AUTOHOME-ORCA at SemEval-2019 Task 8: Application of BERT for Fact-Checking in Community Forums

Zhengwei Lv[1] **Duoxing Liu**[1] **Haifeng Sun**[2] **Xiao Liang**[1]

Tao Lei[1] **Zhizhong Shi**[1] **Feng Zhu**[1] **Lei Yang**[1]

[1] Autohome Inc., Beijing, China
[2] Beijing University of Posts and Telecommunications, Beijing, China
{lvzhengwei,liuduoxing,liangxiao12030,leitao,shizhizhong,zhufeng,yanglei}@autohome.com.cn
hfsun@bupt.edu.cn

Abstract

Fact checking is an important task for maintaining high quality posts and improving user experience in Community Question Answering forums. Therefore, the SemEval-2019 task 8 is aimed to identify factual question (subtask A) and detect true factual information from corresponding answers (subtask B). In order to address this task, we propose a system based on the BERT model with meta information of questions. For the subtask A, the outputs of fine-tuned BERT classification model are combined with the feature of length of questions to boost the performance. For the subtask B, the predictions of several variants of BERT model encoding the meta information are combined to create an ensemble model. Our system achieved competitive results with an accuracy of 0.82 in the subtask A and 0.83 in the subtask B. The experimental results validate the effectiveness of our system.

1 Introduction

The Community Question Answering (CQA) forums are gaining more and more popularity because they can offer great opportunity for users to get appropriate answers to their questions from other users. Meanwhile, the accumulated massive questions and answers in CQA forums present a new challenge to provide valuable information for users more effectively. Therefore, researchers have shown an increased interest in CQA systems (Srba and Bielikova, 2016; Wang et al., 2018), aiming to facilitate efficient knowledge acquisition and circulation. Specifically, a large portion of researches mainly focus on the two tasks: find relevant questions to a new question to reuse corresponding answers (*Question Retrieval*), and search for relevant answers among existing answers to other questions (*Answer Selection*).

Despite a great deal of research on CQA, there are relatively few studies focusing on the quality of questions and answers. Actually, the credibility of answers is an important aspect, which can directly affect the user experience for CQA forums. In order to check the veracity of answers automatically, some recent works (Karadzhov et al., 2017; Mihaylova et al., 2018) attempt to utilize external sources and extract appropriate features for classification. Considering the importance of information veracity in CQA forums, the fact checking of answers is still an issue that is worth investigating further.

Therefore, the SemEval-2019 task 8 aims to conduct fact checking in CQA forums. In order to detect the veracity of answers, it is necessary to identify whether the questions are factual firstly. The task is comprised of two subtasks: the subtask A is targeted to identify whether a question is asking for factual information, an opinion/advice or socializing. Given factual questions, the subtask B is aimed to determine whether the corresponding answers are true, false or not factual.

In order to address the SemEval-2019 task 8, we propose a system based on the BERT model (Devlin et al., 2018). In our system, we extend BERT for integrating some meta information of questions into the BERT encoder, and generate an ensemble model from some potential classification models to achieve very competitive results. To be specific, in subtask A, two outputs of fine-tuned BERT classifiers are obtained from subjects and bodies of questions respectively. Then by combining both outputs with the length of questions as features, the AdaBoost method (Schapire, 1999) is utilized to boost the performance of question classification. As for subtask B, while encoding additional meta information (category and subject of questions) into BERT model, we adopt the bagging method for some variants of BERT model produced by adding additional layers. The experimental results in both subtasks demonstrate the

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 870–876
Minneapolis, Minnesota, USA, June 6–7, 2019.

effectiveness of our system.

The rest of our paper is organized in the following way. The related work about CQA is summarized in Section 2. Section 3 gives a more detailed description of our system. The results and analysis of experiments are demonstrated in Section 4. Finally, Section 5 presents the main conclusions.

