Learning to Rank Question Answer Pairs with Bilateral Contrastive Data Augmentation*

Yang Deng, Wenxuan Zhang, Wai Lam
The Chinese University of Hong Kong
{ydeng,wxzhang,wlam}@se.cuhk.edu.hk

Abstract

In this work, we propose a novel and easy-to-apply data augmentation strategy, namely **Bi**lateral **G**eneration (**BiG**), with a contrastive training objective for improving the performance of ranking question answer pairs with existing labeled data. In specific, we synthesize pseudo-positive QA pairs in contrast to the original negative QA pairs with two pretrained generation models, one for question generation, the other for answer generation, which are fine-tuned on the limited positive QA pairs from the original dataset. With the augmented dataset, we design a contrastive training objective for learning to rank question answer pairs. Experimental results on three benchmark datasets show that our method significantly improves the performance of ranking models by making full use of existing labeled data and can be easily applied to different ranking models.

1 Introduction

Ranking question answer pairs, also known as answer selection, is a fundamental task in question answering (QA) systems. It aims to rank a set of candidate answers for selecting the relevant or correct answers to the given question. Many efforts have been made on developing various neural models to measure the relevance degree between the question and answer pair, including Siamese Structure (Tay et al., 2018; Chen et al., 2020), Attention-based Structure (Chen et al., 2017; Shen et al., 2018; Deng et al., 2021), and Compare-Aggregate Structure (Yoon et al., 2019). Recently, models with contextualized representations, e.g., ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), contribute to major improvement on answer selection (Yoon et al., 2019; Garg et al., 2020).

q: How many music awards has Katy Perry won?
$\mathbf{a}_1^-$: Katy Perry is an American singer from Santa Barbara California.
$\mathbf{a}_2^-$: Perry has released three studio albums including her debut Christian effort Katy Hudson One of the Boys and Teenage Dream.
$\mathbf{a}_3^-$: Perry's mainstream career was kicked off with the massive chart success of her debut single I Kissed a Girl.
$\mathbf{a}^*$: Katy Perry has won 11 awards for her music career including three Grammy awards and two American Music Awards.
$\mathbf{q}_1^*$: What city is Katy Perry from?
$\mathbf{q}_2^*$: How many albums has Katy Perry released in her career?
$\mathbf{q}_3^*$: Who is the singer Katy Perry?

Table 1: An Example from WikiQA with Synthesized Data. $\mathbf{q}, \mathbf{a}_1^-, \mathbf{a}_2^-, \mathbf{a}_3^-$ are the original data, but all these answers are irrelevant to the given question. $\mathbf{a}^*, \mathbf{q}_1^*, \mathbf{q}_2^*, \mathbf{q}_3^*$ are the corresponding synthesized data.

Acquiring the ground-truth positive answer to the given question or manually annotating the relevance degree between a QA pair can be extremely time-consuming. Therefore, the existing answer selection datasets are often limited in scale. Such data scarceness issue poses a great challenge for the modern neural models, which are often data-hungry for the training process. Moreover, the negative samples, i.e., the irrelevant answers to the given question are constructed by random negative sampling from the whole document pool (Tay et al., 2017), leaving less to learn for differentiating the relevant answer from a set of highly irrelevant answers (Zhang et al., 2017; Kumar et al., 2019). Therefore, the data scarceness and label imbalance issues lead to the underutilization of the in-domain knowledge in the original datasets, and even a waste of valuable labeled data.

Some early works (Min et al., 2017; Deng et al., 2018) leverage transfer learning (TL) based approaches to transfer knowledge from large-scale QA datasets, e.g., SQuAD (Rajpurkar et al., 2016), for improving resource-poor answer selection tasks. With similar motivations, Yoon et al. (2019) combine ELMo with transfer learning from a large-scale relevant-sentence-selection dataset to enhance the performance of the compare-aggregate model. Garg et al. (2020) adopt BERT-based mod-

els with a Transfer-and-Adapt (TANDA) approach by using a large and high-quality QA dataset. However, such TL-based methods heavily rely on the existence of large-scale high-quality datasets with similar problem settings, which is also unrealistic in real-world applications. Besides, compared with utilizing cross-domain knowledge via TL, it would be more efficient to take advantage of the in-domain knowledge from the original dataset.

