Improving Answer Selection and Answer Triggering using Hard Negatives

Sawan Kumar *
Indian Institute of Science, Bangalore
sawankumar@iisc.ac.in

Shweta Garg †
Search Customer Experience, Amazon
shwegarg@amazon.com

Kartik Mehta
India Machine Learning, Amazon
kartim@amazon.com

Nikhil Rasiwasia
India Machine Learning, Amazon
rasiwasi@amazon.com

Abstract

In this paper, we establish the effectiveness of using hard negatives, coupled with a siamese network and a suitable loss function, for the tasks of answer selection and answer triggering. We show that the choice of sampling strategy is key for achieving improved performance on these tasks. Evaluating on recent answer selection datasets - InsuranceQA, SelQA, and an internal QA dataset, we show that using hard negatives with relatively simple model architectures (bag of words and LSTM-CNN) drives significant performance gains. On InsuranceQA, this strategy alone improves over previously reported results by a minimum of 1.6 points in P@1. Using hard negatives with a Transformer encoder provides a further improvement of 2.3 points. Further, we propose to use quadruplet loss for answer triggering, with the aim of producing globally meaningful similarity scores. We show that quadruplet loss function coupled with the selection of hard negatives enables bag-of-words models to improve F1 score by 2.3 points over previous baselines, on SelQA answer triggering dataset. Our results provide key insights into answer selection and answer triggering tasks.

1 Introduction

Question answering (QA) is an active field of research, drawing attention from the natural language processing (NLP) and information retrieval (IR) community. Selection-based QA is the task of selecting an answer for a given question, from a set of candidate answers. Two tasks have been proposed for selection-based QA. Given a question and an answer candidate pool, **answer selection** is the task of ranking valid answers higher than irrelevant answers, where it is assumed that there is at least one valid answer in the candidate

pool. **Answer triggering**, defined for cases where the candidate pool may or may not have a valid answer, is the task of finding a valid answer, while being allowed to abstain.

Recently released datasets (Feng et al., 2015; Yang et al., 2015; Jurczyk et al., 2016) have led to significant development in end-to-end neural network Selection-based QA. These networks are learned by posing answer selection as either a ranking problem or a classification problem. When posed as a ranking problem, a common choice is to use a triplet loss defined over a question, a correct answer and a (usually sampled) negative answer. Triplet loss penalises relative distances between positive and negative pairs, making the choice of negatives critical. Recent progress in image understanding tasks suggests that learning in ranking tasks can be enhanced by focusing on sampling strategies in conjunction with the loss function employed (Hermans et al., 2017; Chen et al., 2017). In this work, we investigate the expressive power of siamese architectures (Bromley et al., 1994) coupled with hard negatives (i.e., difficult negative samples for the model being learnt), against more sophisticated models. Specifically, all our models ignore any interaction between question and answer representations, and use a shared encoder to independently encode questions and answers. Siamese architectures provide low-latency solutions in the presence of large candidate answer pools (by caching candidate representations).

In this paper, we first show that using hard negatives with vanilla neural architectures improves over previously reported results on InsuranceQA answer selection task (Feng et al., 2015). The proposed strategy achieves P@1 of 73.3%, as compared to the previously reported 71.1% (dos Santos et al., 2016). Further, using a Transformer encoder coupled with the proposed strategy achieves

---

P@1 of 75.6%. Next, we propose the use of quadruplet loss (Chen et al., 2017) for answer triggering. We show that by employing online selection of hard negatives with quadruplet loss, bag-of-words models improve upon previous baselines on SelQA answer triggering task (Jurczyk et al., 2016). The proposed strategy achieves an F1 score of 53.21, as compared to the previously reported 50.89 (Jurczyk et al., 2016).

We note that the performance gains demonstrated in this paper are obtained with a restricted class of models. Specifically, we restrict to siamese networks, and don't model any interaction between question and answer representations. The previously reported results which we compare against use architectures that are not restricted to be siamese, and model the interaction between question and answer representations.

2 Related Work

For answer selection, CNN, LSTM and LSTM-CNN architectures (Tan et al., 2015; dos Santos et al., 2016), trained with a triplet loss, have been explored. Recent focus has been on designing the interaction layer, between a question and a candidate answer, through various models of attention (Yu et al., 2014; Rao et al., 2016; Zhou et al., 2015; Zhang et al., 2017; Bian et al., 2017). Compare-aggregate architectures have also been studied as sentence matching models and applied to answer selection (Wang and Jiang, 2016; Wang et al., 2017). Similarly, for answer triggering, different architectural changes have been proposed (Zhao et al., 2017; Acheampong et al., 2016; Gupta et al., 2018; Li and Wu, 2017). Often, the architectural advances are coupled with subtle changes in the training process. For example, dos Santos et al. (2016) incorporate a form of negative mining. This makes it difficult to separate the benefits gained from the mining strategy from the architectural advances.

