Multi-Strategies Method for Cold-Start Stage Question Matching of rQA Task

Dongfang Li[1], Qingcai Chen[1], Songjian Chen[2], Xin Liu[1], Buzhou Tang[1], and Ben Tan[2]

[1] Shenzhen Calligraphy Digital Simulation Technology Lab, Harbin Institute of Technology (Shenzhen)
{crazyofapple,hit.liuxin}@gmail.com, {qingcai.chen, tangbuzhou}@hit.edu.cn
[2] Technology Engineering Group, Tencent
{firstchen, bentan}@tencent.com

Abstract. Sentence Semantic Equivalence Identification (SSEI) plays a key role in the Retrieval-based Question Answering (rQA) systems. Nevertheless, for the resource limitation of many real applications, even the best SSEI models may underperform. To enhance the performance, this paper firstly proposes a novel deep neural network named Densely-connected Fusion Attentive Network (DFAN). The key idea behind our model is to learn the interactive semantic information with densely connection and fusion attentive mechanism. Secondly, for the limitation of the available corpus for the given domain, we add an auxiliary classification task, which categorizes questions into domain-specific classes. And pretrained sentence embeddings learned from large unlabeled pairs are integrated as the weakly supervised learning strategy. We conduct experiments on datasets SNLI, Quora, and the domain corpus provided for a real rQA system, achieving competitive results on all. For the domain corpus, as the best F1 value of 93.29% reached by the proposed DFAN model with additional strategies, the measure hit@1 for the real rQA systems is 52.02%, which outperforms all compared methods. This result also shows that, getting satisfied performance for a real rQA system remains a challenging natural language processing task.

Keywords: Enhanced neural approach · Retrieval-based question answering · Sentence matching

1 Introduction

Identifying the semantic equivalence of two sentences is one of the essential tasks for Retrieval-based Question Answering (rQA) systems, which is also known as Sentence Semantic Equivalence Identification (SSEI) [29, 3]. With the development of SSEI techniques, more and more rQA systems are served as domain-specific Automated Customer Service (ACS). However, since most of the SSEI methods are based on supervised deep neural networks [14, 26, 12], the limitation of available corpus becomes the most significant obstacle of building an rQA based ACS system for many specific domains.

To clarify that, we first give the pipeline of rQA in Figure 1. Given question and answer sets (Q_d, A_d) of domain d, for each answer $a \in A_d$, there is a subset $Q_a \subseteq Q_d$

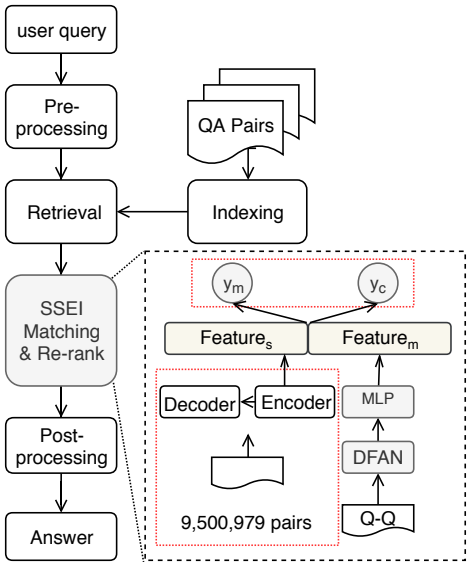

Fig. 1. The pipeline of an rQA system, and the outline of how our model and strategies have been applied.

that could be answered by a. Then the corresponding rQA system is usually composed of the following procedures: 1) index all question and answer (QA, for short) pairs by questions for retrieval, which is usually executed offline; 2) to answer a user question q_u, the rQA system retrieves a candidate intent-similar question set Q_c from indexed QA pairs; 3) the SSEI algorithm is applied for q_u and each question in Q_c to find out the most matching question q^*; 4) the labeled answer for q^* is finally returned as the answer for user question.