Another way to handle the data scarceness and label imbalance issues is to employ hard negative sampling strategies (Zhang et al., 2017; Kumar et al., 2019) or curriculum training approaches (MacAvaney et al., 2020) for taking advantages of those randomly sampled negative answers. The essential motivation behind these methods is that the hard negative samples or the "difficult" samples play a more important role in training a better ranking model. However, those "less" important samples are likely to be overlooked in such cases, leading to a potential waste of data. In order to maximize the utility of the existing labeled data, we aim at investigating data augmentation strategy to construct weak supervision signals for enhancing answer selection models.

In this work, we propose an effective data augmentation strategy, namely **Bi**lateral **G**eneration (**BiG**), with a contrastive training objective for improving the performance of ranking question answer pairs with existing labeled data. In specific, we first fine-tune a Question Generator (QG) and an Answer Generator (AG) from the pre-trained generative language model, such as BART (Lewis et al., 2020), with the positive samples in the training data. Then, as the example shown in Table 1, we use the fine-tuned QG and AG to generate a high-quality question-related answer $\mathbf{a}^*$ and answer-focused questions $\mathbf{q}^*$ from the given negative samples $(\mathbf{q}, \mathbf{a}^-)$. The synthesized QA pairs $(\mathbf{q}, \mathbf{a}^*)$ and $(\mathbf{q}^*, \mathbf{a}^-)$ can be regarded as new "positive" samples in contrast to the original negative sample. By doing so, given a small amount of labeled data even with extremely imbalanced labels, such data augmentation strategy can easily synthesize high-quality data with weak supervision signals. Although such pseudo-positive QA pairs may not be perfectly matched, e.g., $(\mathbf{q}_3^*, \mathbf{a}_3^-)$, they are supposed to be more interrelated than the original negative QA pairs, which can be served as pairwise samples for contrastive estimation. For instance, $(\mathbf{q}_3^*, \mathbf{a}_3^-)$ can be used to establish a contrastive sample with $(\mathbf{q}, \mathbf{a}_3^-)$, rather than be simply regarded as a positive QA pair. Therefore, we further design a contrastive training objective for ranking QA pairs.

We conduct experiments by applying the proposed BiG method with both the attention-based ranking model and the pre-trained BERT ranking model. Experimental results on three benchmark datasets including TREC-QA, WikiQA, and ANTIQUE show the effectiveness and the applicability of our method, which significantly improves the original ranking models across all datasets.

2 Method

2.1 Preliminaries

Let an ad-hoc neural ranking model be denoted as $R_\theta(\mathbf{q}, \mathbf{a})$, which measures the relevance degree of a given question answer pair $(\mathbf{q}, \mathbf{a})$ parameterized by the model parameters θ. Given a set of training samples, there are three mainstream approaches to optimize the ranking performance, namely, pointwise training, pairwise training, and listwise training (Lai et al., 2018; Zhang et al., 2020). Here we focus on pointwise and pairwise training, since listwise training is less practical in applications when there is a large set of candidate answers.

The pointwise training aims to optimize the predicted probability distribution of each QA pair to be consistent with the ground-truth label by minimizing the cross-entropy loss function:

$$p_i = \text{softmax}(R_\theta(\mathbf{q}_i, \mathbf{a}_i)), \quad (1)$$
$$\mathcal{L}^{\text{pt}}(\mathbf{q}_i, \mathbf{a}_i, y_i) = -\left[y_i \log p_i + (1 - y_i) \log (1 - p_i)\right], \quad (2)$$

where y_i is the label of the i-th training sample.

