# Unsupervised FAQ Retrieval with Question Generation and BERT

**Yosi Mass, Boaz Carmeli, Haggai Roitman and David Konopnicki**
IBM Research AI
Haifa University, Mount Carmel, Haifa, HA 31905, Israel
{yosimass,boazc,haggai,davidko}@il.ibm.com

## Abstract

We focus on the task of Frequently Asked Questions (FAQ) retrieval. A given user query can be matched against the questions and/or the answers in the FAQ. We present a fully unsupervised method that exploits the FAQ pairs to train two BERT models. The two models match user queries to FAQ answers and questions, respectively. We alleviate the missing labeled data of the latter by automatically generating high-quality question paraphrases. We show that our model is on par and even outperforms supervised models on existing datasets.

## 1 Introduction

Many websites and online communities publish FAQ to help their users find relevant answers to common questions. An FAQ consists of pairs of questions and answers $\{(q, a)\}$. The FAQ retrieval task involves ranking $\{(q, a)\}$ pairs for a given user query $Q$.[1] Searching over FAQ can leverage multifield indexing and retrieval (Karan and Šnajder, 2016). Hence, a user query $Q$ may be matched with either the question field $q$, the answer field $a$ or the concatenated field $q$+$a$ (Karan and Šnajder, 2016).

The association of questions to answers in the FAQ pairs, can be utilized as weak supervision, for training neural models to predict the similarity between user queries and answers (i.e., $Q$-to-$a$ matching) (Gupta and Carvalho, 2019; Karan and Šnajder, 2018; Sakata et al., 2019). However, FAQ pairs by themselves do not provide the required labeled data for training a model to predict the association between user queries and FAQ questions (i.e., $Q$-to-$q$ matching). Thus, a labeled dataset with user queries $Q$ and their matching $\{(q, a)\}$ pairs is required for supervised learning (Gupta and Carvalho, 2019; Karan and Šnajder, 2018; Sakata et al., 2019). Such a dataset is usually manually generated or obtained from query-log mining. Yet, the construction of such a dataset either requires domain expertise (e.g., enriching the dataset with manually generated question paraphrases (Karan and Šnajder, 2018)) or assumes the availability of query-logs (Kim and Seo, 2006, 2008).

Whenever such a dataset is unavailable, one must resort to utilizing unsupervised retrieval models for Q-to-q matching. Previous unsupervised FAQ retrieval models (Burke et al., 1997; Brill et al., 2002; Karan et al., 2013; Karan and Šnajder, 2018; Wu et al., 2005) have utilized so far "traditional" information retrieval techniques, such as lexical and semantic text matching, query expansion, etc.

In this paper we overcome the aforementioned unsupervised gap, by using distant supervision to train neural models. Our method is composed of a combination of three unsupervised methods. Each method is utilized for re-ranking an initial pool of FAQ pairs obtained by a simple BM25 retrieval (Robertson and Zaragoza, 2009). The first method applies a focused-retrieval approach, utilizing passages for answer re-ranking (Bendersky and Kurland, 2008). Each one of the two other methods fine-tunes a BERT model (Devlin et al., 2019), one for matching $Q$-to-$a$ and one for matching $Q$-to-$q$.

To overcome the lack of training data in the latter's case, we further implement a novel weak-supervision approach using automatically generated question paraphrases, coupled with smart filtering to ensure high-quality paraphrases. We then combine the outcome of the three methods using an unsupervised late-fusion method. Overall, we show that our unsupervised FAQ retrieval approach is on par and sometimes even outperforms state-of-the-art supervised models.

[1] Throughout this paper we use the term "question" ($q$) to denote a question within a given FAQ pair, and "query" ($Q$) to denote an issued user query.



*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 807–812

## 2 Related work

Several previous works have also utilized Deep Neural Networks (DNN) for FAQ retrieval. (Karan and Šnajder, 2016) used Convolution Neural Networks (CNN) for matching user queries to FAQ. (Gupta and Carvalho, 2019) used combinations of Long Short-Term Memory (LSTM) to capture $Q$-to-$q$ and $Q$-to-$a$ similarities. Yet, those works are supervised and use user queries ($Q$) for training.