2 Related Work

So far, most studies about CQA mainly pay attention to two tasks: *Question Retrieval* and *Answer Selection*. In previous works, some traditional methods treat questions or answers as bag of words and measure their similarities based on weighted matching between the words (Robertson et al., 1994) or translation probability learning from language model (Xue et al., 2008). In fact, similar questions often are not phrased with exactly same words, but related words, while there is very little token overlap between questions and answers. These methods essentially consider the question or answer as a bag of words, neglecting semantic information. So it is not surprising that the performance of traditional methods is not very well on aforementioned tasks. Recently, the neural-based models (He et al., 2015; Feng et al., 2015; Tan et al., 2016; Bachrach et al., 2017; Tay et al., 2018), which can capture some semantic relations, are proposed and become mainstream in the research about CQA gradually. The basic idea behind them is to learn the representation of questions and answers based on CNN or LSTM models, then conduct text matching by regarding both tasks as classification or learning to rank.

Furthermore, there are also public CQA datasets and competitions available, which promote relevant researches substantially. The public datasets are collected from various CQA websites, including Quora[1] , Yahoo! Answers[2], Qatar Living[3], etc. As for competitions, there is a kaggle competition[4] to identify the duplicated question pairs collecting from the Quora website. In SemEval-2015 Task 3 "Answer Selection in Community Question Answering" (Nakov et al., 2015), it is mainly targeted on the answer selection task. And there is a more comprehensive competition in SemEval-2016 Task 3 (Nakov et al., 2016) designed for both Question Retrieval and Answer Selection, which is consisted of four subtasks: Question-Comment Similarity, Question-Question Similarity, Question-External Comment Similarity and Reranking the correct answers for a new question. In contrast, in SemEval-2017 task 3 (Nakov et al., 2017), a new duplicate question detection subtask is incorporated on the basis of the SemEval-2016 Task 3.

Although much work has been done in CQA researches, few attentions have been paid on improving the quality of questions and answers. In order to detect true factual answers automatically, Karadzhov et al. (Karadzhov et al., 2017) propose a general framework using external sources, which adopts the LSTM model (Hochreiter and Schmidhuber, 1997) to learn text representation of answers and external sources. Mihaylova et al. (Mihaylova et al., 2018) extract features from multiple aspects (the answer content, the author profile, the rest of the community forum and external authoritative sources) and demonstrate the effectiveness of fact checking of answers. At the same time, the lack of large-scale dataset also restricts the progress on fact checking in CQA forums further.

Recently, there are some of key milestones in the NLP field, such as ELMo (Peters et al., 2018), ULMFiT (Howard and Ruder, 2018), OpenAI GPT (Radford, 2018) and BERT (Devlin et al., 2018). These large-scale models have provided great performance on various NLP tasks, which can be pre-trained on a massive corpus of unlabeled data, and then fine-tuned to downstream tasks. Especially, the BERT model has achieved state-of-the-art results on a variety of language tasks, which allows us to obtain significantly higher performance than models that are only able to leverage a small task-specific dataset. Therefore, we build a system based on the BERT model for the SemEval2019 task 8 and achieve satisfactory results.

3 System Description

3.1 System Overview

The pipeline of our system is shown in Figure 1. Firstly, original input files with questions and answers are preprocessed, including removing redundant information (e.g., HTML tags, URLs and strings exceeding maximum length limit) and extracting the structured contents and

[1]https://www.quora.com
[2]https://answers.yahoo.com
[3]https://www.qatarliving.com
[4]https://www.kaggle.com/c/quora-question-pairs

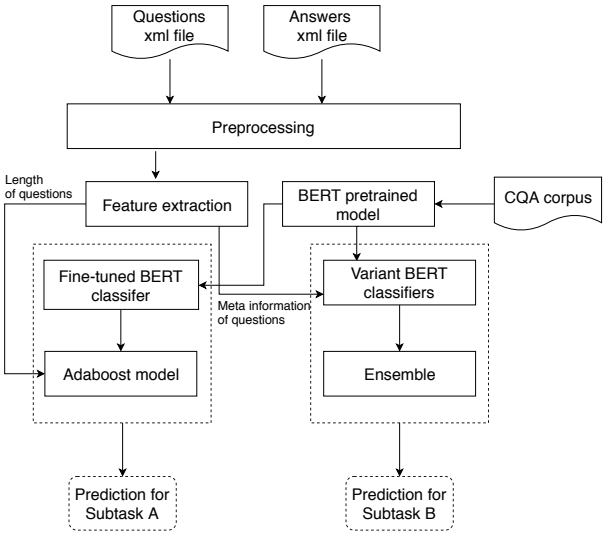

Figure 1: Pipeline of our system.