On the other hand, the pairwise training is to optimize the relative order of a pair of predicted relevance scores, i.e., $R_\theta(\mathbf{q}, \mathbf{a}^+)$ and $R_\theta(\mathbf{q}, \mathbf{a}^-)$, which typically adopt the same question $\mathbf{q}$ with one relevant answer $\mathbf{a}^+$ and one irrelevant answer $\mathbf{a}^-$. The main idea behind pairwise training is contrastive estimation. One widely-used pairwise loss function is the hinge loss:

$$\mathcal{L}^{\text{pr}}(\mathbf{q}, \mathbf{a}^+, \mathbf{a}^-) = \max(0, M - R_\theta(\mathbf{q}, \mathbf{a}^+) + R_\theta(\mathbf{q}, \mathbf{a}^-)), \quad (3)$$

where the model learns to rank the relevant answer $\mathbf{a}^+$ higher than the irrelevant answer $\mathbf{a}^-$, by imposing a margin M between their relevance scores with the given question $\mathbf{q}$.

2.2 BiG Data Augmentation

To maximize the utility of the valuable labeled data as well as alleviate the label imbalance issue, pre-

trained language models, such as BART ([Lewis et al., 2020](#)), preserve promising advantages of synthesizing high-quality pseudo QA pairs through conditional generation, which can serve as weak supervision signals for contrastive learning with original data. Due the nature of question answering, there are two intuitive ways to generate augmentations, *i.e.*, question generation and answer generation, namely **Bi**lateral **G**eneration (**BiG**).

With the pre-trained conditional generation model, a certain number of training data is required to fine-tune a Question Generator (QG) and an Answer Generator (AG). We adopt the positive samples, i.e., the relevant $(\mathbf{q}, \mathbf{a}^+)$ pair, as the training data to fine-tune the QG and AG:

$$\mathbf{H} = \text{QGEnc}(\mathbf{a}^+; [\text{SEP}]), \quad \mathbf{q} = \text{QGDec}(\mathbf{H}), \quad (4)$$

$$\mathbf{H} = \text{AGEnc}(\mathbf{q}; [\text{SEP}]), \quad \mathbf{a}^+ = \text{AGDec}(\mathbf{H}), \quad (5)$$

where [SEP] is a separation token.

Then, the learned QG and AG models can be employed to generate answer-related questions $\mathbf{q}^*$ and question-related answers $\mathbf{a}^*$, respectively. In our case, we are able to synthesize pseudo-positive QA pairs from the original negative QA pairs. For instance, given a negative QA pair. $(\mathbf{q}, \mathbf{a}^-)$, two pseudo-positive QA pairs, i.e., $(\mathbf{q}^*, \mathbf{a}^-)$ and $(\mathbf{q}, \mathbf{a}^*)$, can be generated by the learned question generator QG and answer generator AG:

$$\mathbf{q}^* = \mathbf{QG}(\mathbf{a}^-; [\text{SEP}]), \quad \mathbf{a}^* = \mathbf{AG}(\mathbf{q}; [\text{SEP}]). \quad (6)$$

One one hand, since the generation model is pre-trained on large scale corpus and further fine-tuned on positive QA pairs, we are guaranteed to generate reliable questions or answers related to the source text. In other words, the relevance degree between the synthesized QA pairs is supposed to be sufficiently higher than the original negative QA pairs constructed by randomly sampling. One the other hand, compared with the original positive QA pairs, the synthesized QA pairs are likely to contain certain redundant information or lack of necessary information. Regarding the synthesized QA pairs as ground-truth positive pairs may introduce undesired noises into the original dataset. Therefore, the pseudo-positive nature of the systhesized QA pairs leaves room for contrastive learning.

2.3 Contrastive Training Objective

As mentioned in Section [2.1](#), conventional supervised ranking problem can only optimize the ranking performance with existing labeled data. Here we introduce contrastive learning to rank by leveraging generation-augmented contrastive data for further improving the ranking performance.

