DIM Reader: Dual Interaction Model for Machine Comprehension

Liu Zhuang, Huang Degen, Huang Kaiyu, Zhang Jing

School of Computer Science and Technology, Dalian University of Technology
huangdg@dlut.edu.cn, huangkaiyucs@foxmail.com,
{zhuangliu,zhangjingqf}@mail.dlut.edu.cn

Abstract. Enabling a computer to understand a document so that it can answer comprehension questions is a central, yet unsolved goal of Natural Language Processing, so reading comprehension of text is an important problem in NLP research. In this paper, we propose a novel dual interaction model (called DIM Reader)[1], which constructs dual iterative alternating attention mechanism over multiple hops. The proposed DIM Reader continually refines its view of the query and document while aggregating the information required to answer a query, aiming to compute the attentions not only for the document but also the query side, which will benefit from the mutual information. DIM Reader makes use of multiple turns to effectively exploit and perform deeper inference among queries, documents. We conduct extensive experiments on CNN/DailyMail News datasets, and our model achieves the best results on both machine comprehension datasets among almost published results.

Keywords: machine comprehension; bi-directional attention; dual interaction model; Cloze-style;

1 Introduction

Reading comprehension is the ability to read text, process it, and understand its meaning. How to endow computers with this capacity has been an elusive challenge and a long-standing goal of Artificial Intelligence, so machine comprehension of text is one of the ultimate goals of natural language processing. While the ability of a machine to understand text can be assessed in many different ways, in recent years, several benchmark datasets have been created to focus on Cloze-style questions as a way to evaluate machine reading comprehension[3, 6, 8, 11, 12, 16, 26]. Cloze-style queries are representative problems in machine reading comprehension. To teach the machine to do Cloze-style reading comprehensions, large-scale training data is necessary for learning relationships between the given document and query. Here we mainly focus on the related work in cloze-style datasets[11, 12]. In the past few years, several large-scale datasets of Cloze-style

[1] Our code is available at https://github.com/dlt/mrc-dim

questions over a context document have been introduced which allow the training of supervised machine learning systems[6, 11, 12]. Two large-scale machine comprehension datasets have been released: the CNN/DailyMail corpus, consisting of news articles from those outlets[11], and the Childrens Book Test (CBTest), consisting of short excerpts from books available through Project Gutenberg[12]. The size of these datasets makes them amenable to data-intensive deep learning techniques. Both corpora use Cloze-style questions[28] , which are formulated by replacing a word or phrase in a given sentence with a placeholder token. The task is then to find the answer that "fills in the blank".

Over the past year, the tasks of machine comprehension have gained significant popularity within the natural language processing, and we have seen much progress that is utilizing neural network approach to solve Cloze-style questions. In tandem with these corpora (CNN/DailyMail and CBTest), a host of neural machine comprehension models has been developed[3, 6, 11, 12, 16]. All previous works are focusing on automatically generating large-scale training data for neural network training, which demonstrate its importance. The availability of relatively large training datasets has made it more feasible to train and estimate rather complex models in an end-to-end fashion for these problems, in which a whole model is fit directly with given question-answer tuples and the resulting model has shown to be rather effective.

In this paper, we propose the Dual Interaction Model (DIM), a novel attention-based neural network model called DIM Reader for machine comprehension tasks, designed to study machine comprehension of text, which constructs dual iterative alternating attention mechanism over multiple hops. The model first constructs the representations of the context paragraph at different levels of granularity. DIM Reader includes word-level and character-level embeddings, and uses bi-directional attention for query-aware context representation. Then, DIM Reader's core module, dual inference attention module, begins by deploying a dual multi-hop inference mechanism that alternates between attending query encodings and document encodings, to uncover the inferential links that exist between the document, the missing query word and the query. The results of the alternating attention is gated and fed back into the inference LSTM. After a number of steps, the weights of the document attention are used to estimate the probability of the answer.

To sum up, our contributions can be summarized as follows:

- We propose a novel end-to-end neural network models for machine reading comprehension, which combine a dual inference attention mechanism to handle the Cloze-style reading comprehension task.
- Also, we have achieved the state-of-the-art performance in public reading comprehension datasets such as CNN/DailyMail, and our experimental evaluations also show that our model performs well on machine comprehension datasets.
- Our further analyses with the models reveal some useful insights for further improving the method.