Two main issues make it very challenging to reach satisfied performance for an rQA system: 1) the diversity of intents in the user utterances, 2) among the indexed QA pairs of a specific domain, many semantically close questions may correspond to different intents, thus need different answers. To tackle these, the most effective way is increasing the scale of question subset Q_a for each answer a. It not only improves the coverage rate for various intent expressions but also increases the amounts of training data to enhance the SSEI performance of the rQA system significantly. Unfortunately, considering the complexity in real applications and the cost of manually constructing the QA pair set, there are usually very few questions that correspond to the answer, especially at the cold start stage.

In this paper, to improve the performance of rQA systems in case of lacking domain specific SSEI corpus, a novel neural network named Densely-connected Fusion Attentive Network (DFAN) is firstly proposed. This network encodes each sentence with previous densely-connected vectors via a bidirectional recurrent network to generate respective context representation. The fuse gate is used to combine the representation with its self-attention. As shown in [12] and [11], the intuition of employing the fusion

technique in a single sequence is to work as a skip connection, which helps to traverse information in our model. Then we pass these learned vectors to an interaction layer, which performs the word-by-word alignment as attentive information.

Considering that, in a lot of application domains, though the scale of available corpus is limited, there are usually additional in-domain knowledge available, e.g., the class label of a given question, may be provided. To take advantage of this information, a text classification based multitask learning strategy is designed upon the DFAN architecture. Since external knowledge has been proved useful for SSEI task in [28, 16] etc., in this paper, we use a general-purpose encoder-decoder framework to learn a pre-trained sentence encoding model from large-scale unlabeled data. Unlike pre-training language models [22, 10], domain-independent sentence embeddings are generated via this encoder-decoder model as the external supplementary feature for each sentence. Compared to [28] and [16], the advantage of this approach is that it complements external knowledge without human involvement.

The prime contributions of this work are summarized as follows:

- By providing deeper architecture through stacking with densely connection and fusion attentive mechanism, the DFAN can better capture interactive alignment and self semantic information at multiple sentence interactions without relying on the model's pre-training.
- We propose two additional strategies to overcome the limitation of available corpus in real rQA based applications. The sentence encoding model pre-trained on large unlabeled data is used to supply external information for the base model. Moreover, by taking advantage of question categories given in a specific domain, the multitask learning strategy is proposed.
- We conduct experiments on both the public benchmark corpora SNLI, Quora and the rQA corpus constructed from an online deployed ACS system. The proposed DFAN neural network achieves competitive performance in all evaluations. And its strategies-enhanced version gets best results on F1 and hit@1 measures in the last one compared to other supervised and pre-trained models.

2 Related Work

Recently, most supervised SSEI methods are based on sentence interaction. It enables the encoding of more sophisticated matching patterns for various granularity rather than just sentence level. ESIM [7] is composed of the following main components: input encoding, local inference modeling, and inference composition. In local inference modeling, it uses the dot-product attention to composite relationship of the encoded vectors. BiMPM [26] is a bilateral multi-perspective matching model that matches sentences pairs in two directions, from multiple perspectives. The model uses four different ways of sentence interaction instead of the attention weighted information. Meanwhile, unsupervised methods such as word mover's distance (WMD) [17] and smooth inverse frequency embedding (SIF) [2], etc. are also proposed and applicable for SSEI.

To deal with lack of domain specific corpus, some methods of using external information or transfer learning are proposed [28, 16]. [16] used WordNet and relation embeddings additively to measure the semantic similarity among text snippets. [28]

developed a transfer learning framework to take advantage of other domain-specific labeled text pairs, which models domain relationships via shared layers and a trainable weight matrix. Currently, pre-trained language models [22, 10] by leveraging large amounts of unlabeled data bring significant improvement in various NLP tasks. However, the pre-training requires a large training corpus and time-consuming and it does not mean that we do not need to find an efficient end-to-end model or framework for SSEI. Moreover, some strategies can be integrated into these models to improve the learning of text representation [18].

