Ad-hoc Document Retrieval using Weak-Supervision with BERT and GPT2

Yosi Mass
IBM Research
Haifa University, Mount Carmel, Haifa,
HA 31905, Israel
yosimass@il.ibm.com

Haggai Roitman*
eBay Research
Netanya, Israel
hroitman@ebay.com

Abstract

We describe a weakly-supervised method for training deep learning models for the task of ad-hoc document retrieval. Our method is based on generative and discriminative models that are trained using weak-supervision based solely on the documents in the corpus. We present an end-to-end retrieval system that starts with traditional information retrieval methods, followed by two deep learning re-rankers. We evaluate our method on three different datasets: a COVID-19 related scientific literature dataset and two news datasets. We show that our method outperforms state-of-the-art methods; this without the need for the expensive process of manually labeling data.

1 Introduction

The ad-hoc retrieval task has been extensively studied by the Information Retrieval (IR) community. Traditional IR models evaluate ad-hoc queries against documents mainly on a syntactic (exact) word-matching basis (Manning et al., 2008). Recent years advances in Deep Learning (DL) methods have lead to further improvement in IR tasks, and among others, in ad-hoc document retrieval (Guo et al., 2019). DL methods add a semantic dimension to IR methods. However, such methods usually require large amounts of labeled data for model training.

In this work, we describe a novel weakly-supervised method for training DL methods for ad-hoc document retrieval. Motivated by the recent work of (Mass et al., 2020) on Frequently Asked Questions (FAQ) retrieval, we assume that documents have at least three fields, namely *title*, *abstract* and *content*. Such documents are actually quite common nowadays in the scientific and news domains. Our main hypothesis is that: *titles and abstracts can take the role of questions and answers of FAQs, respectively*.

Whenever a document is missing a title, we consider its first sentence as its augmented title. In a similar way, whenever a document is missing an abstract, we consider the first 512 words of its content as the abstract.

The three fields are used for retrieving candidate documents. Inspired by (Mass et al., 2020), the title and abstract fields are further used as a weak-supervision data source for training two independent BERT (Devlin et al., 2019) models, that are then used to re-rank those candidates documents.

The first model matches user queries to documents' abstracts. Here we use the title-to-abstract associations to fine-tune a BERT model to semantically match queries to abstracts. The second model matches user queries to titles. Here our assumption is that by generating title paraphrases, we can train a model to match user queries to titles. To this end, we use GPT2 (Radford et al., 2018) to generate title paraphrases, which are then utilized for fine-tuning the second BERT model.

While our work is closely related to (Mass et al., 2020), with the lack of human-curated questions (such as in FAQs), we still need to resort to title paraphrases as (noisy) pseudo-questions and transfer (Mass et al., 2020)'s method to the more general task of ad-hoc document retrieval. Moreover, compared to FAQs that are relatively short, the current task deals with documents that can be quite long. Thus, in current paper we use three fields (title, abstract, content) and present a strong IR base line instead of only two fields and a simple IR baseline used in (Mass et al., 2020)

As a proof of concept, we evaluate our method on three benchmarks: *TREC-COVID* - a scientific literature dataset on COVID-19 topics; and TREC's newswire corpora: *Associated Press* (AP) and *Wall Street Journal* (WSJ). By combining the

* Work done while affiliated with IBM.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4191–4197,
November 16–20, 2020.

two weakly-supervised BERT models with an existing strong IR baseline, we demonstrate that the former can help to elevate the performance of the latter. Our approach further outperforms state-of-the-art methods on these benchmarks.

2 Related Work

With the lack of training data, several weakly-supervised alternatives have been explored so far for the task at hand. (Dehghani et al., 2017b,a) and (Nie et al., 2018) have utilized rankings produced by BM25 model as training samples. (MacAvaney et al., 2019) have used pseudo query-document pairs that already exhibit relevance (e.g., newswire headline-content pairs). (Frej et al., 2019) have utilized Wikipedia's internal linkage to define automated queried topics. (Zhang et al., 2020) have used anchor texts and their linked web pages as query-document pairs.