2 Problem notation, datasets

2.1 Definition and notation

The task of the DIM Reader is to answer a Cloze-style question by reading and comprehending a supporting passage of text. Cloze-style questions are formulated by replacing a word or phrase in a given sentence with a placeholder token. The task is then to find the answer that "fills in the blank". The Cloze-style reading comprehension problem[28] aims to comprehend the given context or document, and then answer the questions based on the nature of the document, while the answer is a single word or phrase in the document. Thus, the Cloze-style reading comprehension can be described as a triple:

$$(\mathcal{Q},\ \mathcal{D},\ \mathcal{A})$$

where $\mathcal{Q}$ is the query (represented as a sequence of words), $\mathcal{D}$ is the document, $\mathcal{A}$ is the set of possible answers to the query.

2.2 Reading Comprehension Datasets

Recent advance on reading comprehension has been closely associated with the availability of various datasets. In the past few years, several institutes have released large-scale machine reading comprehension datasets of Cloze-style questions, and these have greatly accelerated the research of machine reading comprehension.

We begin with a brief introduction of the existing Cloze-style reading comprehension datasets. Richardson *et al.*[24] released the MCTest data consisting of 500 short, fictional open-domain stories and 2000 questions. MCTest is challenging because it is both complicated and small. But its size limits the number of parameters that can be trained, and prevents learning any complex language modeling simultaneously with the capacity to answer questions. Typically, there are two main genres of the English Cloze-style datasets publicly available, CNN/Daily Mail[2][11] and Children's Book Test (CBTest)[3][12], which all stem from the English reading materials. The CNN/DailyMail corpus[11], consisting of news articles from those outlets for close style machine comprehension, in which only entities are removed and tested for comprehension, and the Childrens Book Test (CBTest)[12], consisting of short excerpts from books available through Project Gutenberg, which leverages named entities, common nouns, verbs, and prepositions to test reading comprehension. The size of these datasets makes them amenable to data-intensive deep learning techniques. Table 1 provides some statistics on the two English datasets: CNN/Daily Mail and Children's Book Test (CBTest).

[2] CNN and Daily Mail datasets are available at http://cs.nyu.edu/%7ekcho/DMQA

[3] CBTest datasets is available at http://www.thespermwhale.com/jaseweston/babi/CBTest.tgz

Table 1. Data statistics of the CNN datasets and Children's Book Test datasets (CBTest). CBTest CN stands for CBTest Common Nouns and CBTest NE stands for CBTest Named Entites. CBTest had a fixed number of 10 options for answering each question. Statistics provided with the CBTest data set.

	CNN			**CBTest CN**			**CBTest NE**		
	Train	**Valid**	**Test**	**Train**	**Valid**	**Test**	**Train**	**Valid**	**Test**
# queries	380,298	3,924	3,198	879,450	2,000	2,500	108,719	2,000	2,500
Max# options	527	187	396	10	10	10	10	10	10
Avg# options	26.4	26.5	24.5	10	10	10	10	10	10
Avg# tokens	762	763	716	470	448	461	433	412	424
Vocab. size		118,497			53,185			53,063	

3 Proposed Approach

In this section, we will introduce our Dual Interaction Model (DIM Reader) for Cloze-style reading comprehension task. The proposed DIM Reader is shown in Figure 1.

In encoder layer, we first convert the words to their respective word-level embeddings and character-level embeddings. The word embedding is a fixed vector for each individual word, which is pre-trained with word2vec[21]. We also embeds each word by encoding their character sequences with a convolutional neural network followed by max-pooling over time[17], resulting inacharacter-level embedding.

In DIM's core layer, dual inference attention module, our model is primarily motivated by Chen *et al.*[3], Kadlec *et al.*[16] and Sukhbaatar *et al.*[27], which aim to directly estimate the answer from the document, instead of making a prediction over the full vocabularies. But we have noticed that by just concatenating the final representations of the query RNN states are not enough for representing the whole information of query. So we propose to utilize the repeated, tight integration between query attention and document attention, which allows the model to explore dynamically which parts of the query are most important to predict the answer, and then to focus on the parts of the document that are most salient to the currently attended query components.

The top layer, the answer prediction layer, aims to predict the probability of the answer given the document and the query. We aggregate the probabilities for tokens which appear multiple times in a document before selecting the maximum as the predicted answer.

