Sequence-to-Sequence Learning on Keywords for Efficient FAQ Retrieval

Sourav Dutta , Haytham Assem , Edward Burgin
Huawei Research Centre, Dublin, Ireland
{sourav.dutta2, haytham.assem, edwardburgin}@huawei.com,

Abstract

Frequently-Asked-Question (FAQ) retrieval provides an effective procedure for responding to user's natural language based queries. Such platforms are becoming common in enterprise chatbots, product question answering, and preliminary technical support for customers. However, the challenge in such scenarios lies in bridging the *lexical and semantic gap* between varied query formulations and the corresponding answers, both of which typically have a *very short span*.

This paper proposes *TI-S2S*, a novel learning framework combining TF-IDF based keyword extraction and Word2Vec embeddings for training a Sequence-to-Sequence (Seq2Seq) architecture. It achieves high precision for FAQ retrieval by better understanding the underlying *intent* of a user question captured via the representative keywords. We further propose a variant with an additional neural network module for guiding retrieval via relevant candidate identification based on similarity features. Experiments on publicly available dataset depict our approaches to provide around 92% precision-at-rank-5, exhibiting nearly 13% improvement over existing approaches.

1 Introduction

Frequently-Asked-Questions (FAQ) provide a collection of question-answer pairs that are either manually created or automatically extracted from relevant documents. FAQ provide users with an "one-stop" source for the most relevant or most searched information pertaining to a product or service – to enable prompt customer help for general queries.

Motivation. FAQ retrieval systems provide a natural language interface for querying an FAQ collection, and is thus increasingly becoming popular with large-scale service-providing companies for presenting information to customers. Such systems provide two-fold advantages: (i) automation of customer service tasks, e.g., intelligent chatbots [Massaro *et al.*, 2018; Yan *et al.*, 2016] and automated e-mail answering [Karan and Šnajder, 2018; Malik *et al.*, 2007], and (ii) enable efficient access to internal FAQ documents for customer service agents, increasing the quality and efficiency. Further, with the advent of personal assistants (like XiaoIce, Siri, Alexa, Google Assistant, etc.), these "virtual agents" can provide answers to questions and help users solve routine tasks by an additional channel to FAQs, hotlines, and forums – enabling a natural interaction with users [Lommatzsch and Katins, 2019; Santos *et al.*, 2020].

Challenges. FAQ retrieval is a challenging task, majorly attributed to the fact that the question-answer texts are short, making it harder to bridge the *lexical and semantic gap* between a user query and the FAQ questions due to short span with limited context [Karan and Šnajder, 2018; Lee *et al.*, 2008]. Further, precise understanding of user questions can be difficult due to informal representations, domain-specificity, abbreviations, and formal-colloquial term mismatches [Lommatzsch and Katins, 2019]. For example, consider the questions ``How can I seal a hole in the gas tank of my car?'' and ``How to patch a leak in the fuel compartment of my van?'' which are semantically matching but exhibit low lexical overlap and formal-colloquial mismatch. In addition, FAQ retrieval systems should be able to handle both keyword as well as *short span* "natural language" questions. Given the predominantly "customer-centric" nature, such systems generally demand higher precision and interpretability compared to traditional information retrieval methods.

Problem Statement. The task of *FAQ Retrieval* entails the efficient ranking (in terms of relevance) of question-answer pairs of a collection, in response to a user input query. In other words, such retrieval engines attempt to understand the underlying *intent* of users and retrieve the most related answers containing the correct information [Kothari *et al.*, 2009].

Formally, consider a pre-curated collection (or repository) of question-answer (Q, A) pairs to comprise the FAQ $= \{(Q_1, A_1), \cdots, (Q_n, A_n)\}$, where Q_i denotes a question related to the domain, and A_i represents the

corresponding answer. Given a user query q, the task then is to return an ordered list of FAQ (Q, A) pairs, $\{(Q_1^q, A_1^q), \cdots, (Q_k^q, A_k^q)\}$, depicting high semantic and intent similarity with respect to the input user query q.

Contributions. This work proposes *TI-S2S*, a novel keyword based supervised learning framework for efficient FAQ retrieval. Our approach leverages *sequence-to-sequence* model to generate representative labels for user questions to compute query-question similarity. Additionally, a variant incorporating "candidate" identified using a deep learning architecture to guide the retrieval process is shown to further improve performance.

