ISSN 1870-4069

Reasoning with Self-attention and Inference Model for Machine Comprehension

Zhuang Liu, Degen Huang,
Kaiyu Huang, Jing Zhang

Dalian University of Technology,
School of Computer Science and Technology,
China

{zhuangliu, zhangjingqf, huangkaiyu}@mail.dlut.edu.cn
huangdg@dlut.edu.cn

Resumen. Enabling a computer to understand a document so that it can answer comprehension questions is a central, yet unsolved goal of Natural Language Processing (NLP), so reading comprehension of text is an important problem in NLP. Recently, machine reading comprehension has embraced a booming in NLP research. In this paper, we introduce a novel iterative inference neural network based on a matrix sentence embedding with a self-attention mechanism. The proposed approach 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. Experimental results show that our model has achieved significant state-of-the-art performance in public English datasets, such as CNN and Children's Book Test datasets. Furthermore, the proposed model also outperforms state-of-the-art systems by a large margin in Chinese datasets, including People Daily and Children's Fairy Tale datasets, which are recently released and the first Chinese reading comprehension datasets.

Palabras clave: Machine comprehension, matrix sentence, hybrid models of reading comprehension.

1 Introduction

Reading comprehension[1] 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. A recent trend to measure progress towards machine reading is to test a system's ability to answering questions over the text it has to comprehend.

Towards this end, several large-scale datasets of Cloze-style questions over a context document have been introduced recently which allow the training of supervised machine learning systems [4, 6, 7, 23]. Cloze-style queries are representative problems in reading comprehension. Over the past few months, we have seen much progress that is utilizing neural network approach to solve Cloze-style questions.

[1] en.wikipedia.org/wiki/Reading_comprehension

In the past year, 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. Some large-scale reading comprehensions datasets have been released: the CNN/Daily Mail corpus, consisting of news articles from those outlets [6], and the Children's Book Test (CBTest), consisting of short excerpts from books available through Project Gutenberg [7].

Recently, Cui et al. [4] has released the first machine Chinese reading comprehension datasets, including a human-made out-of-domain test set for future research. All previous works are focusing on automatically generating large-scale training data for neural network training, which demonstrate its importance. Furthermore, the more complicated problems the more data is needed to learn comprehensive knowledge from it, such as reasoning over multiple sentences etc.

In this paper, we propose a novel iterative inference neural network model, designed to study machine comprehension of text, which constructs iterative inference mechanism. The Self-Attentive Encoder Model first uses a self-attention mechanism for representing sentences.

Then, the core module, Iterative Inference Model, begins by deploying an iterative inference mechanism that alternates between attending query encodings and document encodings, to uncover the inferential links that exist between the document and the query. The results of the alternating attention are 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 model for machine reading comprehension, which combines self-attentive sentence embedding and an iterative inference mechanism to handle the Cloze-style reading comprehension task.
- Also, we have achieved the state-of-the-art performance in public reading comprehension datasets, including English datasets and Chinese datasets.
- Our further analyses with the models reveal some useful insights for further improving the method.

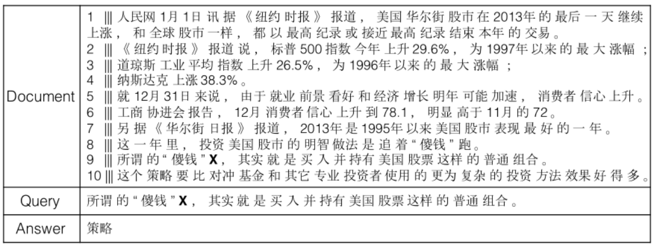
2 Problem Notation, Datasets

2.1 Definition and Notation

The task of the proposed model is to answer a Cloze-style question by reading and comprehending a supporting passage of text. The Cloze-style reading comprehension problem (proposed by Taylor[20]) 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 in the document. Thus, the Cloze-style reading comprehension can be described as a triple:

$$(Q, D, A), \tag{1}$$

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



<table>
<tr><td>Document</td><td>1 ||| 人民网 1月 1日 讯 据 《 纽约 时报 》 报道 ， 美国 华尔街 股市 在 2013年 的 最后 一 天 继续 上涨 ， 和 全球 股市 一样 ， 都 以 最高 纪录 或 接近 最高 纪录 结束 本年 的 交易 。
2 ||| 《 纽约 时报 》 报道 说 ， 标普 500 指数 今年 上升 29.6% ， 为 1997年 以来 的 最 大 涨幅 ；
3 ||| 道琼斯 工业 平均 指数 上升 26.5% ， 为 1996年 以来 的 最 大 涨幅 ；
4 ||| 纳斯达克 上涨 38.3% 。
5 ||| 就 12月 31日 来说 ， 由于 就业 前景 看好 和 经济 增长 明年 可能 加速 ， 消费者 信心 上升 。
6 ||| 工商 协进会 报告 ， 12月 消费者 信心 上升 到 78.1 ， 明显 高于 11月 的 72 。
7 ||| 另 据 《 华尔街 日报 》 报道 ， 2013年 是 1995年 以来 美国 股市 表现 最 好 的 一 年 。
8 ||| 这 一 年 里 ， 投资 美国 股市 的 明智 做法 是 追 着 “ 傻钱 ” 跑 。
9 ||| 所谓 的 “ 傻钱 ” <b>X</b> ， 其实 就 是 买 入 并 持有 美国 股票 这样 的 普通 组合 。
10 ||| 这个 策略 要 比 对冲 基金 和 其它 专业 投资者 使用 的 更为 复杂 的 投资 方法 效果 好 得 多 。</td></tr>
<tr><td>Query</td><td>所谓 的 “ 傻钱 ” <b>X</b> ， 其实 就 是 买 入 并 持有 美国 股票 这样 的 普通 组合 。</td></tr>
<tr><td>Answer</td><td>策略</td></tr>
</table>

Fig. 1. Example training sample in people daily datasets. The ”**X**” represents the missing word. In this example, the document consists of 10 sentences, and the 9th sentence is chosen as the query.

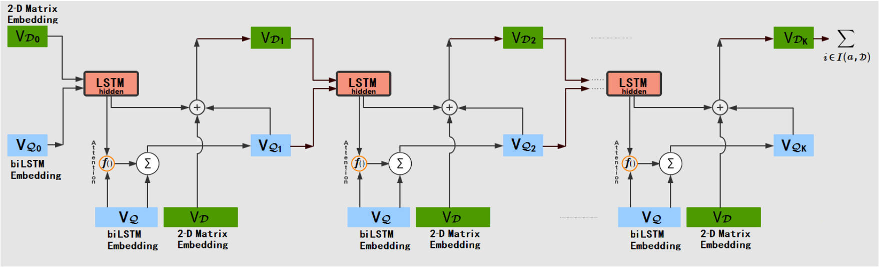


Fig. 2. Architecture of the proposed iterative inference neural networks.

2.2 Reading Comprehension Datasets

Several institutes have released the Cloze-style reading comprehension data, 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, two English datasets and the first Chinese reading comprehension datasets recently released.

Typically, there are two main genres of the English Cloze-style datasets publicly available, CNN/Daily Mail [6] and Children's Book Test (CBTest) [7], which all stem from the English reading materials. Also, there are the first Chinese Cloze-style reading comprehension datasets, People Daily[4] and Children's Fairy Tale (CFT) [4], which are roughly collected 60K news articles from the People Daily website[2] and the Children's Fairy Tale by Cui et al. [4]. Figure 1 shows an example of People Daily datasets[3,4].

Table 1 provides some statistics on the two English datasets: CNN/Daily Mail and Children's Book Test (CBTest). The statistics of People Daily datasets as well as Children's Fairy Tale datasets[5] are listed in the Table 2.

3 Proposed Approach

In this section, we will introduce our iterative inference neural networks for Cloze-style reading comprehension task. The proposed model is shown in Figure 2.

[2] People's Daily: http://www.people.com.cn

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

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

[5] People Daily and CFT datasets are available at http://hfl.iflytek.com/chinese-rc

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 dataset.

			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	

Table 2. Data statistics of people daily datasets and children's fairy tale datasets (CFT).