Following the success of BERT (Devlin et al., 2019) in NLP tasks, (Sakata et al., 2019) have recently used a search engine for $Q$-to-$q$ matching and then combined its results with a supervised BERT model for $Q$-to-$a$ matching. We use a similar BERT model for $Q$-to-$a$ matching, but differently from (Sakata et al., 2019), we use it in an unsupervised way, and we further introduce a second unsupervised BERT model for $Q$-to-$q$ matching.

A somewhat related area of research is Community Question Answering (CQA) (Patra, 2017; Zhou et al., 2015) and the related TREC tracks.[2][3] While CQA shares some common features to FAQ retrieval, in CQA there are additional signals such as votes on questions and answers, or the association of user-answer and user-question. Clearly, in a pure FAQ retrieval setting, such auxiliary data is unavailable. Hence, we refrain from comparing with such works.

## 3 Unsupervised FAQ Retrieval Approach

Our proposed FAQ retrieval approach uses distant supervision to train neural models and is based on an initial candidates retrieval followed by a re-ranking step.

Recall that, the FAQ dataset is composed of $\{(q, a)\}$ pairs. The initial candidate retrieval is based on indexing $\{(q, a)\}$ pairs into a search engine index (Section 3.1) and searching against the index. The re-ranking step combines three unsupervised re-rankers. The first one (Section 3.2) is based on a focused-retrieval approach, utilizing passages for answer re-scoring. The two other re-rankers fine-tune two independent BERT models.

The first BERT model (Section 3.3), inspired by (Sakata et al., 2019), is fine-tuned to match questions ($q$) to answers ($a$). At run time, given a user query $Q$, this model re-ranks top-k $\{(q, a)\}$ candidate pairs by matching the user query $Q$ to the answers ($a$) only.

The second BERT model (Section 3.4) is designed to match user queries to FAQ questions. Here, we utilize weak-supervision for generating high quality question paraphrases from the FAQ pairs. The BERT model is fine-tuned on the questions and their generated paraphrases. At run time, given a user query $Q$, this model gets the top-k $\{(q, a)\}$ candidate pairs and re-ranks them by matching the user query $Q$ to the questions ($q$) only.

The final re-ranking is obtained by combining the three re-rankers using an unsupervised late-fusion step (Section 3.5). The components of our method are described in the rest of this section.

### 3.1 Indexing and initial candidates retrieval

We index the FAQ pairs using the ElasticSearch[4] search engine. To this end, we represent each FAQ pair $(q, a)$ as a multifield document having three main fields, namely: question $q$, answer $a$, and the concatenated field $q$+$a$. Given a user query $Q$, we match it (using BM25 similarity (Robertson and Zaragoza, 2009)) against the $q$+$a$ field[5] and retrieve an initial pool of top-k FAQ candidates.

### 3.2 Passage-based re-ranking

Our first unsupervised re-ranker applies a focused retrieval approach. To this end, following (Bendersky and Kurland, 2008), we re-rank the candidates using a *maximum-passage* approach. Such an approach is simply implemented by running a sliding window (i.e., passage) on each candidate's $q$+$a$ field text, and scoring the candidate according to the passage with the highest BM25 similarity to $Q$ (Gry and Largeron, 2011). We hereinafter term this first re-ranking method as `bm25-maxpsg`.

### 3.3 BERT model for $Q$-to-$a$ similarity

Among the two BERT (Devlin et al., 2019) re-rankers, the first one, `BERT-Q-a`, aims at re-ranking the candidate FAQ pairs $\{(q, a)\}$ according to the similarity between a given user query $Q$ and each pair's answer $a$.

To this end, we fine-tune the BERT model from the FAQ pairs $\{(q, a)\}$, using a triplet network (Hoffer and Ailon, 2015). This network is adopted for BERT fine-tuning (Mass et al., 2019) using triplets $(q, a, a')$, where $(q, a)$ constitutes an FAQ pair and $a'$ is a negative sampled answer as

[2] http://alt.qcri.org/semeval2016/task3/
[3] http://alt.qcri.org/semeval2017/task3/
[4] https://www.elastic.co/
[5] Searching only the $q$ or $a$ fields obtained inferior results

follows. For each question $q$ we have positive answers $\{a_i\}$ from all the pairs $\{(q, a_i)\}$.[6] Negative examples are randomly selected from those FAQ that do not have $q$ as their question. To further challenge the model into learning small nuances between close answers, instead of sampling the negative examples from all FAQ pairs, we run $q$ against the $q$+$a$ field of the search index (from Section 3.1 above). We then sample only among the top-k (e.g., $k = 100$) retrieved pairs, that do not have $q$ as their question.