We show that our proposed framework efficiently captures: (i) domain-specificity of the application, (ii) characteristic words and / or concepts to differentiate between questions, and (iii) semantic similarity for retrieving relevant QA pairs from the FAQ collection.

Experiments on public FAQ dataset depict our framework to outperform existing techniques in terms of accuracy, and also in robustness to limited training data. In effect, it implicitly considers both *document redundancy* and *query redundancy* [Karan and Šnajder, 2018].

2 Related Work

The problem of FAQ retrieval lies at the intersection of information retrieval and question answering and have thus been studied using techniques from both the fields. Initial works on FAQ retrieval relied on manual feature engineering based on text similarities using parsing, edit distance, TF-IDF measures, longest common subsequence [Kothari *et al.*, 2009], match-template construction [Sneiders, 2010], and statistical approaches [Berger *et al.*, 2000] to name a few. The use of both query-question and query-answer vector space similarities within a ranking model was studied in [Jijkoun and de Rijke, 2005]. However observe, over-emphasis on query-answer similarity would be inefficient in scenarios where significant parts of different answers might be similar. For example, answers to both the questions ``How to add an account photo?'' and ``How to change the account name?'' might possibly share the common snippet ``Go to Account > Setting > Profile ...'' or similar. Further, answers might change depending on updates to processes and manuals which might necessitate costly re-training of the entire framework. Such scenarios might degrade the performance of approaches based on query-answer similarities [Otsuka *et al.*, 2018; Sakata *et al.*, 2019]. Thus, in our setting, we do not consider the answer to form a part of the retrieval process.

Contextualized language models like BERT [Devlin *et al.*, 2019] have been shown to capture semantic relatedness, and such embedding techniques have been coupled with traditional IR techniques for FAQ Retrieval [Sakata *et al.*, 2019]. The use of knowledge graphs have also been studied for Question-Answering (Q-A), by use of entity-concept "anchors" in this context [Xie *et al.*, 2019]. Deep Learning has recently enjoyed significant success in classification tasks by constructing high-dimensional latent feature space. A neural network with word embeddings was proposed in [Yan *et al.*, 2016], while a convolution neural network (CNN) based learning-to-rank module was presented in [Karan and Šnajder, 2018]. However, supervised methods require large FAQ-collection with annotations, which are expensive. Hence, in practice such annotated datasets are usually too small to meaningfully train complicated machine learning models. Further, such models tend to face difficulty in handling long-tailed questions. To tackle the problem of limited context in FAQ systems attention-based deep learning models [Gupta and Carvalho, 2019], query expansion [Otsuka *et al.*, 2018] and query generation [Mass *et al.*, 2020] have recently been studied. Document ranking via sequence-to-sequence has also been studied [Nogueira *et al.*, 2020].

Community and non-factoid question answering (CQA) [Surdeanu *et al.*, 2011; Figueroa, 2017] are closely related, but involve larger corpus with broader scope, and hence is not directly applicable to FAQ Retrieval, typically with context brevity and smaller training data.

3 *TI-S2S* Framework

This section describes the working of our proposed *TF-IDF Induced Sequence-to-Sequence* (TI-S2S) algorithm for efficient FAQ Retrieval. It couples TF-IDF score to extract keywords (modeling intents in user queries) and word embeddings (capturing semantic similarity among questions) for learning a *sequence-to-sequence* model to transform syntactically different but semantically similar questions into a *common representative sequence*.

Given an FAQ collection (set of question-answer (QA) pairs), our *TI-S2S* framework hinges on the following:

A. Pre-Processing. The questions in the input FAQ collection are initially pre-processed to remove stopwords and are lemmatized. For each question Q_i, several variations of the question are created (either manually or by automated paraphrasing techniques) or are extracted (from query logs via duplicate detection or similarity measures). Such semantically similar paraphrased questions are added to the FAQ and are annotated to depict that they convey the same user information intent. As proposed in [Karan and Šnajder, 2018], the paraphrased QA pairs in FAQ along with the relevance annotations are used for supervised training. Index structures storing the relevance information between questions are constructed to assist the subsequent modules.