Our work is different from all those works as we train a model to generate title paraphrases that are used to enable query-to-title (question) matching and not only query-to-abstract (answer) matching.

(Ma et al., 2020) have proposed a zero-shot retrieval approach using synthetic query generation by training a generative model on a different Community QA data. Our work differs from (Ma et al., 2020) in three main aspects. First, (Ma et al., 2020) focuses on QA, where answers are very short, while we generate title paraphrases from full abstracts. Second, we train a model to generate title paraphrases which are used to enable not only query-to-abstract (answer) matching, but also query-to-title (question) matching. Third, (Ma et al., 2020) filters the input QA pairs that are used to train the generative model by taking only pairs that were voted by at least one-user on those Community QA (CQA) sites. We do not have such voting so we use a smart filtering on the output data (namely on the generated title-paraphrases) as described in Section 3.3.

The work in (Chang et al., 2020) suggests an efficient neural method for initial retrieval of candidates. Their method uses a two-tower architecture which learns a different representation for passages and for queries. While their method can be used as an initial retrieval (instead of our IR method), the authors of (Chang et al., 2020) still require an additional re-ranking step. Thus it does not replace our two weakly-supervised BERT re-ranking models. Moreover, our two BERT models learn a joint attention-based representation for pairs of (query, abstarct) and (query, title) while in (Chang et al., 2020) they learn a separate representation for queries and passages.

3 Method

Inspired by (Mass et al., 2020), we consider the ad-hoc document retrieval problem as an instance of FAQ retrieval, where a document's title represents the question and its abstract the answer.

Our proposed retrieval approach allows to enhance existing state-of-the-art ad-hoc retrieval methods with weakly-supervised neural models that are completely trained from the documents collection itself without the need to supply manual relevance labels. Following the common approach (Guo et al., 2019), these neural-models are utilized for re-ranking candidate documents retrieved by a given IR baseline.

In what follows, the initial candidate documents retrieval uses pure IR similarities and relevance models (Section 3.1). The re-ranking step exploits two independent weakly-supervised BERT models, namely: **BERT-Q-a** (Section 3.2) for matching queries to abstracts and **BERT-Q-t** (Section 3.2) for matching queries to titles.

The final re-ranking is obtained by combining the outcome of the baseline IR method and the two BERT-based re-rankers using an unsupervised late-fusion step (Section 3.4). The components of our approach are described in the rest of this section.

3.1 Initial retrieval

We first obtain for each query a reasonable pool of candidate documents to be re-ranked using our weakly-supervised models. To this end we retrieve several ranked lists from an Apache Lucene[1] index using various state-of-the-art IR similarities. that are available in Lucene. The various retrieved lists are then combined to generate a single pool of top-k candidates for re-ranking by employing the *PoolRank* (Roitman, 2018) fusion method. We refer to this IR pipeline as **IR-Base**.

The IR similarities and the PoolRank method have few free-parameters that are tuned so to optimize Mean Average Precision (MAP@1000). Details are given in the experimental setup (Section 4.2) below.

[1] https://lucene.apache.org/

3.2 BERT-Q-a

We use pairs of title-abstract (t,a) of documents in the collection as a weak-supervision data source for fine-tuning a pre-trained BERT model which is then used to match user queries to abstracts.

Similar to (Mass et al., 2020), we fine-tune the BERT model (denoted **BERT-Q-a**) using a triplet network (Hoffer and Ailon, 2015). This network is adopted for BERT fine-tuning (Mass et al., 2019) using triplets (t, a, a'), where (t, a) constitutes a document title and its abstract. a' is a negative sampled abstract, obtained as follows. We run t as a query against the index (using the title and abstract fields) and sample n random abstracts from the top-k retrieved documents as negative examples (excluding a) (in our setup we used k=100 and n=2). At run time, given a user query Q, **BERT-Q-a** re-ranks the top-k candidate documents by matching Q to the abstracts (a) only.

3.3 BERT-Q-t

Similar to (Mass et al., 2020), we fine-tune a generative pre-trained (GPT-2) neural network model (Radford et al., 2018) for generating title paraphrases (instead of question paraphrases as in (Mass et al., 2020)).