Label	Subject of questions	Body of questions
Opinion	e.g., *does anyone know good dentist?*	e.g., *can anybody recommend me a dentist? a good one.*
Factual	e.g., *when is eid gonna start?*	e.g., *when will eid start? like holidays*
Socializing	e.g., *What do you like about the person above you?*	e.g., *Hello people...let's play this game...you have to write something good about the person whose 'post' is above you on QL.You can write anything and you can write multiple times. For ex;the person who will respond to my post will write about me ;) and so on. This will be fun...*

Table 1: Samples of questions with different labels.

meta information. Secondly, some important features are obtained from structured information, such as the length and category of questions. Thirdly, based on the pre-trained BERT model released by Google[5], we conduct unsupervised training on specific CQA corpus further to make the model more suitable for the following classification tasks. Finally, the pre-trained BERT model and extracted features are fed into two subsystems to obtain predictions for the subtask A and the subtask B respectively. In the subsystem for subtask A (detailed in Subsection 3.2), the AdaBoost model is adopted to predict the classification of questions by combining the outputs of fine-tuned BERT classifier and the feature of length of questions. In the subsystem for subtask B (described in Subsection 3.3), some variant BERT models which encode meta information of questions are combined to generate an ensemble model for predication of labels of answers.

3.2 Subsystem for Subtask A

In this Subsection, the subsystem for subtask A is described in detail below.

Firstly, the subject and body of questions are encoded into two BERT models separately for fine-tuning on the question classification. The different inputs for both BERT models are represented as

$$[CLS] + text1 + [SEP]$$
$$[CLS] + text2 + [SEP]$$

where $text1$ and $text2$ are the subject and body of question respectively.

Secondly, the outputs of two fine-tuned BERT models are concatenated with the length of questions' body as features for classification. As illustrated in Table 1, it is rather intuitive that the body length of questions for socializing is inclined to be longer than ones for factual or opinion. Therefore, it is reasonable to consider the body length of questions as a suitable feature for classification. In addition, the results of each BERT model are probabilities of questions belonging to different classes (Factual, Opinion and Socializing). Then the feature vector x_{vector} for question classification is represented as follows

$$x_{vector} = [P_s1, P_s2, P_s3, P_b1, P_b2, P_b3, L_b] \quad (1)$$

where P_s1, P_s2, P_s3 are the output of a BERT model encoding the question subject. Similarly, P_b1, P_b2, P_b3 are the output of another BERT model encoding the question body, and L_b is the body length of a question.

Finally, based on the generated feature vector x_{vector}, the AdaBoost algorithm is adopted to obtain the final results of classification. AdaBoost is a typical Boosting algorithm that aims to convert a set of relative weak classifiers to a strong classifier. Therefore, the performance of classification can be strengthened by considering additional length feature, compared to the one that the BERT models

[5]https://github.com/google-research/bert

have achieved.

3.3 Subsystem for Subtask B

The Subsection describes the details of the subsystem for subtask B as follows.

Firstly, the subject and body of question, corresponding reply (i.e., answer) and meta information of question are combined to generate sequences for BERT encoders. In order to identify the true factual reply, the content of corresponding questions and auxiliary information (e.g., the category of question, username of a questioner or replier) should be necessary for classifiers. So in the subsystem, we investigate the influence of different information for the classification performance (see Table 4 for details), including the subject of question (F-subject), the usernames of questioner and replier (F-username) and the category of question (F-category). Ultimately, the text of answer and the information of F-subject and F-category are employed for our BERT based models. The generated sequence for inputs of models are represented as following:

$$[CLS] + text1 + [SEP] + text2 + [SEP]$$

We use [SEP] to separate between the information of question and answer. $text1$ is composed of F-subject, F-category and the body of question separated by the special symbol ($\sim$), while $text2$ is the text of corresponding reply.