3.1 Document and the Query Encoder Layer

In machine reading comprehension task, the document and query are both word sequences. The goal of this layer is to represent each word in the document and query with a vector. We construct the d-dimensional vector with two components: word embeddings and character embeddings. This layer first maps each word to its corresponding word embedding $\{w_t^{\mathcal{D}}\}_{t=1}^{m}$ and $\{w_t^{\mathcal{Q}}\}_{t=1}^{n}$ (Consider a

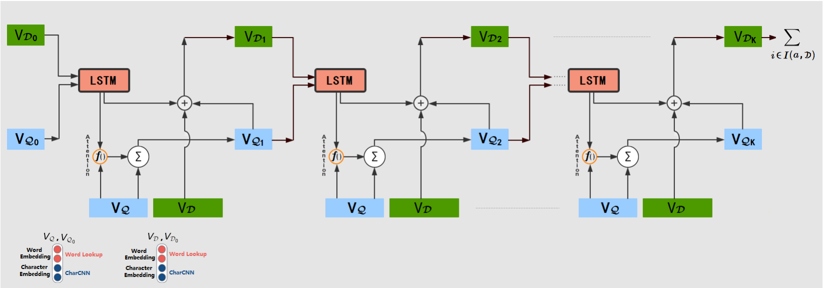

Fig. 1. Architecture of the proposed Dual Interaction Model (DIM Reader).

document $\mathcal{D} = \{x_t^{\mathcal{D}}\}_{t=1}^m$ and a query $\mathcal{Q} = \{x_t^{\mathcal{Q}}\}_{t=1}^n$, where m and n denote the length of document and query respectively), which is typically done by using pre-trained word vectors, which is pre-trained with word2vec[21], to obtain the fixed word embedding of each word. At a more low-level granularity, we also embeds each word by encoding their character sequences with a convolutional neural network followed by max-pooling over time[17]. Characters are embedded into vectors, which can be considered as 1D inputs to the convolutional neural network, and whose size is the input channel size of the convolutional neural network. The outputs of the convolutional neural network are max-pooled over the entire width to obtain a fixed-size vector for each word ($\{x_t^{\mathcal{D}}\}_{t=1}^m$ and $\{x_t^{\mathcal{Q}}\}_{t=1}^n$), resulting in a character-level embedding $\{c_t^{\mathcal{D}}\}_{t=1}^m$ and $\{c_t^{\mathcal{Q}}\}_{t=1}^n$. Each word embedding u_t ($\{u_t^{\mathcal{D}}\}_{t=1}^m$ and $\{u_t^{\mathcal{Q}}\}_{t=1}^n$) is then represented as the concatenation of word-level embedding and character-level embedding, denoted as $u_t = [w_t, c_t] \in R^d$, where d is the total dimensionality of word-level embedding and character-level embedding.

3.2 Dual Inference Interaction Layer

As shown in the previous section, we generates vector representations for the document encodings and query encodings separately. This layer aims to uncover a dual iterative inference chain that starts at the document and the query, and leads to the answer. Figure 1 illustrates dual inference attention module.

Query Attention Module. We use a bilinear attention to compute the importance of each query term (such as as query vector $V_{\mathcal{Q}_0}$) in the current time step t. This bilinear term has been successfully used in [20]. It performs an attentive read on the query encodings, resulting in a query glimpse $\mathbf{q}_t$, and makes the model combine the information in the query with the new information digested from previous iterations. Here $\mathbf{q}_i$ are the query encodings, and we formulate a query glimpse $u_t^{\mathbf{q}}$ at time step t by:

$$u_t^{\mathbf{q}} = \sum_{i=1}^{|\mathcal{Q}|} softmax\, \mathbf{q}_i^T \mathbf{M}_q \mathbf{s}_{t-1} \mathbf{q}_i \tag{1}$$

Document Attention Module. Our method extends the Gated-attention Readers[8] and Attention Sum Reader[16], and performs multiple time-step over the input. The dual iterative alternating attention continues by aggregating the document given the current query glimpse $u_t^{\mathbf{q}}$. The document attention weights are computed based on both the previous search state t-1 and the currently selected query glimpse $u_t^{\mathbf{q}}$:

$$\mathbf{w}_i = \underset{i=1,\ldots,|\mathcal{D}|}{softmax}\, \mathbf{d}_i^T \mathbf{M}_d [\mathbf{s}_{t-1}, u_t^{\mathbf{q}}] \tag{2}$$

$$\mathbf{d}_t = \sum_{i=1}^{|\mathcal{D}|} \mathbf{w}_i \mathbf{d}_i \tag{3}$$

where $\mathbf{w}_i$ are the attention weights for each word in the document, $\mathbf{d}_i$ are the document encodings, and the document attention is conditioned on $\mathbf{s}_{t-1}$, so it reads documents and enriching the query in an iterative fashion, and makes the model perform transitive reasoning on the document side.