Our `BERT-Q-a` is different from that of (Sakata et al., 2019) in two aspects. First, (Sakata et al., 2019) fine tunes a BERT model for Q-to-a matching using both FAQ $(q, a)$ pairs as well as user queries and their matched answers $(Q, a)$. This is, therefore, a supervised setting, since user queries are not part of the FAQ and thus require labeling efforts. Compared to that, we fine tune the `BERT-Q-a` using only FAQ $(q, a)$ pairs. Second, unlike (Sakata et al., 2019), which fine-tunes BERT for a classification task (i.e., point-wise training) we train a triplet network (Hoffer and Ailon, 2015) that learns the relative preferences between a question and a pair of answers. Our network thus implements a pair-wise learning-to-rank approach (Li, 2011).

At inference time, given a user query $Q$ and the top-k retrieved $(q, a)$ pairs, we re-rank the $(q, a)$ pairs using the score of each $(Q, a)$ pair as assigned by the fine-tuned `BERT-Q-a` model (Mass et al., 2019).

### 3.4 BERT model for $Q$-to-$q$ similarity

The second BERT model, `BERT-Q-q`, is independent from the first `BERT-Q-a` model (Section 3.3) and is trained to match user queries to FAQ questions. To fine-tune this model, we generate a weakly-supervised dataset from the FAQ pairs. Inspired by (Anaby-Tavor et al., 2019), we fine-tune a generative pre-training (GPT-2) neural network model (Radford, 2018) for generating question paraphrases. GPT-2 is pre-trained on huge bodies of text, capturing the natural language structure and producing deeply coherent text paragraphs.

Intuitively, we would like to use the FAQ answers to generate paraphrases to questions. Unlike the work of (Anaby-Tavor et al., 2019) which fine tunes a GPT-2 model given classes, where each class has a title and several examples, here we consider each answer $a$ as a class with only one example which is its question $q$.

We thus concatenate all the FAQ pairs into a long text $U = a_1$ `SEP` $q_1$ `EOS` $\cdots$ $a_n$ `SEP` $q_n$ `EOS`, where answers precede their questions,[7] having `EOS` and `SEP` as special tokens. The former separates between FAQ pairs and the latter separates answers from their questions inside the pairs.

The GPT-2 fine-tuning samples a sequence of $l$ consecutive tokens $w^{j-l}, \cdots, w^j$ from $U$ and maximizes the conditional probability $\mathbf{P}(w^j | w^{j-l}, \ldots, w^{j-1})$ of $w^j$ to appear next in the sequence. We repeat this process several times.

Once the model is fine-tuned, we feed it with the text "$a$ `SEP`", ($a$ is an answer in an FAQ pair $(q, a)$), and let it generate tokens until `EOS`. We take all generated tokens until `EOS`, as a paraphrase to $a$'s question $q$. By repeating this generation process we may generate any number of question paraphrases. For example, the paraphrase "*Is there a way to deactivate my account on Facebook?*" was generated for the question "*How do I delete my Facebook account?*".

One obstacle in using generated text is the noise it may introduce. To overcome this problem we apply a filtering step as follows. The idea is to keep only paraphrases that are semantically similar to their original question (i.e., have similar answers). Let $GT(q)$=$\{(q, a_i)\}$ be the FAQ pairs of question $q$ (i.e., the ground truth answers of $q$). For each generated paraphrase $p$ of $q$, we run $p$ as a query against the FAQ index (See section 3.1), and check that among the returned top-$k$ results, there are at least $min(n, |GT(q)|)$ pairs from $GT(q)$ for some $n$. In the experiments (see Section 4 below) we used $k$=10 and $n$=2.