Secondly, based on the generated sequences as inputs, we design three different categories of BERT based models for ensemble. As shown in Figure 2, the specific structures of the three kinds of models are described as follows:

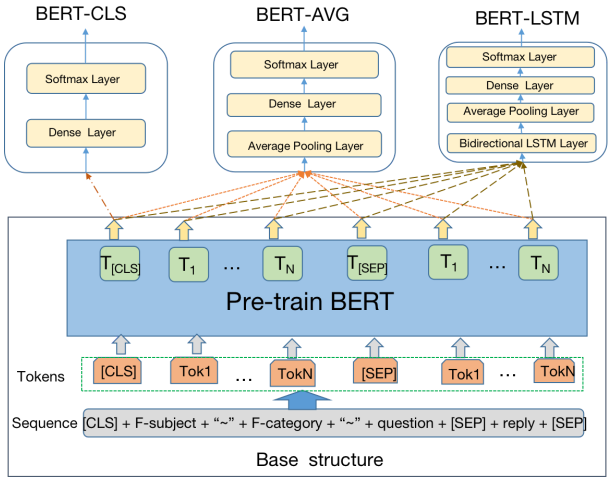

Figure 2: Architecture of BERT based models.

- *BERT-CLS*. The final hidden state for the first token [CLS] in the input is employed for fine-tuning the pre-trained BERT model by adding a classification layer and a standard softmax.

- *BERT-AVG*. Different from BERT-CLS, the final hidden states for all tokens are utilized for classification by conducting an average pooling, and then adding a full connected layer and a standard softmax.

- *BERT-LSTM*. Compared with BERT-AVG, a Bi-LSTM network is added between the pooling layer and the pre-trained BERT encoder. It must be noted that we only obtain the outputs of BERT encoder and the parameters of BERT encoder are not updated when training.

Thirdly, we select a set of competitive classifiers in the training process by the three kinds of BERT based models respectively. The method of five-fold cross-validation is employed. To be specific, the original samples are randomly divided into five sub-samples with equal size. And one of the five sub-samples is retained for validating the performance of classifier, and the rest of four sub-samples are used as training data. For each kind of BERT based model, the cross-validation process is repeated five times and each time no more than five optimal classifiers are obtained. Therefore, we get a total of sixty-five competitive classifiers filtered by certain threshold value on accuracy metric from three kinds of BERT based models for ensemble.

Finally, an effective integration strategy is applied to produce a strong classifier for the subtask. There are two candidate integration strategies:

- *Strategy 1 (Vote-ensemble)*: Each classifier casts a vote, the label of a sample is decided according to the majority of votes.

- *Strategy 2 (Distribution-ensemble)*: If the number of votes for any label exceeds one-half of the total number of classifiers, the sample is classified as the corresponding label. Otherwise, the label of the sample will be determined by considering the actual label distribution of the training data and the label distribution of votes together. For example, if one sample's votes for different labels are very close, then the sample is classified as the

label with the largest proportion of data distribution.

At last, the strategy 2 is employed in our subsystem because it seems that Distribution-ensemble strategy is more robust for variance error, especially for small dataset, which will be discussed in Subsection 4.2 further.

4 Experiment

4.1 Dataset

The dataset is organized in question-answer threads from the Qatar Living forum. Each question, which is annotated by labels: Opinion, Factual and Socializing, has a subject, a body and meta information including question ID, category, posting time, user's ID and name. And each answer, which is classified as Factual-True, Factual-False and Non-Factual, has a body and meta information (answer ID, posting time, user's ID and name). The detailed statistics of the dataset in this task are illustrated in the task description paper (Mihaylova et al., 2019).