	People Daily			**Children's Fairy Tale**	
	Train	**Valid**	**Test**	**Test-auto**	**Test-human**
# queries	870,710	3,000	3,000	1,646	1,953
Max# tokens in docs	618	536	634	318	414
Max# tokens in query	502	153	265	83	92
Avg# tokens in docs	379	425	410	122	153
Avg# tokens in query	38	38	41	20	20
Vocabulary size		248,160			NA

In encoder module, we opt for a new model for extracting an interpretable sentence embedding by introducing self-attention [11]. Instead of using a vector, it uses a 2-D matrix to represent the embedding, with each row of the matrix attending on a different part of the sentence. So we propose to apply the matrix sentence embedding with a self-attention mechanism for the document representation.

In core module, iterative inference module, our model is primarily motivated by Sukhbaatar et al. [19], Kadlec et al. [9] and Chen et al. [1], 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 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.

3.1 Self-Attentive Encoder

Inspired by Lin et al. [11] and Liu et al. [13], we opted for self-Attentive sentence embedding for representing document. The query encodings and document encodings are represented separately as query vector V_Q, document V_D. Let d denote the dimension of word embeddings, and S a document sentence consisting of a sequence of

n words $(w_1, ..., w_n)$ which can be represented by a dense column matrix $\mathbf{W}$, who have the shape $n \times d$. We use a bidirectional LSTM to process the individual sentence, and concatenate each $\overrightarrow{h_t}$ with $\overleftarrow{h_t}$ to obtain a hidden state h_t.

Let the hidden unit number for each unidirectional LSTM be k, and note all the n h_ts as H, which has the shape $n \times 2k$, following Lin et al. [11], we choose a linear combination of the n LSTM hidden vectors in H. Computing the linear combination uses the self-attention mechanism. The attention mechanism takes the whole LSTM hidden states H as input, and outputs a vector of weights $\mathbf{a}$:

$$\mathbf{a} = softmax(\mathbf{w_{s2}}\, tanh(W_{s1}H^T)), \tag{2}$$

where W_{s1} is a weight matrix and $\mathbf{w_{s2}}$ is a vector of parameters. Then we sum up the LSTM hidden states H according to the weight $\mathbf{a}$ to get a vector representation $\mathbf{m}$, then we multiple $\mathbf{m}$ that focus on different parts of the sentence to represent the overall semantics of the sentence.

When we want r different parts to be extracted from the sentence, we extend the $\mathbf{w_{s2}}$ into a matrix, note it as W_{s2}, and the resulting annotation vector $\mathbf{a}$ becomes annotation matrix A:

$$A = softmax(W_{s2}\, tanh(W_{s1}H^T)). \tag{3}$$

We can deem Equation(2) as a 2-layer MLP without bias, whose parameters are $\{W_{s1}, W_{s2}\}$. The embedding vector then becomes an $n \times 2k$ embedding matrix M. We compute the r weighted sums by multiplying the annotation matrix A and LSTM hidden states H, the resulting matrix is the sentence embedding:

$$M = AH. \tag{4}$$

We then use the resulting matrix to represent the embedding, with each row of the matrix attending on a different part of the individual sentences of document $V_{\mathcal{D}}$.

3.2 Iterative Inference Model

This phase aims to uncover a possible inference chain that starts at the query and the document and leads to the answer. Figure 2 illustrates iterative inference module. We use a bilinear term instead of a simple dot product [7, 14, 19] in order to compute the importance of each query term in the current time step *t*. This simple bilinear attention has been successfully used in Luong et al. [15]. We formulate a query glimpse $\mathbf{q}_t$ at time step *t* by:

$$q_{i,t} = \underset{i=1,...,|\mathcal{Q}|}{softmax}\, \tilde{\mathbf{q}}_i^T \mathbf{W}_q \mathbf{s}_{t-1}, \tag{5}$$

$$\mathbf{q}_t = \sum_i q_{i,t}\tilde{\mathbf{q}}_i, \tag{6}$$

where $q_{i,t}$ are the query attention weights and $\tilde{\mathbf{q}}_i$ are the query encodings.