Triplet loss and mining of hard negatives have been studied for computer vision tasks with triplet networks being popular for estimating feature embeddings for images. The key issue with triplet networks is that, for a training set with N samples, the number of triplets is cubic in N. Training becomes intractable for modest sizes of the training set. A solution in the form of importance sampling has been studied. Schroff et al. (2015) learn image embeddings using triplet loss, trained using moderately hard negatives. While other approaches have been studied for the task of person re-identification, which combine classification with a verification loss, Hermans et al. (2017) show that a vanilla CNN with triplet loss and the right sampling strategy could outperform the best models at the time. Wu et al. (2017) present a distance-weighed sampling approach.

Triplet loss, while useful for ranking tasks, doesn't produce globally meaningful scores for tasks such as person re-identification. This can be attributed to the fact that triplet loss doesn't try to learn a global threshold to separate all inter-class pairs from all intra-class pairs, and learns only relative distances with respect to an anchor (a question, in the context of this paper). This has been addressed by accounting for the global structure of the embedding space (Kumar et al., 2016; Ustinova and Lempitsky, 2016), and directly optimizing for inter-class distances to be larger than intra-class distances. Quadruplet loss (Chen et al., 2017) prvoides another way to produce globally meaningful scores, and is suitable for our setting. Quadruplet loss extends triplet loss to ensure smaller intra-class distances and larger inter-class distances. This is achieved by additionally penalizing positive and negative pairs with different probes (questions, in the context of this paper).

While selection of hard negatives has been used for problems in computer vision, its usefulness has not been evaluated for QA tasks. In this work, we first present the selection of hard negatives with triplet loss for answer selection. Next, we show that quadruplet loss is suitable for the task of answer triggering, and when coupled with online selection of hard negatives, improves over previous baselines. As far as we know, we are the first to use quadruplet loss for question answering.

3 Data

Dataset	#Q	#QA	L_q	L_a
InsuranceQA	12887	18540	7.15	95.54
SelQA	5524	6390	11.02	25.39
LargeQA	341869	502763	14.06	30.34

Table 1: Comparison of answer selection datasets. **#Q**: Number of questions; **#QA**: Number of correct question-answer pairs; L_q/L_a: Average number of tokens in a question/answer.

Insurance QA (Feng et al., 2015), a domain-specific non-factoid QA dataset, is suitable for

evaluating answer selection. SelQA (Jurczyk et al., 2016), a recent open-domain factoid QA dataset provides data for evaluating both answer selection and answer triggering.

While these datasets provide standardised comparison, we also evaluate our methods on a large internal answer selection dataset, LargeQA, created using Community Question-Answers (CQnA) asked on a website. The dataset was created similar to InsuranceQA, with the size of candidate pool fixed to 100 answers in test sets.

Table 1 presents some statistics on the training data used in these datasets.

We don't include WikiQA (Yang et al., 2015) and TrecQA (Wang et al., 2007) datasets which have been previously used to report improvements in answer selection. This is due to the large variance in our experiments as well as in the reproduction of existing methods, also noted by Crane (2018), perhaps due to their smaller sizes.

4 Method

For selection-based QA , the training data X, can be characterized as a list of questions $Q = \{q_1, q_2, \ldots , q_s\}$, and sets of correct answer(s) $A_q = \{a_1, a_2, \ldots , a_p\}$ for each question $q \in Q$.

In the following, we discuss online selection of hard negatives with triplet loss for answer selection, and with quadruplet loss for answer triggering. In each case, a siamese architecture is used where questions and answers are encoded independently by identical copies of a neural network denoted by f. Cosine similarity is used to compute the similarity between a question and an answer: $S(x,y) = \text{cosine}(f(x), f(y))$.