4.2 Experimental Results and Analysis

As for pre-training the BERT model, it is trained based on the BERT-Base-Cased model by the forum corpus provided by organizer[6]. The training batch size is 32, the number of train steps is 1e+5 and the learning rate is 2e-5. The detailed experimental results for both subtasks are described as following.

4.2.1 Results for Subtask A

In the subsystem for subtask A, the AdaBoost algorithm is employed to boost the performance on question classification. The number of estimators for the AdaBoost method is 10. To evaluate the performance of question classification, we compare our proposed method against the following models:

- Text-CNN (Kim, 2014): a simple CNN with one layer of convolution on top of word vectors. The subject and body of each question are concatenated as the input of Text-CNN model. When training, the number of epoch is 80, the initial learning rate is 0.001 and the dropout rate is set to 0.4.

- BERT without pre-training: the BERT-Base cased model release by Google. The input of the model is the concatenation of the subject and the body of each question with the symbol [SEP], which is represented as follows:

$$[CLS] + text1 + [SEP] + text2 + [SEP]$$

  $text1$ and $text2$ are the subject and body of a question separately. When training the model, the batch size of training is 32, the initial learning rate is 2e-5 and the number of epoch is 9.

- BERT with pre-training: the BERT model pre-trained by CQA corpus. The settings of hyper-parameters is the same as the BERT model without pre-training.

Models	Acc. (Dev)	Acc. (Test)
Text-CNN	0.6569	0.6502
BERT without pre-training	0.6862	0.7370
BERT with pre-training	0.7197	0.7922
Our method	**0.7283**	**0.8181**

Table 2: Performance of different models in the subtask A.

The comparison results are shown in Table 2. From the table, it can be observed that the accuracy of the Text-CNN model is much lower than the other three BERT-based models. Even if only the BERT model without pre-training is used to predict the final result, it is 2.93% and 8.68% higher than Text-CNN model on development dataset and test dataset, respectively. Considering the size of dataset is relative small, it seems to demonstrate the potential advantage of BERT based models. Compared with the BERT model without pre-training, the BERT model with pre-training has 3.35% and 5.52% increase respectively. It is illustrated that the step of pre-training the BERT model is very important. Furthermore, the accuracy achieved by our method is 0.86% and 2.59% higher than the one by the BERT with pre-training model on two datasets separately. It shows that the AdaBoost algorithm can make better use of the probability outputs from the fine-tuned BERT models for prediction. What's more, the body length of questions can be considered as an effective feature for training model and predicting results.

[6]http://alt.qcri.org/semeval2016/task3/data/uploads/QL-unannotated-data-subtaskA.xml.zip

4.2.2 Results for Subtask B

In the experiments for subtask B, the three kinds of BERT based models are implemented with TensorFlow and trained with Adam optimizer. The maximum length of sequence is set to 150 and the batch size is 4. The initial learning rates are 3e-5 for parameters of BERT encoder and 1e-3 for others.

Models	Acc.(Dev)	Acc.(Test)
BERT-AVG	0.6732	–
BERT-CLS	0.6667	–
BERT-LSTM	0.656	–
Vote-ensemble	0.6693	0.7935
Distr.-ensemble	**0.6845**	**0.8322**

Table 3: Performance of different models in the subtask B.

The experimental results of different kinds of models are shown in Table 3. From the table, it can be observed that the BERT-AVG model achieves the best performance in the three single models. By conducting average pooling operation on final hidden states of all tokens, the BERT-AVG model can capture more semantic information than the BERT-CLS model which can only pay attention to the hidden state of the [CLS] token. As for the BERT-LSTM model, it performs the worst, which may be caused by the highest model complexity and the lack of adequate training dataset, resulting in somewhat overfitting. In addition, it is indicated that ensemble models can obtain higher accuracy than single models and the strategy of Distribution-ensemble is more robust than the strategy of Vote-ensemble. This is because that when the numbers of votes for different labels are close to each other, it is difficult to identify the correct class only by the majority. By considering actual classification distribution in training dataset additionally, the Distribution-ensemble can show its potential advantage.