B. Intent Target Keyword Learning. Based on the relevance annotations among the questions in the FAQ, *TI-S2S* creates groups or clusters of questions that are semantically similar to (or paraphrases of) each other. For each group of such similar questions (or annotated paraphrased variants) in the FAQ, we extract words that have TF-IDF score [Aizawa, 2003] (computed on the entire FAQ collection) greater than a thresholding parameter τ, denoted as *intent target keywords*. Intuitively, these intent keywords capture the context and topic of the question groups. Hence, these *intent keywords* enable a "common representative sequence" for

each group of similar questions (refer Table 2 for example), providing cues for *weak supervision* in training the subsequent modules.

C. Seq2Seq Learning. A sequence-to-sequence (Seq2Seq) model [Sutskever *et al.*, 2014] utilizes an encoder-decoder architecture for learning to transform an input sequence to a corresponding output sequence (possibly of differing lengths). *TI-S2S* uses a Seq2Seq module to *learn to transform* a question $Q_i \in FAQ$ (i.e., a sequence of pre-processed words) to the *representative intent target keyword sequence* associated with the question group to which Q_i belongs to. Word embeddings of the questions (using Word2Vec [Mikolov *et al.*, 2013]) are fed to the input layer of the Seq2Seq module for training with *teacher-forcing* technique [Bengio *et al.*, 2015] and Luong attention mechanism [Luong *et al.*, 2015].

It is interesting to note that the transformation of questions into a common keyword space bridges the lexical gap, while the use of word embeddings (of the question) bridges the semantic gap. For example, both the words 'image' and 'photo' (similar in the embedding space) in different questions would be trained to generate the same output word 'picture' (a common representative keyword) from the seq2seq module – addressing the lexical and semantic gap between user and FAQ questions with short spans.

D. Translated FAQ. The above trained Seq2Seq model is then used to transform the questions in FAQ to *intent representative format*. That is, this module translates the input FAQ into a collection of 3-tuples $\{(Q_i, \overline{Q}_i, A_i)\}$ – where $(Q_i, A_i) \in$ FAQ is the original QA pair and $\overline{Q}_i$ is the *predicted intent keyword sequence* for Q_i obtained from the Seq2Seq module.

Ideally, the predicted $\overline{Q}_i$ should be the same as the intent target keywords (provided during training) associated to the question group to which Q_i (and other similar or paraphrased questions) belongs to. However, in practice, training losses and presence of noise might lead to deviations. Through this, *TI-S2S* aims to minimize the impact of such error propagation to the final phase.

E. FAQ Retrieval. The trained *TI-S2S* framework along with the translated FAQ forms the proposed FAQ Retrieval platform for user queries. Given a new user query q, it is initially pre-processed and its word embeddings (as in *Modules A* and *C* above) are provided as input to *TI-S2S*. The "predicted intent target keyword sequence" ($\overline{q}$) from the Seq2Seq module is then compared with all $\overline{Q}_i$ in the translated FAQ. A similarity score between $\overline{q}$ and $\overline{Q}_i$ is used to obtain the final ranked list of QA pairs of the FAQ. To capture syntactic and semantic similarity between the keyword sequences, we use the average of Word Mover's Distance [Kusner *et al.*, 2015] and Levenstein distance between $\overline{q}$ and $\overline{Q}_i$.

Since, the final stage uses Word Mover's Distance and Levenstein distance to compute the similarities between the representative sequences (treated as bag-of-significant-words), the order of the predicted representative sequence (obtained from the Seq2Seq module) is not important and our framework is not sensitive to it. This provides flexibility to our framework and does not enforce strict order in the seq2seq generation process. Further, the modular structure of our framework enables it to be easily adapted to diverse application scenarios with algorithmic variants – attention mechanisms for Seq2Seq learning or combinations of different similarity measures.

3.1 *GTI-S2S* Variant

We now present *Guided TF-IDF Induced Sequence-to-Sequence* (GTI-S2S), a variant of the *TI-S2S* framework to cater to scenarios with high domain-specificity and noisy training process. *GTI-S2S* (along with Seq2Seq module) employs an additional recurrent neural network (RNN) to learn to predict *question-question relevance* using features like entity overlap, Levenstein distance and embedding space similarity between two input questions.

Thus, given the groups of similar or paraphrased questions (as discussed in *Module A*), the RNN is trained as a binary classifier to predict if two questions are similar and / or relevant (using the relevance annotations), thus providing *"guided candidate QA selection"* during the retrieval phase of *TI-S2S* framework (Section 3).