Using N (t_i,a_i)-pairs, we concatenate titles and their abstracts into a long text $U = a_1$ [SEP] t_1 [EOS] $\cdots$ a_N [SEP] t_N [EOS], where [SEP] and [EOS] are special tokens. The GPT-2 fine-tuning samples sequences of l consecutive tokens in U (in our setup we used l=256), aiming to maximize the Language Model (LM) probability for generating the last token on each sequence, given its $l-1$ preceding tokens.

Once the model is fine-tuned, we feed it with the text "a [SEP]", (a is an abstract), and let it generate tokens until [EOS] is generated. We take all generated tokens excluding [EOS], as a paraphrase to a's title t. We repeat the generation process n times (e.g., n=10) to generate n paraphrases to each title.

The generated paraphrases are filtered to ensure high quality paraphrases (Mass et al., 2020). Each paraphrase is run as a query against the Lucene index and only paraphrases that return the exact same documents as their original title are kept.

The filtered paraphrases are then used to fine-tune a second BERT model (denoted **BERT-Q-t**), using a triplet network (similar to **BERT-Q-a**), with triplets (p, t, t'), where p is a paraphrase of t and t' is a randomly selected title from the corpus.

At run time, given a user query Q, **BERT-Q-t** re-ranks the top-k candidate documents by matching Q to titles (t) only.

3.4 Enhanced ad-hoc retrieval using Fusion

To enhance ad-hoc retrieval quality, we now propose to combine the two weakly-supervised fine-tuned BERT models with the baseline IR method (**IR-Base**, see again Section 3.1). To this end, following (Roitman, 2018), we utilize the *Two-Step PoolRank* (denoted **TSPR**) unsupervised fusion method – an extended PoolRank method that estimates document relevance using the three ranked lists (obtained by **IR-Base**, **BERT-Q-a** and **BERT-Q-t**) as pseudo-relevance evidence sources.

4 Evaluation

4.1 Datasets and Indexing

We evaluated our proposed approach using three different benchmarks. The first benchmark, **TREC-COVID**[2], is based on the CORD-19 dataset[3], which contains scientific documents related to the recent Coronavirus pandemic. We used the Round-1 challenge which consists of 43K documents[4] and 30 topics (queries) with their query relevance sets (qrels). Documents in this dataset have three fields (title, abstract and content). The two other benchmarks are based on news articles datasets: **AP** (Association Press, about 242K docs) and **WSJ** (Wall Street Journal, about 160K docs). These datasets are part of the TREC ad-hoc retrieval newswire collection[5]. Here we used topics 51-150 and topics 151-200 (with their respective qrels) for the AP and WSJ datasets, respectively. Those two datasets have only title and content so we created the abstract by taking the first 512 tokens of the content.

We used Apache Lucene to process and indexed the (multi-field) documents, employed with English analysis (tokenization, lower-casing, Porter stemming and stopping). Each indexed document has three main fields: *title*, *abstract* and *content*.

4.2 Experimental Setup

We used an initial candidate pool of $k = 1000$ documents retrieved by **IR-Base** and re-ranked by the two BERT models. We detail below the setup of each of the three rankers and their fusion.

[2] https://bit.ly/2ApmLcz

[3] https://bit.ly/3dxyZ1i

[4] Round-1 contained about 51K documents, but we kept only those that have a non-empty content

[5] https://bit.ly/3gJcF6X

IR-Base. The following Lucene similarities configurations were used: i) BM25Similarity (Robertson and Zaragoza, 2009) with $k1 = 1.2$ and $b = 0.7$. ii) LMDirichletSimilarity (Zhai, 2009) with Dirichlet-smoothing parameter $\mu = 200$ and $\mu = 1000$ for TREC-COVID and news datasets, respectively. iii) DFRSimilarity (Amati and Van Rijsbergen, 2002) with BasicModelIF, AfterEffectB and NormalizationH3. iv) AxiomaticF1LOG (Fang and Zhai, 2005) with growth parameter $s = 0.25$ and $s = 0.1$ for TREC-COVID and news datasets, respectively.