Feature	Acc.(Dev)
Baseline	0.6559
+F-category	0.6606(+0.47)
+F-username	0.6547(-0.12)
+F-subject	0.6642(+0.83)
+F-category, +F-subject	**0.6667(+1.08)**

Table 4: Performance of different features on development dataset in the subtask B. "+" means to add current features to the main feature.

In order to explore the effectiveness of different information for classification, a series of experiments based on the BERT-CLS model are conducted. The baseline model (BERT-CLS) is established only by encoding the information of the body of question and the corresponding answer. Therefore, the influence of other information can be discussed individually. By considering different information, the performance of the model validated on development dataset is shown in Table 4. It is observed that the F-username can not contribute to the increase of accuracy, which may be caused by existing many anonymous users in the forum. By encoding the information F-subject and F-category into the model, it can achieve the best performance.

5 Conclusion

Detecting the veracity of answers is vital to maintain high quality information in CQA forums. In order to address this problem, a system based on BERT model is developed for participating in the SemEval-2019 Task 8. In the system, the meta information of questions is encoded into the BERT model and an ensemble with multiple variants of BERT model are produced to accomplish better performance. In subtask A, we utilize the AdaBoost algorithm to the features that is consisted of fine-tuned results of BERT models and length of questions. In subtask B, after encoding the auxiliary information of questions and answers into the BERT model, fine-tuned BERT model and two variant models by adding average-pooling or LSTM layers are combined to reduce the variance error. Finally, our system achieved great performance with an accuracy of 0.82 and 0.83 in the two subtasks respectively.

To our surprise, the system has impressive results in the subtask B without using external sources. It may be explained by the potential advantage of BERT model over other models only trained on a small task-specific dataset. In the future, we will explore to retrieve relevant information from the Web efficiently and then integrate the external information into our BERT based model.

References

Yoram Bachrach, Andrej Zukov Gregoric, Sam Coope, Ed Tovell, Bogdan Maksak, José Rodríguez, Conan McMurtie, and Mahyar Bordbar. 2017. An attention mechanism for neural answer selection using a combined global and local view. *2017 IEEE 29th*

International Conference on Tools with Artificial Intelligence (ICTAI), pages 425–432.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 813–820.

Hua He, Kevin Gimpel, and Jimmy Lin. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1576–1586. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339. Association for Computational Linguistics.

Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. Fully Automated Fact Checking Using External Sources. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 344–353. INCOMA Ltd.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Tsvetomila Mihaylova, Georgi Karadzhov, Atanasova Pepa, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. SemEval-2019 task 8: Fact checking in community question answering forums. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '19, Minneapolis, MN, USA.

Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Mitra Mohtarami, Georgi Karadzhov, and James Glass. 2018. Fact Checking in Community Forums. In *AAAI Conference on Artificial Intelligence*.

Preslav Nakov, Lluís Arquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak Abed, Alhakim Freihat, James Glass, Bilal Randeree, and Qatar Living. 2016. SemEval-2016 Task 3: Community Question Answering. In *Proceedings of SemEval-2016*, pages 525–545.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 Task 3: Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 27–48.

Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 2015. Semeval-2015 task 3: Answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 269–281. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

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

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, pages 109–126.

Robert E. Schapire. 1999. A brief introduction to boosting. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, pages 1401–1406, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ivan Srba and Maria Bielikova. 2016. A Comprehensive Survey and Classification of Approaches for Community Question Answering. *ACM Transactions on the Web*, 10(3):1–63.

Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2016. Improved representation learning for question answer matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 464–473. Association for Computational Linguistics.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Multi-cast attention networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 2299–2308, New York, NY, USA. ACM.

Xianzhi Wang, Chaoran Huang, Lina Yao, Boualem Benatallah, and Manqing Dong. 2018. A survey on expert recommendation in community question answering. *Journal of Computer Science and Technology*, 33(4):625–653.

Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. 2008. Retrieval models for question and answer archives. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 475–482, New York, NY, USA. ACM.