Specifically, on arrival of a user query q, the predicted intent target keyword sequence ($\overline{q}$) are generated by *TI-S2S* (as in *Module E*). Additionally, for each question $Q_i \in$ FAQ, *GTI-S2S* now computes its relevance to q (using the above trained RNN module). Based on the predicted classification probabilities, the top-k FAQ questions (with probabilities above a threshold) are extracted as "prime candidates" for the user query. Finally, the similarity scores between the obtained candidate questions' predicted keyword sequence and $\overline{q}$ are computed to obtain the final rank list.

In a nutshell, *GTI-S2S* can be viewed as a two-stage framework: (a) generation of candidates using RNN and (b) use of TI-S2S framework for ranking the candidates. While in the TI-S2S framework, the final similarity of the user question (after prediction phase using seq2seq) is computed against all the questions in the FAQ collection; in GTI-S2S, the final similarity is computed only with the candidates identified from the stage (a). Later in Section 4.1 we show the performance advantages of "candidate generation guided retrieval" in *GTI-S2S* in certain settings.

4 Experimental Results

We now compare the performance of our proposed framework with competing state-of-the-art approaches for FAQ retrieval on open dataset.

Dataset Used. We perform experiments on the publicly available *StackExchange* FAQ dataset [Karan and Šnajder, 2018] (from www.takelab.fer.hr/data/StackFAQ/). It contains 125 QA threads pertaining to popular Web applications, with each thread containing an original query and 10 different manual paraphrasings (annotated as relevant to the original question) – a total of 1375

Table 1: (a) Performance of algorithms on *MAP* and *P@5* measures. (b) Effect of TF-IDF threshold (τ) on MAP and P@5.

(a)

Approaches	MAP	P@5
CNN-Rank [Karan and Šnajder, 2018]	0.74	0.62
TSU-BERT [Sakata *et al.*, 2019]	0.897	0.776
BERT [Devlin *et al.*, 2019]	0.614	0.583
RoBERTa [Liu *et al.*, 2019]	0.712	0.796
SBERT [Reimers and Gurevych, 2019]	0.686	0.774
TI-S2S	*0.929*	*0.92*
GTI-S2S	**0.934**	**0.924**

(b)

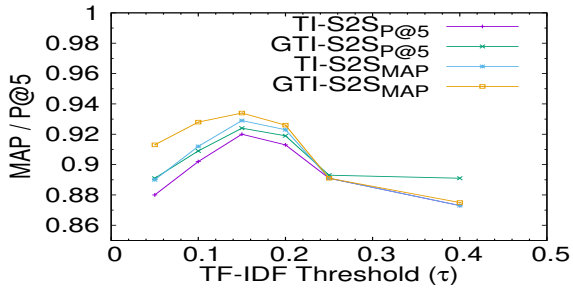

Table 2: Representative intent keywords extracted for different question clusters on StackExchange data.

Sample Questions	$\tau \geq 0.4$	$\tau \geq 0.25$	$\tau \geq 0.15$	$\tau \geq 0.05$
How secure is my sensitive data on dropbox Are there security threats to dropbox Is sensitive data on dropbox secure Does dropbox have good security against attackers How safe is my data on dropbox	dropbox; security;	dropbox; security; threat;	attacker; data; dropbox; secure; security; sensitive; threat;	attacker; concern; data; dropbox; eavesdrop; file; good; information; know; like; malicious; safe; secure; steal; threat; tight; transfer; sensitive; use; user;
Is splitting conversations possible in mail threads on gmail Can i split a conversation in gmail How do I split two merged gmail conversations Splitting conversations in gmail How to divide a conversation on gmail	conversation;	conversation; gmail; split; thread;	gmail; manage; conversation; one; merge; separate; split; thread;	assign; bcc; break; confusing; divide; conversation; easy; gmail; hard; keep; large; mail; manage; merge; people; one; possible; reply; response; small; separate; split; thread; time; track;

(Q, A) pairs. The task is then to return a ranked result of the QA pairs in terms of their relevance to a query, as in the setup of [Karan and Šnajder, 2018; Sakata *et al.*, 2019].