To select the best paraphrases for each question $q$, we further sort the paraphrases that passed the above filter, by the score of their top-1 returned $(q, a)$ pair (when running each paraphrase $p$ as a query against the FAQ index). The motivation is that a higher score of a returned $(q, a)$ for a query $p$, implies a higher similarity between $p$ and $q$. [8]

Similar to the `BERT-Q-a`, this model is fine-tuned using triplets $(p, q, q')$, where $p$ is a paraphrase of $q$ and $q'$ is a randomly selected question

[6] Usually $i = 1$, i.e., there is a single answer for each FAQ question $q$. Yet, it is possible that $i > 1$.

[7] FAQ questions with more than one answer are treated here as different questions.

[8] The filtered paraphrases can be downloaded from https://github.com/YosiMass/faq-retrieval

from the FAQ questions. At inference time, given a user query $Q$ and the top-k retrieved $(q, a)$ pairs, we re-rank the answers $(q, a)$ answers, using the score of each $(Q, q)$ pair as assigned by the fine-tuned `BERT-Q-q` model (Mass et al., 2019).

### 3.5 Re-rankers combination

We combine the three re-ranking methods (i.e., `bm25-maxpsg` and the two fined-tuned BERT models) using two alternative late-fusion methods. The first one, `CombSUM` (Kurland and Culpepper, 2018), calculates a combined score by summing for each candidate pair the scores that were assigned to it by the three re-ranking methods.[9]

Following (Roitman, 2018), as a second alternative, we implement the `PoolRank` method. PoolRank first ranks the candidate pairs using `CombSUM`. The top pairs are then used to introduce an unsupervised query expansion step (RM1 model (Lavrenko and Croft, 2001)) which is used to re-rank the whole candidates pool. [10]

## 4 Experiments

### 4.1 Datasets

We use two FAQ datasets in our evaluation, namely: FAQIR (Karan and Šnajder, 2016)[11] and StackFAQ (Karan and Šnajder, 2018).[12] The FAQIR dataset was derived from the "*maintenance & repair*" domain of the Yahoo! Answers community QA (CQA) website. It consists of 4313 FAQ pairs and 1233 user queries. The StackFAQ dataset was derived from the "*web apps*" domain of the StackExchange CQA website. It consists of 719 FAQ pairs (resulted from 125 threads; some questions have more than one answer) and 1249 user queries.

### 4.2 Baselines

On both datasets, we compare against the results of the various methods that were evaluated in (Karan and Šnajder, 2018), namely: `RC` – an ensemble of three unsupervised methods (BM25, Vector-Space and word-embeddings); `ListNet` and `LambdaMART` – two (supervised) learning-to-rank methods that were trained over a diverse set of text similarity features; and `CNN-Rank` – a (supervised) learning-to-rank approach based on a convolutional neural network (CNN).

On the StackFAQ dataset, we further report the result of (Sakata et al., 2019), which serves as the strongest supervised baseline. This baseline combines two methods: `TSUBAKI` (Shinzato et al., 2008) – a search engine for $Q$-to-$q$ matching; and a supervised fine-tuned `BERT` model for $Q$-to-$a$ matching. We put the results of this work (that were available only on the StackFAQ dataset), just to emphasize that our approach can reach the quality of a supervised approach, and not to directly compare with it.

### 4.3 Experimental setup

We used ElasticSearch to index the FAQ pairs. For the first ranker (Section 3.1) we used a sliding window of size 100 characters with 10% overlap. For fine-tuning the `BERT-Q-a` model, we randomly sampled 2 and 5 negative examples for each positive example $(q, a)$ on FAQIR and StackFAQ datasets, respectively.

To fine-tune GPT-2 for generating the question paraphrases (Section 3.4), we segmented $U$ into consecutive sequences of $l = 100$ tokens each. We used OpenAI's Medium-sized GPT-2 English model: 24-layer, 1024-hidden, 16-heads, 345M parameters. We then used the fine-tuned model to generate 100 paraphrases for each question $q$ and selected the top-10 that passed filtering (as described in Section 3.4). Overall on FAQIR, 22,736 paraphrases passed the filter and enriched 3,532 out of the 4,313 questions. On StackFAQ, 856 paraphrases passed the filter and enriched 109 out of the 125 thread questions. Similar to the `BERT-Q-a` fine-tuning, we selected 2 and 5 negative examples for each $(p, q)$ (paraphrase-question) pair on FAQIR and StackFAQ, respectively.