Given a question $\mathbf{q}$ from the original dataset, there are a set of relevant answers $\mathbf{A}^+ = \{\mathbf{a}^+\}$ and a set of irrelevant answers $\mathbf{A}^- = \{\mathbf{a}^-\}$, correspondingly. The pairwise training for conventional supervised learning to rank can be defined as:

$$\mathcal{L}_{\text{set}}^{\text{pr}}(\mathbf{q}, \mathbf{A}^+, \mathbf{A}^-) = \frac{\sum_{\mathbf{a}^+ \in \mathbf{A}^+} \sum_{\mathbf{a}^- \in \mathbf{A}^-} \mathcal{L}^{\text{pr}}(\mathbf{q}, \mathbf{a}^+, \mathbf{a}^-)}{|\mathbf{A}^+| \cdot |\mathbf{A}^-|}. \quad (7)$$

With the synthesized pseudo-positive samples, we leverage such contrastive weak supervision signals to enhance the pairwise training in Eq.[7](#):

$$\mathcal{L}_{\text{set}}^{\text{ctrst}}(\mathbf{q}, \mathbf{A}^+, \mathbf{A}^-) = \mathcal{L}_{\text{set}}^{\text{pr}}(\mathbf{q}, \mathbf{A}^+, \mathbf{A}^-)$$
$$+ \frac{1}{|\mathbf{A}^-|} \sum_{\mathbf{a}^- \in \mathbf{A}^-} \left[\mathcal{L}^{\text{pr}}(\mathbf{q}, \mathbf{a}^*, \mathbf{a}^-) + \mathcal{L}^{\text{pr}}(\mathbf{q}^*, \mathbf{q}, \mathbf{a}^-) \right], \quad (8)$$

where the second term implements the contrastive training between the synthesized pseudo-positive QA pairs, $(\mathbf{q}, \mathbf{a}^*)$ and $(\mathbf{q}^*, \mathbf{a}^-)$, and the original negative QA pairs, $(\mathbf{q}, \mathbf{a}^-)$.

3 Experiment

3.1 Experimental Setup

3.1.1 Datasets and Evaluation Metrics

We evaluate our method on three widely-used benchmark datasets for answer selection: TREC-QA, WikiQA, and ANTIQUE. The statistics of these datasets are described in Table [2](#).

- TREC-QA ([Wang et al., 2007](#)), collected from TREC QA track 8-13 data, is a widely-adopted benchmark for factoid question answering.

- WikiQA ([Yang et al., 2015](#)) is an open-domain factoid answer selection benchmark, in which the answers are collected from the Wikipedia.

Following previous studies ([Garg et al., 2020](#)), we adopt the RAW version of training data and evaluate on the cleaned test data for both TREC-QA and WikiQA. The mean average precision (MAP) and mean reciprocal rank (MRR) are adopted as the evaluation metrics.

- ANTIQUE ([Hashemi et al., 2020](#)) is an open-domain non-factoid QA dataset collected from community question answering services with a diverse set of categories.

Dataset (train/dev/test)	#Question	#QA Pairs	%Positive
TREC-QA	1160/65/68	53417/1117/1442	12.0/18.4/17.2
WikiQA	2118/126/243	20360/1130/2351	12.0/12.4/12.5
ANTIQUE	2426/-/200	27422/-/6589	72.2/-/38.4

Table 2: Summary statistics of datasets.

Although there are four-level relevance labels for the candidate answers, we follow previous studies (Hashemi et al., 2020; MacAvaney et al., 2020) to assume that the label 3 and label 4 are relevant, while 1 and 2 are non-relevant. MRR and P@1 are adopted for evaluation on the binary labels, while nDCG is adopted for evaluation on the four-level relevance labels.

3.1.2 Baseline Methods

Since TREC-QA and WikiQA have been extensively studied for answer selection, we are able to compare with more diverse baseline methods, including traditional methods and Transfer Learning based methods.