Competing Approaches. We benchmark the performance of our proposed framework against the following methods:
(1) *CNN-Rank* [Karan and Šnajder, 2018] – uses learning-to-rank via convolutional NN architecture.
(2) *TSU-BERT* [Sakata *et al.*, 2019] – combination of TSUBAKI IR engine for computing query-question and BERT based embeddings for query-answer similarities (from github.com/ku-nlp/bert-based-faqir).
(3) *BERT* [Devlin *et al.*, 2019] – bidirectional language representation fine-tuned to capture contextual similarities using cosine score (from github.com/hanxiao/bert-as-service).
(4) *RoBERTa* [Liu *et al.*, 2019] – fine-tuned optimized version of BERT for better contextual similarity computation (from github.com/pytorch/fairseq/blob/master/examples/roberta).
(5) *SBERT* [Reimers and Gurevych, 2019] – Siamese network structure for sentence embeddings using *roberta-large-nli-stsb-mean-tokens*), particularly suitable for FAQ retrieval given the short span of texts (from github.com/UKPLab/sentence-transformers.

Fine-Tuning. The BERT and RoBERTa baselines were fine-tuned on the training data to identify similarities between different text or question representations. The CLS token was used as the overall representation of the input questions. No observable difference was found while using mean pooling strategy.

Evaluation Measures. We evaluate the performance of the algorithms using the following measures:
(i) *Mean Average Precision* (MAP) – computes the mean (over the query set) of average precision using the rank position of relevant QA pairs returned.
(ii) *Precision-at-Rank-5* (P@5) – reports the number of relevant answers among the top-5 retrieved QA pairs, providing a more practical measure as users typically tend to inspect the top few results.

Experimental Setup. We adopt the setup of [Karan and Šnajder, 2018], with 80-20 train-test data split and report the averaged results across five-fold cross-validation runs. Further, for supervised model training, FAQ pairs in the train set were provided with relevance annotations with respect to other questions in the form of *relevance matrix*, i.e., if a FAQ pair (Q_j, A_j) is relevant to $Q_i \in$ FAQ, its corresponding annotation is set to 1 (i.e., the $(ij)^{th}$ element of the relevance matrix is set to 1), otherwise is considered as 0.

The Seq2Seq module of *TI-S2S* consisting of an LSTM model with 2048 encoder nodes, *concat* Luong Attention mechanism [Luong *et al.*, 2015], and a dropout factor of 0.4 as regularizer. The decoder uses a *tanh* activation function optimized for *Sparse Categorical Cross-Entropy* loss function. Additionally, for candidate generation, *GTI-S2S* stacks a Gated Recurrent Unit (GRU) with 1024 units, a 512 node fully-connected layer having *SoftMax* activation, and 0.5 dropout layer. The models are trained with 32 batch size over 30 epochs with TF-IDF threshold τ set to 0.15 (refer Section 4.2), and top-20 candidates were considered in *GTI-S2S*. Publicly

available pre-trained Google Word2Vec embeddings were used. For all algorithms, the input questions were pre-processed to remove stopwords and were lemmatized.

4.1 Overall Results

The obtained performance results of the competing algorithms are presented in Table 1(a). The use of TF-IDF to obtain discriminating words characterizing the different question groups and learning the transformation of questions to representative keywords via a sequence-to-sequence model provide a proxy to understanding the context, topic and intent of the questions. This enables our proposed algorithms *TI-S2S* and *GTI-S2S* to achieve more than **92%** accuracy on both the MAP and P@5 measures. We observe that our framework outperforms the existing approaches with nearly **13**% improvements in terms of *P@5* over RoBERTa, and around 3% better *MAP* score over *TSU-BERT*.

The *GTI-S2S* framework depicts a slight increase in performance over *TI-S2S*, which can be attributed to the "guided candidate selections" from the additional recurrent neural network based learning module. Although the overall gain is marginal for *GTI-S2S*, note that this variant provides robustness against sub-optimal parameter settings or minor prediction errors. For example, in Table 1(b), for a sub-optimal TF-IDF threshold setting (e.g., $\tau = 0.05$) the performance of *GTI-S2S* is still efficient ($\sim$ 91% MAP) compared to *TI-S2S*. Further, as seen in Figure 1, *GTI-S2S* also performs better in scenarios with limited training data availability. Thus, the *GTI-S2S* provides a robust variant of our algorithm.

4.2 Parameter Setting

The working of our proposed *TI-S2S* and *GTI-S2S* algorithms depends on the thresholding hyper-parameter τ on TF-IDF score to extract *representative intent target keywords* characterizing the various contexts presents in the questions. We study the performance of our algorithm (on MAP and P@5) for different values of τ. Figure 1(b) shows a "bell-like" curve with $\tau = 0.15$ (used in our experiments) providing the best empirical results.