BERT models. We used the pytorch huggingface implementation of BERT and GPT2[6]. For the two BERT models we used bert-base-uncased (12-layers, 768-hidden, 12-heads, 110M parameters). Fine-tuning was done with a learning rate of 2e-5 and 3 training epochs. For training BERT-Q-a on each of the three datasets, we used a subset of their first 20K documents. For TREC-COVID, we used SciBERT model (Beltagy et al., 2019) (that was pre-trained on 1M scientific documents), as it yields better results than using the vanilla pre-trained BERT model. This is mainly due to the scientific nature of the documents in this benchmark.

GPT2. For generating title paraphrases we used GPT2 small model (12-layers, 768-hidden, 12-heads, 110M parameters). For fine-tuning we used (title, abstract) pairs from all documents of TREC-COVID and a subset of the first 20K documents of the other two datasets. We generated 10 paraphrases for the first 20K documents of each of the three datasets. After filtering the generated paraphrases, we were left with 18K, 4.5K and 3.5K paraphrases for TREC-COVID, WSJ and AP respectively. [7]

Fusion. We fine-tuned the PoolRank (Roitman, 2018) method's parameters for all datasets as follows: For *Base fusion* we used *CombSUM* (Nuray and Can, 2006) with sum-normalization. The other parameters were set as: *Pseudo-relevance set size*: 5 documents. *Term clip size*: 100. Document re-ranking using KL-score (equally interpolated with the CombSUM score) with Dirichlet-smoothing parameter $\mu = 200$ and $\mu = 1000$ for TREC-COVID and news datasets, respectively.

We assessed retrieval quality using the following metrics: *Precision* (P@5), *Normalized Discounted Cumulative Gain* (NDCG@10) and *Mean Average Precision* (MAP@1000). All experiments were run on two 32GB V100 GPUs. The re-ranking times of 1000 documents for each query were 11 sec for **BERT-Q-a** (using BERT's max_seq_len of 512) and 5 sec for **BERT-Q-q** (max_seq_len = 256).

4.3 Results

We now report the evaluation results of the TREC-COVID benchmark and the two news benchmarks (AP and WSJ) in Table 2 and Table 3, respectively. We compared our three rankers (**IR-Base, BERT-Q-a and BERT-Q-t**) and their fusion (**TSPR**). We further evaluated two additional **TSPR** versions, namely: **TSPR-Q-a** and **TSPR-Q-t** where we only fused the **IR-Base** ranked-list with either **BERT-Q-a** or **BERT-Q-t**, respectively.

To demonstrate the relative effectiveness of our proposed approach, we compared its quality to state-of-the-art alternative baselines. On TREC-COVID, we directly compared against the three best automatic performing systems[8] (out of 141 system runs submitted to the Round-1 challenge by 56 different teams), namely: **sabir**, **IRIT_markers** and **unipd.it**.

On the news benchmarks (AP and WSJ), we compared against quality metrics (when available) that were previously reported for the following state-of-the-art unsupervised and semi-supervised IR methods: **ClustMRF** (Raiber and Kurland, 2013), **NVSM** (Gysel et al., 2018), **LBDM** (Wei and Croft, 2006), **PGR** (Krikon et al., 2011) and **CRM** (Gelfer Kalmanovich and Kurland, 2009).

The symbols $\triangle$ and $\blacktriangle$ in both tables denote a statistical significant ($p < 0.05$) result with **IR-Base** and the best alternative baseline, respectively.

4.3.1 Retrieval enhancement

The first and most important observation that we now make is that, consistently over the three benchmarks, the proposed method **TSPR**, which fuses the initial IR retrieval (**IR-Base**) and the two weakly-supervised BERT models, performs significantly better than each of the three separately, on all measures. As a second observation, we note that, **TSPR** employed with both BERT models significantly outperforms **TSPR-Q-a** and **TSPR-Q-t**.