- We consider the methods without using external resources during the training phase as traditional methods. aNMM (Yang et al., 2016), RNN-POA (Chen et al., 2017), and KABLSTM (Shen et al., 2018) develop different kinds of attention-based structures for enhancing the interaction between the question and answer. HyperQA (Tay et al., 2018) and SD (Chen et al., 2020) study the similarity aggregation approaches in siamese neural network structure. Currently, models based on pre-trained contextualized representations, e.g., BERT (Devlin et al., 2019), achieve state-of-the-art performance among the traditional methods.

- TL-based methods refer to those methods that transfers cross-domain knowledge from external large-scale resources to enhance the resource-poor tasks. Here we compare our results with two state-of-the-art TL-based methods, i.e., CA+LM+LC+TL (Yoon et al., 2019) and TANDA (Garg et al., 2020), which adopts QNLI (86K QA pairs) and ASNQ (20M QA pairs) as source dataset, respectively, for transfer learning.

As for ANTIQUE, we compare to the benchmark performance on BERT and a newly developed curriculum learning strategy (MacAvaney et al., 2020) for learning to handle difficult cases.

3.1.3 Implementation Details

We adopt the Large version of BART (Lewis et al., 2020) as the conditional generation model for data

Method	TREC-QA		WikiQA	
	MAP	MRR	MAP	MRR
aNMM	0.750	0.811	0.610	0.628
RNN-POA	0.781	0.851	0.721	0.731
KABLSTM	0.804	0.885	0.732	0.749
HyperQA	0.784	0.865	0.712	0.727
SD	0.783	0.878	0.704	0.712
BERT$_{base}$	0.857	0.937	0.813	0.828
BERT$_{large}$	0.904	0.946	0.836	0.853
CA+LM+LC+TL(QNLI)	0.875	0.940	0.834	0.848
BERT$_{base}$ TANDA(QNLI)	0.863	0.906	0.832	0.852
BERT$_{base}$ TANDA(ASNQ)	0.912	0.951	0.893	0.903
BERT$_{large}$ TANDA(ASNQ)	0.912	0.967	0.904	0.912
aNMM-BiG	0.792†	0.879†	0.698†	0.710†
BERT$_{base}$-BiG	**0.899**†	**0.961**†	**0.825**†	**0.836**†
BERT$_{large}$-BiG	**0.913**†	**0.966**†	**0.855**†	**0.863**†

Table 3: Experimental Results on TREC-QA and WikiQA. † indicates significant improvement ($p < 0.05$) over the corresponding base ranking models.

augmentation. During the fine-tuning of the question generator and answer generator, we set the maximum sequence length for question and answer as 30 and 50, respectively. Both of the question generator and answer generator are fine-tuned for 5 epochs. All other hyper-parameters are set to default as the original BART. Since the proposed method can be applied to any ranking model, we choose two widely-adopted ranking models, namely, BERT (Devlin et al., 2019) and aNMM (Yang et al., 2016), as the base models for evaluation. The maximum sequence length of the concatenated QA pair is set to be 128. We train all the models for 5 epochs. All other hyper-parameters are also set to default as the original BERT. And we use MatchZoo (Guo et al., 2019) to implement the aNMM model.

3.2 Overall Performance

Table 3 summarizes the method comparisons on TREC-QA and WikiQA. Here we combine our BiG method with both an attention-based ranking model, namely aNMM, and the pre-trained BERT ranking model. Two groups of methods on answer selection, in terms of whether to use large-scale external resources, are adopted for comparisons.