The two BERT models used the pre-trained BERT-Base-Uncased model (12-layer, 768-hidden, 12-heads, 110M parameters). Fine-tuning was done with a learning rate of 2e-5 and 3 training epochs. Similar to previous works, we used the following metrics: P@5, Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR), calculated on an initial candidate list of 100 FAQs retrieved by the search engine using standard BM25.

[9] Each re-ranker's scores are first max-min normalized.

[10] Further following (Roitman, 2018), we use the normalized `CombSUM` fusion scores as the weak-relevance labels for the RM1 model estimation.

[11] http://takelab.fer.hr/data/faqir/

[12] http://takelab.fer.hr/data/StackFAQ

### 4.4 Results

Table 1 reports the results for the two datasets.[13] We compare the base BM25 retrieval (`bm25(q+a)`), our three proposed unsupervised re-ranking methods (`bm25-maxpsg`, `BERT-Q-a` and `BERT-Q-q`) and their fusion-based combinations (`CombSUM` and `PoolRank`) with the state-of-the-art unsupervised and supervised baselines. We also compare to `PoolRank+`, which is same as `PoolRank` except that the two BERT models (i.e., `BERT-Q-a` and `BERT-Q-q`) are fine-tuned on the union of the respective training sets of both the FAQIR and StackFAQ datasets.

We observe that, among our three re-rankers, `BERT-Q-q` was the best. For example, on FAQIR it achieved 0.67, 0.61 and 0.90 for P@5, MAP and MRR, respectively. This in comparison to 0.54, 0.50 and 0.81, obtained by `bm25-maxpsg` for P@5, MAP and MRR, respectively. This confirms previous findings (Karan and Šnajder, 2016), that Q-to-q matching gives the best signal in FAQ retrieval. Furthermore, on both datasets, the fusion methods achieved better results than the individual re-rankers, with better performance by the `PoolRank` variants over `ComboSum`.

An exception is FAQIR, where `BERT-Q-q` achieved same results as the `ComboSUM` fusion. As mentioned above, `BERT-Q-q` has a significantly better performance on FAQIR than the other two individual rankers, thus a simple fusion method such as `CombSUM` can not handle such cases well. PoolRank, which uses relevance model, is a better approach and thus gives better fusion results.

Further comparing with the baselines, we can see that, on FAQIR, our unsupervised `PoolRank` outperformed all other methods; including the supervised methods on all three metrics. On StackFAQ, `PoolRank` outperformed all other methods, except the supervised `TSUBAKI+BERT` (Sakata et al., 2019). We note that, our unsupervised results `PoolRank+` achieved (0.75, 0.88 and 0.90 for P@5, MAP and MRR, respectively), which is quite close to the supervised results (0.78, 0.90 and 0.94 respectively) of (Sakata et al., 2019).

[13] Similar to (Karan and Šnajder, 2018), the FAQIR initial retrieval is done against a subset of 789 FAQ pairs that are relevant to at least one user query.

| **FAQIR** | **P@5** | **MAP** | **MRR** |
|---|---|---|---|
| bm25 (q+a) | 0.48 | 0.44 | 0.74 |
| bm25-maxpsg | 0.54 | 0.50 | 0.81 |
| BERT-Q-a | 0.53 | 0.46 | 0.81 |
| BERT-Q-q | 0.67 | 0.61 | **0.90** |
| CombSUM | 0.67 | 0.61 | **0.90** |
| PoolRank | **0.69** | **0.62** | 0.88 |
| PoolRank+ | **0.69** | **0.62** | 0.88 |
| RC | 0.58 | 0.53 | 0.80 |
| ListNet | 0.57 | 0.53 | 0.80 |
| LambdaMART | 0.61 | 0.57 | 0.84 |
| CNN-Rank | 0.66 | 0.58 | 0.85 |

| **StackFAQ** | **P@5** | **MAP** | **MRR** |
|---|---|---|---|
| bm25 (q+a) | 0.56 | 0.67 | 0.79 |
| bm25-maxpsg | 0.63 | 0.75 | 0.81 |
| BERT-Q-a | 0.54 | 0.63 | 0.81 |
| BERT-Q-q | 0.68 | 0.82 | 0.80 |
| CombSUM | 0.72 | 0.85 | 0.91 |
| PoolRank | 0.74 | 0.87 | 0.88 |
| PoolRank+ | 0.75 | 0.88 | 0.90 |
| RC | 0.52 | 0.63 | 0.8 |
| ListNet | 0.51 | 0.54 | 0.70 |
| LambdaMART | 0.60 | 0.74 | 0.84 |
| CNN-Rank | 0.62 | 0.74 | 0.84 |
| TSUBAKI+BERT | **0.78** | **0.9** | **0.94** |