For interpretability, we list the target intent keywords identified (for training the sequence-to-sequence module) at different values of τ. As seen in Table 2, a high threshold value extracts only a few *representative intent* words which fail to properly model the full context of the QA pairs. For example, in the second row of Table 2 only ``conversation'' is identified as the representative keyword (with $\tau \geq 0.4$), completely ignoring the vital context of ``gmail''. On the other hand, a very low value of τ is seen to extract non-informative words which possibly overlaps with other QA groups, diminishing the discriminative power of the framework. Both scenarios are seen to degrade the overall accuracy performance of our algorithm. Thus, this parameter captures the domain-specificity and can be suitably tuned for different application domains.

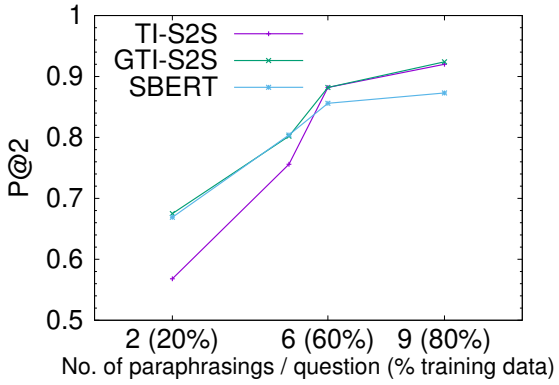

Figure 1: Effect of training size on the performance of the proposed algorithm.

4.3 Robustness Study

A major challenge for supervised systems is the availability of large annotated training data, and the ensuing associated costs. The *StackExchange* dataset used also contains 10 manual paraphrasing for each original user questions. In this regard, we now study the robustness of our approach in presence of limited training data, by varying the number of relevance training questions (paraphrasings with same meaning) provided for each question. We compare the performance on P@2 (the smallest training subset has only 2 variants per QA) with SBERT (demonstrating the best P@2 results on the full dataset). From Figure 1, we observe that *GTI-S2S* can robustly handle limited supervision scenarios, demonstrating a graceful degradation with performance similar to SBERT ($\sim$ 68% accuracy) with only 2 training examples for each of the 125 QA threads. However, the accuracy of *TI-S2S* is seen to be more affected. Hence, *GTI-S2S* with "guided candidate generation" can robustly handle applications with limited supervision needs.

Overall, we observe that the proposed *TI-S2S* and *GTI-S2S* frameworks enable efficient FAQ retrieval by capturing query intent via representative target keywords. Experimental results demonstrate the transformation of questions onto a common keyword space provides improved accuracy as well as robustness.

5 Conclusion

We propose a novel FAQ Retrieval system using sequence-to-sequence framework to compute the similarity between user queries and FAQ based on "predicted representative intent keywords". We show how the filter-and-refine approach utilizing TF-IDF scores to obtain the representative keywords of questions act as weak supervision cues for capturing semantic similarities bridging the lexical and contextual gap in short span FAQ retrieval systems. Further, we also show that the use of "candidate identification" from an additional learning module boosts the performance of our framework by enabling early pruning. Experimental results on open-source FAQ dataset demonstrated the efficacy and robustness of our algorithm over existing approaches.

References

[Aizawa, 2003] A. Aizawa. An Information-Theoretic Perspective of TF-IDF Measures. *Information Processing & Management*, 39(1):45–65, 2003.

[Bengio *et al.*, 2015] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled Sampling for Sequence prediction with Recurrent Neural Networks. In *NIPS*, pages 1171–1179, 2015.

[Berger *et al.*, 2000] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. Bridging the Lexical Chasm: Statistical Approaches to Answer-finding. In *SIGIR*, pages 192–199, 2000.

[Devlin *et al.*, 2019] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidir. Transformers for Lang. Understanding. In *NAACL-HLT*, pages 4171–4186, 2019.

[Figueroa, 2017] A. Figueroa. Automatically Generating Effective Search Queries Directly from Community QA Questions for Finding Related Questions. *Expert Systems with Applications*, 77:11–19, 2017.

[Gupta and Carvalho, 2019] S. Gupta and V. R. Carvalho. FAQ Retrieval Using Attentive Matching. In *SIGIR*, pages 929–932, 2019.

[Jijkoun and de Rijke, 2005] V. Jijkoun and M. de Rijke. Retrieving Answers from Frequently Asked Questions Pages on the Web. In *CIKM*, pages 76–83, 2005.