Table 3. 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
Impatient Reader (Hermann $et\ al.$ [6])	61.8	63.8	-	-	-	-
MemNN (window + self-sup.) (Hill $et\ al.$ [7])	63.4	66.8	70.4	66.6	64.2	63.0
AS Reader (Kadlec $et\ al.$ [9])	68.6	69.5	73.8	68.6	68.8	63.4
Stanford AR (Chen $et\ al.$ [1])	72.4	72.4	-	-	-	-
Iterative Attention (Sordoni $et\ al.$ [18])	72.6	73.3	75.2	68.6	72.1	69.2
CAS Reader (avg mode)(Cui $et\ al.$ [4])	68.2	70.0	74.2	69.2	68.2	65.7
GA Reader (Dhingra $et\ al.$ [5])	73.0	73.8	74.9	69.0	69.0	63.9
EpiReader (Trischler $et\ al.$ [21])	**73.4**	74.0	75.3	69.7	71.5	67.4
AoA Reader (Cui $et\ al.$ [3])	73.1	74.4	**77.8**	72.0	72.2	69.4
Our proposed model	73.0	**74.5**	77.1	**72.1**	**72.4**	**69.4**

Table 4. Results on people daily datasets and children's fairy tale (CFT) datasets. The result that performs best is depicted in bold face. CAS Reader (marked with †) are the most recent works.

	People Daily		**Children's Fairy Tale**	
	Valid	Test	Test-auto	Test-human
AS Reader	64.1	67.2	40.9	33.1
CAS Reader (avg mode)	65.2	68.1	1.3	35.0
CAS Reader (sum mode)	64.7	66.8	3.0	34.7
CAS Reader (max mode)	63.3	65.4	8.3	32.0
Our proposed model	**66.6**	**69.8**	**45.0**	**37.0**

Our method extends the Attention Sum Reader [9], and performs multiple hops over the input. The alternating attention continues by probing the document given the current query glimpse $\mathbf{q}_t$.

The document attention weights are computed based on both the previous search state $t-1$ and the currently selected query glimpse $\mathbf{q}_t$:

$$d_{i,t} = \underset{i=1,\ldots,|\mathcal{D}|}{softmax} \tilde{\mathbf{d}}_i^T \mathbf{W}_d[\mathbf{q}_t, \mathbf{s}_{t-1}], \tag{7}$$

$$\mathbf{d}_t = \sum_i d_{i,t} \tilde{\mathbf{d}}_i, \tag{8}$$

where $\tilde{\mathbf{d}}_i$ are the query encodings, $d_{i,t}$ are the attention weights for each word in the document, and the document attention is also conditioned on $\mathbf{s}_{t-1}$, so it makes the model perform transitive reasoning on the document side. This use previously obtained document information to future attended locations, which is particularly important for natural language inference tasks[19].

In iterative inference module, the inference is modeled by an additional LSTM (proposed by Hochreiter and Schmidhuber[8]). The recurrent network iteratively performs an alternating search step to gather information that may be useful to predict the answer.

The module performs an attentive read on the query encodings, resulting in a query glimpse $\mathbf{q}_t$ at each time step, then gives the current query glimpse $\mathbf{q}_t$, it extracts a conditional document glimpse $\mathbf{d}_t$, representing the parts of the document that are relevant to the current query glimpse. 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.

Finally, after a fixed number of time-steps 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:

$$P(a|Q,\ D) = \sum_{i \in I(a,D)} d_{i,K}, \tag{9}$$

where $I(a, D)$ is the set of positions where token a appears in the document D, we then use cross-entropy loss between the predicted probabilities and true answers for training.

4 Experiments

4.1 Experimental Setups

In this section we present our experimental setup for assessing the performance of our iterative inference neural networks. For the self-attentive encoder module, the biLSTM is 300 dimension in each direction, the attention MLP has 180 hidden units instead, and sentence embeddings of document has 20 rows (the r).

We use 300 dimensional GloVe embeddings [16] to represent words, projected down to 200 dimensions, a number determined via hyperparameter tuning. In hidden Layer, we initialized the LSTM units with random orthogonal matrices [17]. In order to minimize the hyper-parameter tuning, we used stochastic gradient descent with the ADAM update rule [10] and learning rate of 0.01 or 0.025, with an initial learning rate of 0.01.