4.1 Answer Selection

Siamese networks can be trained for answer selection using triplets (q, a^+, a^-), where $a^+ \in A_q$ is a correct answer and $a^- \notin A_q$ an incorrect answer chosen randomly from the entire answer pool. A triplet loss with margin m is formulated as:

$$L_{\text{triplet}} = \max(0, m - S(q, a^+) + S(q, a^-)). \quad (1)$$

In this work, we employ online selection of hard negatives within a batch of sampled question-answer pairs. The model is trained using batch gradient descent, with batches $B = \{(q_i, a_i, a_i^H) : 1 \le i \le b\}$, where b is the batch size, $a_i \in A_{q_i}$ and a_i^H is the hardest negative answer chosen as:

$$a_i^H = \underset{j : a_j \notin A_{q_i},\ 1 \le j \le b}{\operatorname{argmax}} (S(q_i, a_j)). \quad (2)$$

Selection of negatives from within a batch of sampled question-answer pairs has several advantages. First, it does not require any extra computation from the embedding model. Second, there is no need to employ additional heuristics to avoid the hardest negatives. Selection from within a stochastic batch ensures that the selected hardest negatives in a batch are not dominated by false negatives or noisy hard negatives during training (See Appendix C for experimental results).

4.2 Answer Triggering

Answer triggering differs from answer selection in that we need to identify if a valid answer is present in the candidate pool. Triplet loss is not suitable as it promotes scores to be discriminative only relative to a given question. While classification based methods have been studied, we hypothesize that globally meaningful scores can be obtained in a ranking setup. Essentially, we need the similarity scores for all wrongly paired question-answers to be smaller than the scores for correct question-answer pairs. We propose to use quadruplet loss for answer triggering. Quadruplet loss, introduced by Chen et al. (2017) for Person re-identification, aims at keeping all inter-class distances larger than all intra-class distances. This, we believe, is what we need to obtain globally meaningful scores for answer triggering.

The network can be trained using quadruplets of the form (q, a^+, a^-, q'), where $a^+ \in A_q$, $a^- \notin A_q$, and q' is a negative question for both a^+ and a^- chosen randomly from the entire question pool, with $a^+ \notin A_{q'}$, $a^- \notin A_{q'}$. Quadruplet loss for answer triggering can be formulated as:

$$L_{\text{quad}} = \max(0, m_1 - S(q, a^+) + S(q, a^-)) + \\ \max(0, m_2 - S(q, a^+) + S(q', a^-)) \quad (3)$$

where m_1 and m_2 are fixed margins.

As with triplet loss, we believe the selection of negatives is critical for learning with quadruplet loss. We propose to select both the negative answer and the negative question using online selection of hard negatives. The model is trained using batch gradient descent with batches $B = \{(q_i, a_i, a_i^H, q_i^H) : 1 \le i \le b\}$, where b is the batch size, $a_i \in A_{q_i}$, a_i^H is chosen as in Equation 2, and q_i^H is selected to be the hardest negative question for a_i^H:

$$q_i^H = \underset{k : a_i^H \notin A_{q_k},\ 1 \le k \le b}{\operatorname{argmax}} (S(q_k, a_i^H)). \quad (4)$$

Method	P@1		
	Dev	Test1	Test2
Attentive Pooling + CNN (dos Santos et al., 2016)	**68.8**	69.8	66.3
Attentive Pooling + BiLSTM (dos Santos et al., 2016)	68.4	**71.7**	**66.4**
Max-Pooling + Random Negatives	61.5	61.6	57.9
Max-Pooling + Hard Negatives	64.4	65.3	64.4
LSTM-CNN + Random Negatives	68.6	70.3	67.2
LSTM-CNN + Hard Negatives	**72.5**	**73.3**	**69.1**
Transformer encoder + Random Negatives	65.8	67.1	63.8
Transformer encoder + Hard Negatives	**75.7**	**75.6**	**73.4**

Table 2: *Answer selection on InsuranceQA*. **Top**: Baselines. **Middle** & **Bottom**: Our Max-Pooling, LSTM-CNN and Transformer encoder results. When coupled with hard negatives, simpler architectures such as Max-Pooling become competitive, while LSTM-CNN improves over previously reported results. The effectiveness of using hard negatives, compared to random negatives, is also evident when training the Transformer encoder model.

During inference, for a given question, we predict the highest scoring answer, provided the score exceeds the threshold. The optimal threshold is obtained using performance on the development set.

5 Results and Analysis

We present quantitative results using MRR and P@1 for answer selection on InsuranceQA, LargeQA and SelQA, and F1 score (Yang et al., 2015) for answer triggering on SelQA, followed by an ablation study of the proposed method for answer triggering.

5.1 Experimental Setup

Following Tan et al. (2015), we experiment with an LSTM-CNN model, where a CNN layer is employed on top of LSTM encoding of the input sentence, followed by Max Pooling to obtain a fixed length representation for the sentence.