Inference Attention Module. The inference is modeled by an additional LSTM[13]. The recurrent network iteratively performs an alternating search step to probe information that may be useful to predict the answer. The module performs an attentive read on the query encodings, resulting in a query glimpse $u_t^{\mathbf{q}}$ at each time step, then gives the current query glimpse $u_t^{\mathbf{q}}$, it extracts a conditional document glimpse $\mathbf{d}_t$, representing the parts of the document that are relevant to the current query glimpse. It produces a new query glimpse and document glimpse in each iteration and utilizes them alternatively in the next iteration, then combines the information in the query with the new information digested from previous iterations. Both attentive reads are conditioned on the previous hidden state of the inference LSTM $\mathbf{s}_{t-1}$, summarizing the information that has been gathered from the query and the document up to time t, making it easier to determine the degree of matching between them. The inference LSTM uses both glimpses to update its recurrent state and thus decides which information needs to be gathered to complete the inference process. It explores the idea of using both attention-sum to aggregate candidate attention scores and multiple turns to attain a better reasoning capability.

3.3 Answer Prediction Layer

The top layer, the answer prediction layer, aims to predict the probability of the answer given the document and the query. After a maximum number of hops K, the document attention weights obtained in the last search step $d_{i,K}$ are used to predict the probability of the answer. We aggregate the probabilities for tokens which appear multiple times in a document before selecting the maximum as the predicted answer:

$$P(a|\mathcal{Q},\ \mathcal{D}) = \sum_{i \in I(a,\mathcal{D})} d_{i,K} \tag{4}$$

where $I(a, \mathcal{D})$ is the set of positions where token a appears in the document $\mathcal{D}$, the model is trained by minimizing the cross-entropy loss using the softmax-weights of candidate scores as the predicted probabilities.

4 Experiments

4.1 Experimental Setups

The general settings of our neural network model are detailed below.

- Embedding Layer: The embedding weights are randomly initialized with the uniformed distribution in the interval [-0.05, 0.05].
- Hidden Layer: We initialized the LSTM units with random orthogonal matrices[25].
- Vocabulary: For training efficiency and generalization, we truncate the full vocabulary (about 200K) and set a shortlist of 100K. During training we randomly shuffled all examples in each epoch. To speedup training, we always pre-fetched 10 batches worth of examples and sorted them according to the length of the document. This way each batch contained documents of roughly the same length.
- Optimization: In order to minimize the hyper-parameter tuning, we used stochastic gradient descent with the ADAM update rule[18] and learning rate of 0.001 or 0.0005, with an initial learning rate of 0.001.

Due to the time limitations, we only tested a few combinations of hyper-parameters, while we expect to have a full parameter tuning in the future. The results are reported with the best model, which is selected by the performance of validation set. Our model is implemented with Tensorflow[1] and Keras[4], and all models are trained on GTX Titan x GPU.

4.2 Results

We compared the proposed model with several baselines as summarized below. To verify the effectiveness of our proposed model, we tested our model on public english reading comprehension datasets. Our evaluation is carried out on CNN news datasets[11] and CBTest NE/CN datasets[12], and the statistics of these datasets are given in Table 2. As we can see that, the proposed DIM Reader achieves the state-of-the-art results in all types of test set, among almost published results.

In CNN news datasets, our model is on par with the AoA Reader[5], with 0.1% improvements in validation set. But we failed to outperform EpiReader[29]. In CBTest NE, though there is a drop in the validation set with 0.7% declines, there is a boost in the test set with an absolute improvements over other models, which suggest our model is effective. In CBTest CN dataset, our model outperforms all the state-of-the-art systems, where a 0.3% and 0.6% absolute accuracy improvements over the most recent state-of-the-art AoA Reader[5] in

Table 2. Results on the CNN news, CBTest NE (named entity) and CN (common noun) datasets. The result that performs best is depicted in bold face.