Among **traditional methods**, models using contextualized representations, i.e., BERT$_{base}$ and BERT$_{large}$, achieve distinguishable state-of-the-art performance, showing the power of pre-trained language models. All the ranking models combined with the proposed data augmentation strategy, BiG, consistently and substantially improve the perfor-

Method	MRR	P@1	nDCG@1	nDCG@3	nDCG@10
aNMM	0.625	0.485	0.529	0.513	0.490
BERT$_{base}$	0.797	0.709	0.713	0.657	0.642
BERT$_{large}$	0.802	0.690	0.708	0.669	0.693
BERT$_{base}$ Pair.	0.700	0.585	-	-	-
- w/ $\mathcal{D}_{recip}$	0.734	0.645	-	-	-
aNMM-BiG	0.662†	0.565†	0.601†	0.586†	0.589†
BERT$_{base}$-BiG	**0.847**†	**0.765**†	**0.750**†	**0.710**†	**0.720**†
BERT$_{large}$-BiG	**0.859**†	**0.785**†	**0.772**†	**0.719**†	**0.738**†

Table 4: Experimental Results on ANTIQUE.

mance of the original ranking models. For instance, BERT$_{base}$-BiG outperforms the original BERT$_{base}$ by about 5% and 2% on TREC-QA and WikiQA datasets, respectively. The results indicate that the proposed data augmentation strategy effectively synthesize valuable contrastive data for learning to rank question answer pairs without the need of large-scale annotated data.

As for **transfer learning based methods**, the performances on target tasks step up to the next stage with large-scale external resources. Compared with CA+LM+LC+TL(QNLI) and TANDA(QNLI), which employs 86K external annotated QA pairs from QNLI for training, BiG-enhanced models achieve a competitive performance by only manipulating the limited original data. Even if comparing to TANDA(ASNQ), which employs 20M external QA pairs from Natural Questions for training, BiG-enhanced models can still outperform it on the MRR metric of TREC-QA. However, since Natural Questions and WikiQA are both derived from Wikipedia, they are likely to share a lot of related in-domain knowledge, leading to a great performance boosting on WikiQA.

Table 4 presents the experimental results on ANTIQUE. In general, BiG significantly improves the performance of BERT$_{base}$ by about 6%. Especially for nDCG, we observe a large performance increase in the multi-level ranking metrics. This result shows that the proposed method is also effective on non-factoid QA.

3.3 Ablation Study

According to the ablation studies in Table 5, we observe that both the BiG data augmentation and the contrastive learning objective contribute to the final performance more or less.

As for TREC-QA and ANTIQUE, pairwise training achieves a better performance than pointwise training. Since there is a large number of questions without any positive answer in the WikiQA dataset, the lack of pairwise training samples causes that the performance of pairwise training is much worse

Method	TREC-QA		WikiQA		ANTIQUE	
	MAP	MRR	MAP	MRR	MRR	P@1
BERT$_{base}$ Point.	0.857	0.937	0.813	0.828	0.797	0.709
BERT$_{base}$ Pair.	0.872	0.925	0.734	0.747	0.806	0.715
BERT$_{base}$-BiG	**0.899**	**0.961**	**0.825**	**0.836**	**0.847**	**0.765**
- w/o QG	0.891	0.958	0.758	0.772	0.844	0.755
- w/o AG	0.882	0.949	0.824	0.833	0.834	0.740
- w/o Contrastive	0.878	0.942	0.818	0.831	0.802	0.690

Table 5: Ablation Study.

than pointwise training. This issue is appropriately solved by our data augmentation strategy by constructing synthesized pseudo-positive samples.

The QG-augmented data contributes much more to the performance on WikiQA than AG, while the AG-augmented data is slightly more beneficial to TREC-QA and ANTIQUE than QG. Since there are only 873 out of 2118 questions containing positive answers in the training data of WikiQA, the rest questions are all underutilized for pairwise training. The synthesized contrastive data by QG effectively evoke such valuable knowledge for learning.

Compared with pointwise training with only the original data, "w/o contrastive learning", where the synthesized QA pairs are regarded as ground-truth positive samples for pointwise training, can achieve relatively better performance. However, there is still a great margin to reach the performance of BERT$_{base}$-BiG. The results indicate that the synthesized QA pairs are actually more relevant than the original negative samples, but the proposed contrastive learning objective can better utilize the synthesized data than treating them as positive samples for pointwise training.