[6] https://bit.ly/2Me0Gk1

[7] The filtered paraphrases can be downloaded from https://github.com/YosiMass/ad-hoc-retrieval

[8] The details of these systems as well as other competing systems are available in https://bit.ly/2XjkE2T

These two observations confirm our hypothesis that: 1) BERT contributes a semantic understanding of the data and thus improves the ad-hoc retrieval task over pure IR methods; and 2) each of the two BERT models contributes a different semantic aspect. **BERT-Q-a**, which was trained on the relation between titles and abstracts, allows to consider the semantic similarity between a user query and abstracts. Moreover, **BERT-Q-t**, which was trained on titles and their paraphrases, can successfully match a user query to titles.

To examine the semantic differences of our three rankers, we report their P@1 performance. On **TREC-COVID**, there were 12 queries in which **BERT-Q-t** and **IR-Base** differed in their P@1, and 9 queries in which **BERT-Q-a** and **IR-Base** differed. On **AP**, differences from **IR-Base** were on 32 and 31 queries for **BERT-Q-t** and **BERT-Q-a** respectively, and on **WSJ**, differences were on 11 and 23 queries for **BERT-Q-t** and **BERT-Q-a** respectively.

Table 1 shows some example queries from **TREC-COVID**, where **BERT-Q-t** returned a correct top-1 answer (showing its title), while **IR-Base** returned a wrong one.

Query	how does the coronavirus respond to changes in the weather
BERT-Q-t	The Effects of Temperature and Relative Humidity on the Viability of the SARS Coronavirus
Query	how long can the coronavirus live outside the body
BERT-Q-t	Microbes, Transmission Routes and Survival Outside the Body

Table 1: Example queries and titles of correct top-1 documents retrieved by **BERT-Q-t** on **TREC-COVID**

Looking further at the effect of each of the two BERT models as a standalone ranker, we can see that on **TREC-COVID**, **BERT-Q-t** performed better than **BERT-Q-a**, while on the two news datasets it was the other way around. This can be attributed to the length of the titles. In **TREC-COVID** titles are much longer (13 words on average compared to 9.8 and 8.2 words on **WSJ** and **AP** respectively) and hence carry more information.

4.3.2 Comparison with alternative baselines

Looking further down the tables, we notice that our proposed method, **TSPR**, outperforms all alternative baselines in most of the cases and metrics.

On the **TREC-COVID** benchmark, **TSPR** provides a better retrieval quality compared to the best systems. Interestingly, some systems (such as **IRIT_markers**) fine-tuned a BERT model (including SciBERT) using an auxiliary largely annotated dataset such as MS-Marco, yet still fall behind **TSPR**'s quality. This serves as another strong empirical evidence on the importance of our weakly-supervised BERT fine-tuning directly on the domain's data.

Finally, on the two news benchmarks, **TSPR** overpass most of the quality metrics that were previously reported for state-of-the-art alternatives.

Table 2: Retrieval quality on TREC-COVID.

Method	P@5	NDCG@10	MAP
IR-Base	.753	.597	.297
BERT-Q-a	.466	.373	.148
BERT-Q-t	.620	.506	.186
TSPR-Q-a	.693	.555	.270
TSPR-Q-t	.747	.625	.254
TSPR	**.827**$^{\triangle\blacktriangle}$	**.652**$^{\triangle\blacktriangle}$	**.315**$^{\triangle}$
sabir	.780	.608	.313
IRIT_markers	.733	.586	.248
unipd.it	.727	.572	.208

Table 3: Retrieval quality on news benchmarks.