Motivated by the performance of bag-of-words like sentence representations (Arora et al., 2016), we also experiment with distributional bag-of-words models where sentence embeddings are obtained by pooling across the dimensions of corresponding word embeddings. In particular, we used Max-Pooling and Max-Min-Pooling, the latter being obtained by concatenating the outputs of Max and Min Pooling.

Finally, we experiment with the Transformer encoder (Vaswani et al., 2017; Devlin et al., 2019) for answer selection.

Training details are available in Appendix A.

5.2 Answer Selection

InsuranceQA: LSTM-CNN model coupled with hard negatives outperforms the best reported numbers on both test sets by a significant margin (1.6

and 2.7 points gain in P@1 with p-values of 0.066 and 0.008 on Test1 and Test2 respectively) on InsuranceQA (Table 2). Finally, the Transformer encoder model with hard negatives provides significant performance gains for the task (3.9 and 7.0 points gain in P@1), while still learning with siamese networks without any interaction between question and answer representations.

Method	P@1	MRR
LSTM-CNN + Random Negatives	+0	+0
LSTM-CNN + Hard Negatives	**+2.5**	**+0.015**

Table 3: *Answer selection on LargeQA*; reported numbers are relative to the baseline.

LargeQA: The gains due to hard negatives are also observed on LargeQA (Table 3), with an improvement of 2.5 in P@1 and 0.015 in MRR.

Model	MRR
RNN + Attention Pooling (Jurczyk et al., 2016)	0.876
CNN-Hinge (dos Santos et al., 2017)	0.881
ACNN (Shen et al., 2018)	0.880
AdaQA (Shen et al., 2018)	0.910
DRCN (Kim et al., 2019)	**0.930**
Max-Pooling + Random Negatives	0.843
Max-Pooling + Hard Negatives	0.896

Table 4: *Answer selection on SelQA*. **Top**: Baselines. **Bottom**: Max-Pooling results from this work.

SelQA: For answer selection on SelQA (See Table 4), Max-Pooling models performed best among the models we investigated, perhaps due to the smaller size of the dataset (other results omitted).

Key Observations: First, siamese bag-of-words models, coupled with hard negatives, are competitive with models which rely on various ways of question-answer interaction. Second, siamese networks coupled with hard negatives are particularly suited for domain specific QA (In-

suranceQA). We believe open-domain QA has a greater need for interaction between question and answer representations. Third, we believe employing hard negatives is especially useful for the Transformer encoder model, which has a much larger set of parameters. With the Transformer encoder model, there is an improvement of 8.5 points on InsuranceQA Test1 dataset when using hard negatives as opposed to using random negatives. With the LSTM-CNN model, the corresponding gain is 3.0 points.

5.3 Answer Triggering

Method	F1	
	Dev	Test
CNN (Jurczyk et al., 2016)	49.16	50.89
RNN + AP (Jurczyk et al., 2016)	44.02	45.67
Max-Min-Pooling + Quadruplet loss + Hard negatives	**54.95**	**53.21**

Table 5: *Answer triggering on SelQA*. **Top**: Baselines. **Bottom**: Our approach using quadruplet loss with Max-Min-Pooling model. The choice of the loss function and the sampling strategy proposed in this paper is justified by the improved results obtained using a relatively simpler Max-Min-Pooling model.

The proposed approach with the Max-Min-Pooling model for answer triggering improves upon the previously reported baselines significantly (2.32 points) on SelQA (See Table 5).
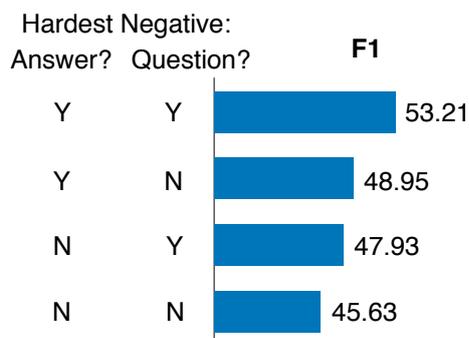

Figure 1: *Ablation study for answer triggering.* The columns on the left indicate whether online selection of hardest negative answer/question was done. To obtain best results with quadruplet loss, it is critical to sample both hard negative answers and hard negative questions.

An ablation study (Figure 1) reveals that the selection of both hardest negative answers and questions is needed to get the majority improvement. Quadruplet loss with the Max-Min-Pooling model without hard negatives is competitive with base-

lines (Table 5), justifying its usage for answer triggering.