4 Conclusions

In this paper, we study the data augmentation for improving the performance of ranking question answer pairs when encountering data scarceness and label imbalance issues. We propose an easy-to-apply and effective data augmentation strategy, namely Bilateral Generation, with a contrastive learning objective to maximize the utility of limited labeled data. Experimental results show that the proposed method significantly improves the performance of ranking models on three factoid and non-factoid QA benchmark datasets.

The proposed data augmentation strategy can be easily extended to few-shot learning scenario to further handle more severe resource-poor applications, such as domain-specific question answering, e-Commerce question answering, etc.

References

Dong Chen, Shaoliang Peng, Kenli Li, Ying Xu, Jinling Zhang, and Xiaolan Xie. 2020. Re-ranking answer selection with similarity aggregation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020*, pages 1677–1680.

Qin Chen, Qinmin Hu, Jimmy Xiangji Huang, Liang He, and Weijie An. 2017. Enhancing recurrent neural networks with positional attention for question answering. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 993–996.

Yang Deng, Ying Shen, Min Yang, Yaliang Li, Nan Du, Wei Fan, and Kai Lei. 2018. Knowledge as A bridge: Improving cross-domain answer selection with external knowledge. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, pages 3295–3305.

Yang Deng, Yuexiang Xie, Yaliang Li, Min Yang, Wai Lam, and Ying Shen. 2021. Contextualized knowledge-aware attentive neural network: Enhancing answer selection with knowledge. *CoRR*, abs/2104.05216.

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, NAACL-HLT 2019*, pages 4171–4186.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. TANDA: transfer and adapt pre-trained transformer models for answer sentence selection. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 7780–7788.

Jiafeng Guo, Yixing Fan, Xiang Ji, and Xueqi Cheng. 2019. Matchzoo: A learning, practicing, and developing system for neural text matching. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019*, pages 1297–1300.

Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W. Bruce Croft. 2020. ANTIQUE: A nonfactoid question answering benchmark. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020*, pages 166–173.

Sawan Kumar, Shweta Garg, Kartik Mehta, and Nikhil Rasiwasia. 2019. Improving answer selection and answer triggering using hard negatives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 5910–5916.

Tuan Manh Lai, Trung Bui, and Sheng Li. 2018. A review on deep learning techniques applied to answer selection. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, pages 2132–2144.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 7871–7880.

Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. 2020. Training curricula for open domain answer re-ranking. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020*, pages 529–538.

Sewon Min, Min Joon Seo, and Hannaneh Hajishirzi. 2017. Question answering through transfer learning from large fine-grained supervision data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, pages 510–517.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*, pages 2227–2237.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, pages 2383–2392.

Ying Shen, Yang Deng, Min Yang, Yaliang Li, Nan Du, Wei Fan, and Kai Lei. 2018. Knowledge-aware attentive neural network for ranking question answer pairs. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018*, pages 901–904.

Yi Tay, Minh C. Phan, Anh Tuan Luu, and Siu Cheung Hui. 2017. Learning to rank question answer pairs with holographic dual LSTM architecture. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 695–704.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Hyperbolic representation learning for fast and efficient neural question answering. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018*, pages 583–591.

Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. 2007. What is the jeopardy model? A quasi-synchronous grammar for QA. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 22–32.

Liu Yang, Qingyao Ai, Jiafeng Guo, and W. Bruce Croft. 2016. anmm: Ranking short answer texts with attention-based neural matching model. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016*, pages 287–296.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 2013–2018.

Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2019. A compare-aggregate model with latent clustering for answer selection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019*, pages 2093–2096.

Haotian Zhang, Jinfeng Rao, Jimmy J. Lin, and Mark D. Smucker. 2017. Automatically extracting high-quality negative examples for answer selection in question answering. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 797–800.

Wenxuan Zhang, Yang Deng, and Wai Lam. 2020. Answer ranking for product-related questions via multiple semantic relations modeling. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020*, pages 569–578.