiong, Chenyan, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power (2017). “End-to-End Neural Ad-Hoc Ranking with Kernel Pooling”. In: *CoRR* abs/1706.06613. arXiv: 1706.06613. URL: http://arxiv.org/abs/1706.06613.

Yang, Peilin, Hui Fang, and Jimmy Lin (Oct. 2018). “Anserini: Reproducible Ranking Baselines Using Lucene”. In: *J. Data and Information Quality* 10.4. ISSN: 1936-1955. DOI: 10.1145/3239571. URL: https://doi.org/10.1145/3239571.

Ye, Xin, Hui Shen, Xiao Ma, Razvan Bunescu, and Chang Liu (2016). “From Word Embeddings to Document Similarities for Improved Information Retrieval in Software Engineering”. In: *Proceedings of the 38th international conference on software engineering*, pp. 404–415.

Zhai, ChengXiang and Sean Massung (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Association for Computing Machinery and Morgan & Claypool.

Zhang, Xin, An Yang, Sujian Li, and Yizhong Wang (2019). “Machine Reading Comprehension: A Literature Review”. In: *arXiv preprint arXiv:1907.01686*.