3 Methodology

Figure 1 shows the architecture of the rQA system, supported by the densely-connected fusion attentive network and the integrated strategies. The following parts will present the structure and the strategies in detail.

3.1 Densely-connected Fusion Attentive Network

Embedding Layer In the embedding layer, in order to construct the word representation effectively and informatively, existing pre-trained word embedding, such as Word2vec [19] or Glove [21] vector representations could be combined, with the character features and the exactly matched feature (EM) [6].

Encoder Layer A bidirectional LSTM (BiLSTM) is deployed in the encoder layer to enhance the context influence in both the forward and the backward direction.

Fusion Layer Each word relative position is represented with o after encoding. These representations input to a self-attention layer to calculate the relationship between the words in context. Then the self-attention representation and their original encoded vector are passed into a fuse gate to determine whether the concatenation of input text could achieve a good semantic composition for the single sentence. Unlike previous work [12], our work uses addition connection to consider both the new and the old information that reduces the redundant gate. As an advantage of such modification, we generate the deeper network by keeping the same scale of parameters. The details of self-attention and fuse gate mechanism are as follows:

$$c_{i,j} = f(o_i, o_j), \forall i, j \in [1, ..., l] \tag{1}$$

$$\bar{o}_i = \sum_{j=1}^{l} \frac{exp(c_{i,j})}{\sum_{k=1}^{l} exp(c_{k,j})} o_j \tag{2}$$

$$z_i = tanh(W_1[o_i, \bar{o}_i] + b_1) \tag{3}$$

$$r_i = \sigma(W_2[o_i, \bar{o}_i] + b_2) \tag{4}$$

$$\hat{o}_i = r_i \odot o_i + (1 - r_i) \odot z_i \tag{5}$$

where $f(o_i, o_j) = [o_i, o_j, o_i \odot o_j]$, and $o \in R^{l*d}$ is the vector output from the encoder layer in words sequential order, $\hat{o} \in R^{l*d}$ is the output of the fusion layer and W_1, W_2, b_1, b_2 are trainable weights, σ is sigmoid activation function, l refers to max sentence length, d refers to size of hidden unit in encoder layer. In practice, two sentences can obtain respective output by the same operation.

Interaction Layer Then we apply inter-attention operation to interact two sentences to get attentive vectors respectively. These attentive vectors represent soft alignment between two sentences as follows:

$$e_{i,j} = g(\hat{o}_i, \hat{o}_j) \tag{6}$$

$$\tilde{o}_i^1 = \sum_{j=1}^{l} \frac{exp(e_{i,j})}{\sum_{k=1}^{l} exp(e_{i,k})} \hat{o}_j^2 \tag{7}$$

$$\tilde{o}_j^2 = \sum_{i=1}^{l} \frac{exp(e_{i,j})}{\sum_{k=1}^{l} exp(e_{k,j})} \hat{o}_i^1 \tag{8}$$

where $g(\hat{o_i}, \hat{o_j}) = \hat{o_i} \odot \hat{o_j}$, and $\tilde{o_i}$ is the output of the interaction layer.

Aggregated Layer We aggregate the matching information from the interaction layer by performing several operations. All the operations are performed element-wise. Let o, $\tilde{o}$ be the input respectively, two representations are concatenated with their subtractions and their multiplications together as the feature vector v, i.e.,

$$v = [o; \tilde{o}; o - \tilde{o}; o \odot \tilde{o}] \tag{9}$$

Inspired by ResNet [13] and DenseNet [15], we also concatenate the input of the current encoder layer t with v as an additional connection. Since we repeat middle layers 3 times, the input of each encoder layer would be different. For example, in the first time of repetition, t is the output of the embedding layer. In the next iteration, t is the output of the previous aggregated layer.

Output Layer After aggregating the information from the previous layer, we convert the representations of all positions in two sentences to a fixed-length vector by max pooling and mean pooling operations. The low-dimensional result will be fed into two fully connected layers to calculate the relationship between two sentences.