	CNN News		**CBTest NE**		**CBTest CN**	
	Valid	Test	Valid	Test	Valid	Test
Deep LSTM Reader (Hermann *et al.*[11])	55.0	57.0	-	-	-	-
Attentive Reader (Hermann *et al.*[11])	61.6	63.0	-	-	-	-
Impatient Reader (Hermann *et al.*[11])	61.8	63.8	-	-	-	-
LSTMs (context+query) (Hill *et al.*[12])	-	-	51.2	41.8	62.6	56.0
MemNN (window + self-sup.) (Hill *et al.*[12])	63.4	66.8	70.4	66.6	64.2	63.0
AS Reader (Kadlec *et al.*[16])	68.6	69.5	73.8	68.6	68.8	63.4
Stanford AR (Chen *et al.*[3])	72.4	72.4	-	-	-	-
Iterative Attention(Sordoni *et al.*[26])	72.6	73.3	75.2	68.6	72.1	69.2
GA Reader (Dhingra *et al.*[8])	73.0	73.8	74.9	69.0	69.0	63.9
EpiReader (Trischler *et al.*[29])	**73.4**	74.0	75.3	69.7	71.5	67.4
CAS Reader (avg mode)(Cui *et al.*[6])	68.2	70.0	74.2	69.2	68.2	65.7
AoA Reader (Cui *et al.*[5])	73.1	**74.4**	**77.8**	72.0	72.2	69.4
DIM Reader (our)	73.2	**74.4**	77.1	**72.2**	**72.5**	**70.0**

the validation and test set respectively. We have also noticed that, our model has an absolute improvement over EpiReader[29]. When compared with AoA Reader[5], GA Reader[8], EpiReader[29], our model shows a similar result, with improvements on validation and test set, experimental results demonstrate that our model is more general and powerful than previous works. This demonstrates that our model is powerful enough to compete with english reading comprehension, to tackle the Cloze-style reading comprehension task.

So far, we have good results in machine reading comprehension, all higher than most baselines above, verifying that dual interaction model is useful, suggesting that our DIM Reader performed better on relatively difficult reasoning questions.

5 Related Work

Neural attention models have been applied recently to machine learning and natural language processing problems. Cloze-style reading comprehension tasks have been widely investigated in recent studies. We will take a brief revisit to the previous works.

Hermann *et al.*[11] have proposed a methodology for obtaining large quantities of ($\mathcal{Q}$, $\mathcal{D}$, $\mathcal{A}$) triples through news articles and its summary. Along with the release of Cloze-style reading comprehension dataset, they also proposed an attention-based neural network to tackle the issues above. Experimental results showed that the proposed neural network is effective than traditional baselines. Hill *et al.*[12] released another dataset, which stems from the children's books. Different from Hermann *et al.*[11]'s work, the document and query are all generated from the raw story without any summary, which is much more general than previous work. To handle the reading comprehension task, they proposed a window-based memory network, and self-supervision heuristics is also applied to learn hard-attention. Kadlec *et al.*[16] proposed a simple model that directly pick the answer from the document, which is motivated by the Pointer Network[30]. A restriction of this model is that, the answer should be a single word and appear in the document. Results on various public datasets showed that the proposed model is effective than previous works. Liu *et al.*[19] proposed to exploit these reading comprehension models into specific task. They first applied the reading comprehension model into Chinese zero pronoun resolution task with automatically generated large-scale pseudo training data. Trischler *et al.*[29] adopted a re-ranking strategy into the neural networks and used a joint-training method to optimize the neural network. Sordoni *et al.*[26] have proposed an iterative alternating attention mechanism and gating strategies to accumulatively optimize the attention after several hops, where the number of hops is defined heuristically. Seo *et al.*[22] proposed the BiDAF model, which computes both the context-to-query attention and the query-to-context attention by using second-order attention, and Cui *et al.*[5] computed a query-level average attention based on the alignment matrix, which is then used to further compute a weighted sum of context-level attention. Document Reader[7] and Dynamic Coattention Networks[2] utilized a multi-hop pointing decoder to indicate the answer span iteratively, and Answer Pointer[31] and Ruminating Reader[10] focused on the query-aware context representation and use query-independent pointer vector to select the answer boundary. Weissenborn *et al.*[9] and Zhang *et al.*[15] used one-hop reasoning models to emphasis relevant parts between the context and the query, and the architecture of these models is quite shallow, usually containing only one interaction layer.

6 Conclusions

In this paper we presented the novel dual interaction model (called DIM Reader), and showed it offered improved performance for machine comprehension tasks. Among the large, public english machine comprehension datasets, our model could give significant improvements over various state-of-the-art baselines.