Method	**AP**			**WSJ**		
	P@5	NDCG@10	MAP	P@5	NDCG@10	MAP
IR-Base	.480	.460	.237	.564	.557	.319
BERT-Q-a	.406	.411	.179	.444	.455	.204
BERT-Q-t	.380	.382	.168	.452	.470	.195
TSPR-Q-a	.528	.527	.268	.592	.609	.361
TSPR-Q-t	.512	.507	.267	.592	.588	.342
TSPR	**.592**$^{\triangle}$	**.570**$^{\triangle}$	.275$^{\triangle}$	**.676**$^{\triangle}$	**.664**$^{\triangle}$	.368$^{\triangle}$
ClustMRF	.559	-	-	-	-	-
NVSM	-	-	.257	-	-	.208
LBDM	-	-	.265	-	-	-
PGR	.537	-	-	.612	-	-
CRM	.521	-	**.301**	.620	-	**.409**

5 Conclusions and Future work

We have cast a solution for FAQ retrieval to a solution for ad-hoc document retrieval, where titles and abstracts took the role of questions and answers in FAQs. We have shown that, using the corpus itself, we could generate weakly-supervised title paraphrases for training a BERT model that matches queries to titles. Coupled with a second BERT model that was trained to match queries to abstracts, we have experimentally shown on three different benchmarks that our proposed method outperformed state-of-the-art alternatives.

As a future work, we plan to utilize automatic summarization for missing abstracts, instead of taking the first 512 content tokens.

References

Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval.

Mostafa Dehghani, Aliaksei Severyn, Sascha Rothe, and Jaap Kamps. 2017a. Avoiding your teacher's mistakes: Training neural networks with controlled weak supervision.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017b. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 65–74, New York, NY, USA. Association for Computing Machinery.

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

Hui Fang and ChengXiang Zhai. 2005. An exploration of axiomatic approaches to information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, page 480–487, New York, NY, USA. Association for Computing Machinery.

Jibril Frej, Didier Schwab, and Jean-Pierre Chevallet. 2019. Wikir: A python toolkit for building a large-scale wikipedia-based english information retrieval dataset.

Inna Gelfer Kalmanovich and Oren Kurland. 2009. Cluster-based query expansion. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 646–647, New York, NY, USA. Association for Computing Machinery.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W. Bruce Croft, and Xueqi Cheng. 2019. A deep look into neural ranking models for information retrieval.

Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. 2018. Neural vector spaces for unsupervised information retrieval. *ACM Trans. Inf. Syst.*, 36(4).

Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

Eyal Krikon, Oren Kurland, and Michael Bendersky. 2011. Utilizing inter-passage and inter-document similarities for reranking search results. *ACM Trans. Inf. Syst.*, 29(1).

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2020. Zero-shot neural retrieval via domain-targeted synthetic query generation. *CoRR*, abs/2004.14503.

Sean MacAvaney, Andrew Yates, Kai Hui, and Ophir Frieder. 2019. Content-based weak supervision for ad-hoc re-ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 993–996, New York, NY, USA. Association for Computing Machinery.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

Yosi Mass, Boaz Carmeli, Haggai Roitman, and David Konopnicki. 2020. Unsupervised FAQ retrieval with question generation and BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 807–812, Online. Association for Computational Linguistics.

Yosi Mass, Haggai Roitman, Shai Erera, Or Rivlin, Bar Weiner, and David Konopnicki. 2019. A study of bert for non-factoid question-answering under passage length constraints. *CoRR*, abs/1908.06780.

Yifan Nie, Alessandro Sordoni, and Jian-Yun Nie. 2018. Multi-level abstraction convolutional model with weak supervision for information retrieval. In *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, SIGIR '18, page 985–988, New York, NY, USA. Association for Computing Machinery.

Rabia Nuray and Fazli Can. 2006. Automatic ranking of information retrieval systems using data fusion. *Information processing & management*, 42(3):595–614.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Fiana Raiber and Oren Kurland. 2013. Ranking document clusters using markov random fields. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, page 333–342, New York, NY, USA. Association for Computing Machinery.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Haggai Roitman. 2018. Utilizing pseudo-relevance feedback in fusion-based retrieval. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '18, pages 203–206, New York, NY, USA. ACM.

Xing Wei and W. Bruce Croft. 2006. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 178–185, New York, NY, USA. Association for Computing Machinery.

C. Zhai. 2009. Statistical language models for information retrieval. Morgan Claypool Publsihers.

Kaitao Zhang, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2020. Selective weak supervision for neural information retrieval. In *Proceedings of The Web Conference 2020*, WWW '20, page 474–485, New York, NY, USA. Association for Computing Machinery.