3.2 Strategies

In this paper, two strategies are proposed to supply more comprehensive semantic knowledge for the base model, including weakly supervised features and a related auxiliary task.

Pre-trained Feature Extractor With the development of community question answering web sites such as Yahoo! Answers[3], Baidu Zhidao[4], etc., tremendous amount of QA pairs have been produced by community users. Through well-designed methods, these type of QA pairs could be a very useful complement for the manually constructed domain QA corpus. In this paper, we use the Baidu Zhidao as the complementary source. Each of the question on the Baidu Zhidao web site and its possible matching questions from "Other similar questions" section reported on the web site are crawled. We crawl 9,500,979 question-question pairs under broad topics, such as "scientific education", "laws and regulations", "social and livelihood" etc., as the training set. For example, we assume the question "how to correctly understand the concept of deep learning" is a duplicate sentence to "what is deep learning". After filtering and pre-processing text, we train an attention-based Seq2Seq model using these pairs. Here the hypothesis is that, in a text pair, one's intent information can be generated by the other one. We use a general-purpose encoder-decoder framework provided by [5] for training. All hyperparameters are configured as default in the original code.[5] After this Seq2Seq model is trained, it is used to generate the weakly supervised representation vectors for each sentence. In our experiments, the last state of the encoder is used as the pre-trained feature and is directly concatenated to the middle representation generated by the last aggregated layer. Though other combinations are tried, there are no obvious positive gains acquired and thus not reported here.

Auxiliary Task In many cases, though abundant of various expression questions for a given answer are hard to collect for a real application, the domain-specific categories may available for each question and answer in the QA database. To full use of such information in SSEI modeling, inspired by [8], we add an auxiliary task into our model, i.e., text classification for each sentence that is simultaneously trained with text matching task. Learning with auxiliary tasks restricts the parameter space during training, which can be regarded as a regularizer. In spite of being seemingly unrelated, text classification task is expected to assist in finding a robust and rich semantic representation of the input text, from which improve the ultimately desired main task performance by forcing the network to generalize to other tasks. For details, the mean-pooling and max-pooling vectors of each sentence are respectively passed to a fully-connected layer with ReLU activation followed by another fully-connected layer. Softmax function is applied to predict the most suitable class of each sentence in the final layer.

4 Experiments

We compared our model with other methods on three public datasets, two public datasets in English and one dataset in Chinese sampled from an online domain-specific rQA system. On the two public English datasets, the current state-of-the-art models were selected for comparison, while on the Chinese dataset, the following models were selected

[3] https://answers.yahoo.com
[4] https://zhidao.baidu.com/
[5] https://github.com/google/seq2seq

for comparison: WMD [17], ABCNN [27], DecompATT [20], BiMPM [26], ESIM [7], BiLSTM+MaxPool [9] and BERT [10].

4.1 Datasets

The Stanford Natural Language Inference Corpus [4] The train set consists of 549,367 text pairs, while development set has 9,842 pairs and test set has 9,824 pairs.

Quora Question Pairs [1] In the end, we have 384,348 pairs for training, 5,000 matched pairs and 5,000 mismatched pairs for development, and another 5,000 matched pairs and 5,000 mismatched pairs for test.

Szga FAQ Corpus This corpus is in Chinese. We collect FAQs from the real online customer service system of a public sector, and grouped together questions by the same answer. Each group was double-checked by human annotators. All question pairs in the same group form positive samples. The negative samples are constructed in the following way: for each question, find out the top k (set to 100 in this study) questions not in the same group as it by BM25-based searching from all questions. In order to simulate the lack of data, here we set the maximum group size to be 3. In the end, we obtain a dataset of 21,357 matched pairs and 53,504 mismatched pairs, which are randomly split into two parts: a training set of 20,237 matched pairs and 350,130 mismatched pairs, and a test set of 1,120 matched pairs and 18,374 mismatched pairs.

4.2 Results of the DFAN model

Table 1. Results for natural language inference on the SNLI dataset.