As future work, we need to consider how we can utilize new corpora (such as SQuAD[23] and TriviaQA[14]) to solve more complex machine reading comprehension tasks, and we are going to investigate hybrid reading comprehension models to tackle the problems that rely on comprehensive induction of several sentences. We also plan to augment our framework with a more powerful model for question answering.

Acknowledgments

We would like to thank the reviewers for their helpful comments and suggestions to improve the quality of the paper. This research is supported by National Natural Science Foundation of China (No.61672127).

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016)
2. Caiming Xiong, V.Z., Socher, R.: Dynamic coattention networks for question answering. In Proceedings of ICLR (2017)
3. Chen, D., Bolton, J., Manning, C.D.: A thorough examination of the cnn/daily mail reading comprehension task. arXiv preprint arXiv:1606.02858 (2016)
4. Chollet, F.: Keras (2015)
5. Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., Hu, G.: Attention-over-attention neural networks for reading comprehension. arXiv preprint arXiv:1607.04423 (2016)
6. Cui, Y., Liu, T., Chen, Z., Wang, S., Hu, G.: Consensus attention-based neural networks for chinese reading comprehension. arXiv preprint arXiv:1607.02250 (2016)
7. Danqi Chen, Adam Fisch, J.W., Bordes, A.: Reading wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051 (2017)
8. Dhingra, B., Liu, H., Cohen, W.W., Salakhutdinov, R.: Gated-attention readers for text comprehension. arXiv preprint arXiv:1606.01549 (2016)
9. Dirk Weissenborn, G.W., Seiffe, L.: Fastqa: A simple and efficient neural architecture for question answering. arXiv preprint arXiv:1703.04816 (2017)
10. Gong, Y., Bowman, S.R.: Ruminating reader: Reasoning with gated multi-hop attention. arXiv preprint arXiv:1704.07415 (2017)
11. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. In: Advances in Neural Information Processing Systems. pp. 1693–1701 (2015)
12. Hill, F., Bordes, A., Chopra, S., Weston, J.: The goldilocks principle: Reading children's books with explicit memory representations. arXiv preprint arXiv:1511.02301 (2015)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
14. Joshi, M., Choi, E., Weld, D.S., Zettlemoyer, L.: Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. arXiv preprint arXiv:1705.03551 (2017)
15. Junbei Zhang, Xiaodan Zhu, Q.C.L.D.S.W., Jiang, H.: Exploring question understanding and adaptation in neural-network-based question answering. arXiv preprint arXiv:1703.04617 (2017)
16. Kadlec, R., Schmid, M., Bajgar, O., Kleindienst, J.: Text understanding with the attention sum reader network. arXiv preprint arXiv:1603.01547 (2016)
17. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
18. Kinga, D., Adam, J.B.: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
19. Liu, T., Cui, Y., Yin, Q., Wang, S., Zhang, W., Hu, G.: Generating and exploiting large-scale pseudo training data for zero pronoun resolution. arXiv preprint arXiv:1606.01603 (2016)
20. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
21. 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)

22. Minjoon Seo, Aniruddha Kembhavi, A.F., Hajishirzi, H.: Bidirectional attention flow for machine comprehension. In Proceedings of ICLR (2017)
23. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
24. Richardson, M., Burges, C.J., Renshaw, E.: Mctest: A challenge dataset for the open-domain machine comprehension of text. In: EMNLP. vol. 3, p. 4 (2013)
25. Saxe, A.M., McClelland, J.L., Ganguli, S.: Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv preprint arXiv:1312.6120 (2013)
26. Sordoni, A., Bachman, P., Trischler, A., Bengio, Y.: Iterative alternating neural attention for machine reading. arXiv preprint arXiv:1606.02245 (2016)
27. Sukhbaatar, S., Weston, J., Fergus, R., et al.: End-to-end memory networks. In: Advances in neural information processing systems. pp. 2440–2448 (2015)
28. Taylor, W.L.: cloze procedure: a new tool for measuring readability. Journalism Bulletin 30(4), 415–433 (1953)
29. Trischler, A., Ye, Z., Yuan, X., Suleman, K.: Natural language comprehension with the epireader. arXiv preprint arXiv:1606.02270 (2016)
30. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. In: Advances in Neural Information Processing Systems. pp. 2692–2700 (2015)
31. Wang, S., Jiang, J.: Machine comprehension using match-lstm and answer pointer. In Proceedings of ICLR (2017)