Table 1: Evaluation results

## 5 Summary and Conclusions

We presented a fully unsupervised method for FAQ retrieval. The method is based on an initial retrieval of FAQ candidates followed by three re-rankers. The first one is based on an IR passage retrieval approach, and the others two are independent BERT models that are fine-tuned to predict query-to-answer and query-to-question matching. We showed that we can overcome the "unsupervised gap" by generating high-quality question paraphrases and use them to fine-tune the query-to-question BERT model. We experimentally showed that our unsupervised method is on par and sometimes even outperforms existing supervised methods.

## References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2019. Not enough data? deep learning to the rescue! *CoRR*, abs/1911.03118.

Michael Bendersky and Oren Kurland. 2008. Utilizing passage-based language models for document retrieval. In *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*, ECIR'08, pages 162–174, Berlin, Heidelberg. Springer-Verlag.

Eric Brill, Susan Dumais, and Michele Banko. 2002. An analysis of the askmsr question-answering system. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 257–264, Stroudsburg, PA, USA. Association for Computational Linguistics.

Robin Burke, Kristian Hammond, Vladimir Kulyukin, Steven Lytinen, Noriko Tomuro, and Scott Schoenberg. 1997. Question answering from frequently asked question files: Experiences with the FAQ FINDER system. *AI Magazine*, 18:57–66.

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.

Sparsh Gupta and Vitor R. Carvalho. 2019. FAQ retrieval using attentive matching. In *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 929–932, New York, NY, USA. ACM.

Mathias Gry and Christine Largeron. 2011. BM25t: A BM25 extension for focused information retrieval. *Knowledge and Information Systems - KAIS*, 32:1–25.

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*.

Mladen Karan and Jan Šnajder. 2016. FAQIR - a frequently asked questions retrieval test collection. In *Text, Speech, and Dialogue (TSD)*, volume 9924. Springer.

Mladen Karan and Jan Šnajder. 2018. Paraphrase-focused learning to rank for domain-specific frequently asked questions retrieval. *Expert Systems with Applications*, 91:418 – 433.

Mladen Karan, Lovro Žmak, and Jan Šnajder. 2013. Frequently asked questions retrieval for Croatian based on semantic textual similarity. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 24–33, Sofia, Bulgaria. Association for Computational Linguistics.

Harksoo Kim and Jungyun Seo. 2006. High-performance faq retrieval using an automatic clustering method of query logs. *Information Processing and Management*, 42(3):650 – 661.

Harksoo Kim and Jungyun Seo. 2008. Cluster-based faq retrieval using latent term weights. *IEEE Intelligent Systems*, 23(2):58–65.

Oren Kurland and J. Shane Culpepper. 2018. Fusion in information retrieval: Sigir 2018 half-day tutorial. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '18, pages 1383–1386, New York, NY, USA. ACM.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 120–127, New York, NY, USA. ACM.

Hang Li. 2011. A short introduction to learning to rank. *IEICE Transactions*, 94-D:1854–1862.

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.

Barun Patra. 2017. A survey of community question answering. *CoRR*, abs/1705.04009.

Alec Radford. 2018. Improving language understanding by generative pre-training.

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.

Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. 2019. FAQ retrieval using query-question similarity and bert-based query-answer relevance. *CoRR*, abs/1905.02851.

Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. 2008. TSUBAKI: An open search engine infrastructure for developing new information access methodology. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Chung-Hsien Wu, Jui-Feng Yeh, and Ming-Jun Chen. 2005. Domain-specific faq retrieval using independent aspects. *ACM Transactions on Asian Language Information Processing*, 4(1):117.

Xiaoqiang Zhou, Baotian Hu, Qingcai Chen, Buzhou Tang, and Xiaolong Wang. 2015. Answer sequence learning with neural networks for answer selection in community question answering. *CoRR*, abs/1506.06490.