[Karan and Šnajder, 2018] M. Karan and J. Šnajder. Paraphrase-focused Learning to Rank for Domain-specific FAQ Retrieval. *Expert Systems With Applications*, 91:418–433, 2018.

[Kothari *et al.*, 2009] G. Kothari, S. Negi, T. A. Faruquie, V. T. Chakaravarthy, and L. V. Subramaniam. SMS Based Interface for FAQ Retrieval. In *ACL-IJCNLP*, pages 852–860, 2009.

[Kusner *et al.*, 2015] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. From Word Embeddings to Document Distances. In *ICML*, pages 957–966, 2015.

[Lee *et al.*, 2008] J. T. Lee, S. B. Kim, Y. I. Song, and H. C. Rim. Bridging Lexical Gaps between Queries and Questions on Large Online Q&A Collections with Compact Translation Models. In *EMNLP*, pages 410–418, 2008.

[Liu *et al.*, 2019] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692, 2019.

[Lommatzsch and Katins, 2019] A. Lommatzsch and J. Katins. An Information Retrieval-based Approach for Building Intuitive Chatbots for Large Knowledge Bases. In *LWDA*, pages 343–352, 2019.

[Luong *et al.*, 2015] T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, and W. Zaremba. Addressing the Rare Word Problem in Neural Machine Translation. In *ACL-IJCNLP*, pages 11–19, 2015.

[Malik *et al.*, 2007] R. Malik, L. V. Subramaniam, and S. Kaushik. Automatically Selecting Answer Templates to Respond to Customer Emails. In *IJCAI*, pages 1659–1664, 2007.

[Mass *et al.*, 2020] Y. Mass, B. Carmeli, H. Roitman, and D. Konopnicki. Unsupervised FAQ Retrieval with Question Generation and BERT. In *ACL*, pages 807–812, 2020.

[Massaro *et al.*, 2018] A. Massaro, V. Maritati, and A. Galiano. Automated Self-learning Chatbot Initially Build as a FAQs Database Information Retrieval System. *Informatica (Slovenia)*, 42(4):515–525, 2018.

[Mikolov *et al.*, 2013] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Represt. of Words and Phrases and their Compositionality. In *NIPS*, pages 3111–3119, 2013.

[Nogueira *et al.*, 2020] R. Nogueira, Z. Jiang, R. Pradeep, and J. Lin. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of EMNLP*, pages 708–718, 2020.

[Otsuka *et al.*, 2018] A. Otsuka, K. Nishida, K. Bessho, H. Asano, and J. Tomita. Query Expansion with Neural Question-to-Answer Translation for FAQ-Based QA. In *WWW*, page 1063–1068, 2018.

[Reimers and Gurevych, 2019] N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP*, pages 3982–3992, 2019.

[Sakata *et al.*, 2019] W. Sakata, T. Shibata, R. Tanaka, and S. Kurohashi. FAQ Retrieval using Query-Question Simi. and BERT-Based QA Relevance. In *SIGIR*, pages 1113–1116, 2019.

[Santos *et al.*, 2020] J. Santos, L. Duarte, J. Ferreira, A. Alves, and H. G. Oliveira. Developing Amaia: A Conversational Agent for Helping Portuguese Entrepreneurs—An Extensive Exploration of Question-Matching Approaches for Portuguese. *Information*, 11(9):428, 2020.

[Sneiders, 2010] E. Sneiders. Automated Email Answering by Text Pattern Matching. In *ICNLP*, pages 381–392, 2010.

[Surdeanu *et al.*, 2011] M. Surdeanu, M. Ciaramita, and H. Zaragoza. Learning to Rank Answers to Non-factoid Questions from Web Collections. *Comp. Ling.*, 37(2):351–383, 2011.

[Sutskever *et al.*, 2014] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to Sequence Learning with Neural Networks. In *NIPS*, pages 3104–3112, 2014.

[Xie *et al.*, 2019] R. Xie, Y. Lu, F. Lin, and L. Lin. FAQ-based Question Answering via Knowledge Anchors. *CoRR*, abs/1911.05930, 2019.

[Yan *et al.*, 2016] Z. Yan, N. Duan, J. Bao, P. Chen, M. Zhou, Z. Li, and J. Zhou. DocChat: An Information Retrieval Approach for Chatbot Engines using Unstructured Docs. In *ACL*, pages 516–525, 2016.