Further, we also compared against using triplet loss with the Max-Min-Pooling model for answer triggering. We found that quadruplet loss outperforms triplet loss (F1 of 51.89 on development set, and 50.08 on Test set) by 3.13 points in F1.

6 Conclusion & Future Work

We have shown that selection of hard negatives is a powerful tool for answer selection. We improve over previously reported results on recent benchmarks using siamese architectures and hard negatives, outperforming interaction-based models. We show the generality of the approach using shallow as well as deep neural network models.

For answer triggering, we presented results supporting the hypothesis that quadruplet loss with hard negatives is suitable for the task, and improves upon previous baselines. Our ablation study confirms the importance of using hard negatives.

As future work, we plan to further investigate the generality of the approach with other tasks and base models.

Acknowledgments

We would like to thank the reviewers for their valuable feedback.

References

Kingsley N Acheampong, Zhen-Hao Pan, Er-Qiang Zhou, and Xiao-Yu Li. 2016. Answer triggering of factoid questions: A cognitive approach. In *Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2016 13th International Computer Conference on*, pages 33–37. IEEE.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings.

Weijie Bian, Si Li, Zhao Yang, Guang Chen, and Zhiqing Lin. 2017. A compare-aggregate model with dynamic-clip attention for answer selection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1987–1990. ACM.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744.

Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. 2017. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412.

Matt Crane. 2018. Questionable answers in question answering research: Reproducibility and variability of published results. *Transactions of the Association for Computational Linguistics*, 6:241–252.

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

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. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 813–820. IEEE.

Deepak Gupta, Sarah Kohail, and Pushpak Bhattacharyya. 2018. Combining graph-based dependency features with convolutional neural network for answer triggering. *CoRR*, abs/1808.01650.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737.

Tomasz Jurczyk, Michael Zhai, and Jinho D Choi. 2016. Selqa: A new benchmark for selection-based question answering. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 820–827. IEEE.

Seonhoon Kim, Inho Kang, and Nojun Kwak. 2019. Semantic sentence matching with densely-connected recurrent and co-attentive information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6586–6593.

BG Kumar, Gustavo Carneiro, Ian Reid, et al. 2016. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5385–5394.

Wei Li and Yunfang Wu. 2017. Hierarchical gated recurrent neural tensor network for answer triggering. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 287–294. Springer.

Jinfeng Rao, Hua He, and Jimmy Lin. 2016. Noise-contrastive estimation for answer selection with deep neural networks. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1913–1916. ACM.

Cícero Nogueira dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. 2016. Attentive pooling networks. *CoRR*, abs/1602.03609.

Cícero Nogueira dos Santos, Kahini Wadhawan, and Bowen Zhou. 2017. Learning loss functions for semi-supervised learning via discriminative adversarial networks. *CoRR*, abs/1707.02198.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Dinghan Shen, Martin Renqiang Min, Yitong Li, and Lawrence Carin. 2018. Learning context-sensitive convolutional filters for text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1839–1848, Brussels, Belgium. Association for Computational Linguistics.

Ming Tan, Bing Xiang, and Bowen Zhou. 2015. Lstm-based deep learning models for non-factoid answer selection. *CoRR*, abs/1511.04108.

Evgeniya Ustinova and Victor Lempitsky. 2016. Learning deep embeddings with histogram loss. In *Advances in Neural Information Processing Systems*, pages 4170–4178.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

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

Shuohang Wang and Jing Jiang. 2016. A compare-aggregate model for matching text sequences. *CoRR*, abs/1611.01747.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4144–4150. AAAI Press.

Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krähenbühl. 2017. Sampling matters in deep embedding learning. In *Proc. IEEE International Conference on Computer Vision (ICCV)*.

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*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep learning for answer sentence selection. *CoRR*, abs/1412.1632.

Xiaodong Zhang, Sujian Li, Lei Sha, and Houfeng Wang. 2017. Attentive interactive neural networks for answer selection in community question answering. In *AAAI*, pages 3525–3531.

Jie Zhao, Yu Su, Ziyu Guan, and Huan Sun. 2017. An end-to-end deep framework for answer triggering with a novel group-level objective. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1276–1282, Copenhagen, Denmark. Association for Computational Linguistics.

Xiaoqiang Zhou, Baotian Hu, Qingcai Chen, Buzhou Tang, and Xiaolong Wang. 2015. Answer sequence learning with neural networks for answer selection in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 713–718, Beijing, China. Association for Computational Linguistics.