Models	Accuracy (%)
ESIM [7]	88.0
DIIN [12]	88.0
MwAN [23]	88.3
CAFE [24]	88.5
KIM [16]	88.6
DFAN	88.6
DFAN+BERT$_{embedding}$	89.1
MT-DNN [18]	**91.1**

Table 2. Results for paraphrase identification on the Quora Question Pairs dataset. The first 8 rows are reported in [12].

Models	Accuracy (%)
Siamese-CNN	79.60
MP-CNN	81.38
Siamese-LSTM	82.58
MP-LSTM	83.21
L.D.C	85.55
BiMPM	88.17
pt-DecAttchar.c	88.40
DIIN [12]	89.06
MwAN [23]	89.12
DFAN	**89.91**

Table 1 shows the accuracies of DFAN and other state-of-the-art models on the SNLI test set. DFAN achieves an accuracy of 88.6%, better than most of the models for comparison. KIM that used external linguistic inference knowledge is the model of the same accuracy as DFAN. The one model better than DFAN is MT-DNN, which is a multi-task fine-tuned model based on pre-trained BERT. It may be unfair to compare

DFAN with KIM and MT-DNN, as we know that external knowledge, multi-task and pre-training can bring extra improvement. For example, when integrating BERT embeddings into DFAN, we obtained an accuracy of 89.1%, higher than the base DFAN model by 0.5%. Table 2 shows the results of our base model on the Quora Question Pair dataset. [6] We achieve the improved results of 89.91% accuracy, surpassing the previous works like MwAN.

4.3 Comparisons between Strategies

Table 3. Results for sentence semantic equivalence identification on Szga FAQ corpus. *Seq2Seq*, *AUX* indicate the pre-trained feature extractor, the auxiliary task respectively.

Models	hit@1	P	R	F1
WMD	19.74	/	/	/
DecompATT	36.92	88.79	44.55	59.33
ABCNN	39.47	**95.41**	57.50	71.75
BiLSTM+MaxPool	41.85	79.91	68.00	73.48
BiMPM	45.11	77.88	77.95	77.91
ESIM	48.28	93.11	85.62	89.21
BERT_{base}	48.63	94.27	88.21	91.14
DFAN	48.37	93.04	85.69	89.22
DFAN+Seq2Seq	49.77	94.53	86.34	90.24
DFAN+AUX	51.98	92.11	90.80	91.46
DFAN+AUX+Seq2Seq	**52.07**	91.64	**95.00**	**93.29**

Intrinsic Evaluation Table 3 shows the results of our base model on the Szga FAQ corpus. The F1 of DFAN is 89.21%, higher than that of all other models except BERT_{base}[7]. When Seq2seq features and the other auxiliary task are separately added, DFAN obtain improvements of 1.03% and 2.25% in F1 respectively. When both of them are added together, DFAN is further improved and achieves the highest F1 of 93.29%, higher than BERT_{base} by 2.15%. That indicates that the effectiveness of the proposed strategies of weakly supervised learning and the related auxiliary task.

Extrinsic Evaluation The results of the intrinsic evaluation and the extrinsic evaluation are pretty consistent. But there are exceptions, fine-tuned BERT_{base} has 91.14% F1 value but doesn't achieve the higher hit@1 compared to DFAN+Seq2Seq. There are two reasons why some models have an inconsistent situation. On the one hand, the test set has labeling noise. On the other hand, the two metrics may not be fully aligned. Compared with other methods, our model performs much better on this dataset.

[6] The result of BERT and MT-DNN in this dataset is 89.3% and 89.6%, as they used other data split of [25].

[7] We use the pre-trained model released by authors. There is only a base model in Chinese.

Through DFAN is slightly lower than fine-tuned BERT$_{base}$, the model with the combination of the pre-trained feature extractor and the auxiliary task achieves the best performance in the extrinsic evaluation. Furthermore, to compare the predictive accuracy of the two methods, we conduct the McNemar's test between results of our base model and the strategies-enhanced DFAN+AUX+Seq2Seq model and calculate the p-value equals 0.037, which shows that our strategies can give the model a significant improvement in the real application.

5 Analysis and Discussion

5.1 Ablation Study on DFAN

Table 4. Ablation study.

DFAN	88.6
rm one middle layer	88.3
rm two middle layers	86.2
w/o EM	88.3
w/o fuse gate	87.7
w/o dense connect	88.2
w/o max pool	87.5
w/o mean pool	87.5
w/o dot att. w/ cosine att.	87.2

Table 5. Pairs in the Szga FAQ corpus with different question size.

# samples	positive	negative
upper3	20,237	350,130
upper5	30,073	397,316
upper10	41,058	423,749
upper30	68,416	443,807
upper50	89,367	449,250

With the purpose of examining the effectiveness of each component of our base model, we conduct an ablation study on the SNLI test set, as shown in Table 4. We use the validation score on the development set as the standard for model selection. First, we report the performance of models having different number of middle layers. Then we explore how exact match signal contributes to the model. The accuracy of our base model degrades to 88.3% on SNLI test set slightly. It proves that a simple feature can help the model to understand the text semantic similarity better. Then we remove the fuse gate and obtain 87.7% on test set. The result implies the addition of the fuse gate can have an effective impact to capture semantic information. Then we remove our densely connection; the accuracy is getting lower. To verify the effectiveness of the two pooled operations, we first replace the output layer with only the max pooling. Next, replace the output layer with only the mean pooling. We find that the contribution of these two pooling form is almost equal. To show that the impact of different attention, we replace the dot attention matrix with cosine similarity matrix. The results show that dot attention has a stronger influence than cosine-attention for modeling text semantic similarity.

5.2 Effect of Data Size

To further investigate the performance of our base model and strategies under different amount of training data, we compare the extrinsic evaluation performance of different

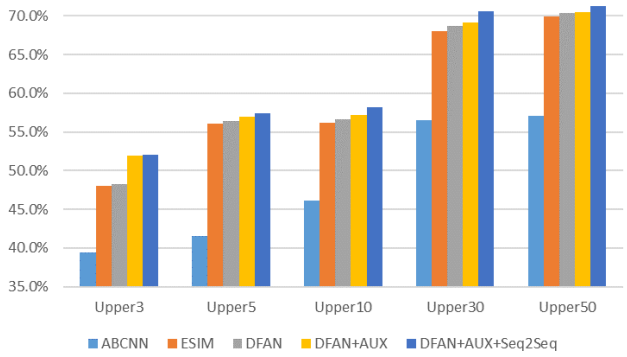

Fig. 2. Hit@1 on the Szga FAQ corpus with different question size.

models. We assume the number of the equivalence questions set of the same answer in our FAQ set is 3, 5, 10, 30, 50 at most respectively. Therefore, the number of matched pairs in the training set would be different. As we can see in Table 5, matched pairs increases with the number of the equivalence questions. The overall performance of ABCNN, ESIM, DFAN, and strategies-enhanced DFAN are shown in Figure 2. Compared with other models, our strategies-enhanced DFAN performs best with a small amount of data. And we observe that our stratgies-enhance model is superior to others consistently. The results indicate that the model with proposed strategies can learn better representations in different scenarios.

6 Conclusion and Future Work

In this paper, we firstly clarified the task and challenges of an rQA system. Then a deep and densely connected neural network DFAN is proposed. Its performance is verified through two public datasets SNLI and Quora Question Pairs. On the corpus that is constructed from the real ACS system to evaluate the overall performance of an rQA system, we show the efficacy of the DFAN model and two additional strategies proposed to tackle the corpus lacking issue. Finally, the proposed method has been deployed as an online ACS system and is serving for millions of requests each day. While the best SSEI performance reached by our method is 93.29%, the best hit@1 value acquired by the same method is 52.07%, which shows the great space of rQA performance improvement for future research. Our future works include combining our strategies with other models, using alternative encoder and applying our methods to more NLP tasks.

Acknowledgments. We would like to thank the anonymous reviewers and Li Gui and Fiona Liu for their helpful feedback. This work is supported by Natural Science Foundation of China (Grant No.61872113), and the joint project foundation of Tencent Group.

References

1. Aghaebrahimian, A.: Quora question answer dataset. In: Text, Speech, and Dialogue - 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings. pp. 66–73 (2017)
2. Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings. In: International Conference on Learning Representations, ICLR, 2017 (2017)
3. Bogdanova, D., dos Santos, C.N., Barbosa, L., Zadrozny, B.: Detecting semantically equivalent questions in online user forums. In: Proceedings of the 19th Conference on Computational Natural Language Learning, CoNLL 2015, Beijing, China, July 30-31, 2015. pp. 123–131 (2015)
4. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015. pp. 632–642 (2015)
5. Britz, D., Goldie, A., Luong, M.T., Le, Q.: Massive exploration of neural machine translation architectures. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1442–1451 (2017)
6. Chen, D., Fisch, A., Weston, J., Bordes, A.: Reading wikipedia to answer open-domain questions. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. pp. 1870–1879 (2017)
7. Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H., Inkpen, D.: Enhanced LSTM for natural language inference. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. pp. 1657–1668 (2017)
8. Collobert, R., Weston, J.: A unified architecture for natural language processing: deep neural networks with multitask learning. In: Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008. pp. 160–167 (2008)
9. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017. pp. 670–680 (2017)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186 (2019)
11. Gong, Y., Bowman, S.R.: Ruminating reader: Reasoning with gated multi-hop attention. In: Proceedings of the Workshop on Machine Reading for Question Answering@ACL 2018, Melbourne, Australia, July 19, 2018. pp. 1–11 (2018)
12. Gong, Y., Luo, H., Zhang, J.: Natural language inference over interaction space. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings (2018), https://openreview.net/forum?id=r1dHXnH6-
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 770–778 (2016)
14. Hu, B., Lu, Z., Li, H., Chen, Q.: Convolutional neural network architectures for matching natural language sentences. In: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. pp. 2042–2050 (2014)

15. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 2261–2269 (2017)
16. Inkpen, D., Zhu, X., Ling, Z., Chen, Q., Wei, S.: Neural natural language inference models enhanced with external knowledge. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. pp. 2406–2417 (2018)
17. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances. In: Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015. pp. 957–966 (2015)
18. Liu, X., He, P., Chen, W., Gao, J.: Multi-task deep neural networks for natural language understanding. CoRR **abs/1901.11504** (2019)
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
20. Parikh, A.P., Täckström, O., Das, D., Uszkoreit, J.: A decomposable attention model for natural language inference. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. pp. 2249–2255 (2016)
21. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. pp. 1532–1543 (2014)
22. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
23. Tan, C., Wei, F., Wang, W., Lv, W., Zhou, M.: Multiway attention networks for modeling sentence pairs. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden. pp. 4411–4417 (2018)
24. Tay, Y., Luu, A.T., Hui, S.C.: Compare, compress and propagate: Enhancing neural architectures with alignment factorization for natural language inference. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. pp. 1565–1575 (2018)
25. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018. pp. 353–355 (2018)
26. Wang, Z., Hamza, W., Florian, R.: Bilateral multi-perspective matching for natural language sentences. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017. pp. 4144–4150 (2017)
27. Yin, W., Schütze, H., Xiang, B., Zhou, B.: ABCNN: attention-based convolutional neural network for modeling sentence pairs. TACL **4**, 259–272 (2016)
28. Yu, J., Qiu, M., Jiang, J., Huang, J., Song, S., Chu, W., Chen, H.: Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018. pp. 682–690 (2018)
29. Zhang, X., Sun, X., Wang, H.: Duplicate question identification by integrating framenet with neural networks. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. pp. 6061–6068 (2018)