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. The code in this paper has been implemented with Keras framework [2] and our source code also has be released.

4.2 Results

We compared the proposed model with several baselines as summarized below. To verify the effectiveness of our proposed model, we first tested our model on public English datasets. Our evaluation is carried out on CNN news datasets [6] and CBTest NE/CN datasets[7], and the statistics of these datasets are given in Table 3.

The results on Chinese reading comprehension datasets are listed in Table 4, as we can see that, the proposed model significantly outperforms the most recent state-of-the-art CAS Reader in all types of test set, with a maximum.

English Reading Comprehension Datasets. In CNN news datasets, our model is almost on par with the AoA Reader, but we failed to outperform EpiReader. Meantime, In CBTest NE, though there is a drop in the validation set with 0.6% 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 gives modest improvements over the state-of-the-art systems. When compared with AoA Reader, our model shows a similar result, with slight improvements on test set, which demonstrate that our model is more general and powerful than previous works.

Chinese Reading Comprehension Datasets. In People Daily and CFT datasets, our proposed model outperforms all the state-of-the-art systems by a large margin, where a 1.4%, 1.7%, 2.0% and 2.0% absolute accuracy improvements over the most recent state-of-the-art CAS Reader in the validation and test set respectively. This demonstrates that our model is powerful enough to compete with Chinese 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 our proposed model is useful, suggesting that iterative inference neural networks 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. [6] proposed a methodology for obtaining large quantities of $(Q,\ D,\ A)$ 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. [7] released another dataset, which stems from the children's books.

Different from Hermann et al. work [6], 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. [9] proposed a simple model that directly pick the answer from the document, which is motivated by the Pointer Network [22]. 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. [12] 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. [21] adopted a re-ranking strategy into the neural networks and used a joint-training method to optimize the neural network.

6 Conclusions

In this paper we presented the novel iterative inference neural network by introducing a fixed size, matrix sentence embedding with a self-attention mechanism, and showed it offered improved performance for machine comprehension tasks. Among the large, public Chinese and English datasets, our model could give significant improvements over various state-of-the-art baselines, especially for Chinese reading comprehension corpora, on which our model outperformed state-of-the-art systems by a large margin.

As future work, we need to consider how we can utilize these datasets to solve more complex machine reading comprehension tasks (with less annotated data), 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 natural language inference.

Acknowledgments. We thank the reviewers for their instructive feedback. This work is supported by National Natural Science Foundation of China (No.61672127).

References

1. Chen, D., Bolton, J., Manning, C.D.: A thorough examination of the CNN/daily mail reading comprehension task. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. vol. 1, pp. 2358–2367 (2016)
2. Chollet, F.: Building autoencoders in keras. The Keras Blog 14, 1–17 (2016)
3. Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., Hu, G.: Attention-over-attention neural networks for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. vol. 1, pp. 593–602 (2016)
4. Cui, Y., Liu, T., Chen, Z., Wang, S., Hu, G.: Consensus attention-based neural networks for chinese reading comprehension. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (2016)
5. Dhingra, B., Liu, H., Cohen, W.W., Salakhutdinov, R.: Gated-attention readers for text comprehension 1, 1832–1846 (2016)
6. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. vol. 1, pp. 1693–1701 (2015)
7. Hill, F., Bordes, A., Chopra, S., Weston, J.: The goldilocks principle: Reading children's books with explicit memory representations. In: International Conference on Learning Representations (2015)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)

9. Kadlec, R., Schmid, M., Bajgar, O., Kleindienst, J.: Text understanding with the attention sum reader network. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. vol. 1, pp. 908–918 (2016)
10. Kingma, D., Ba, J.: A method for stochastic optimization. In: International Conference on Learning Representations (2015)
11. Lin, Z., Feng, M., Dos-Santos, C.N., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A structured self-attentive sentence embedding. In: International Conference on Learning Representations (2017)
12. Liu, T., Cui, Y., Yin, Q., Wang, S., Zhang, W., Hu, G.: Generating and exploiting large-scale pseudo training data for zero pronoun resolution. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. vol. 1, pp. 102–111 (2016)
13. Liu, Y., Liu, Z., Chua, T.S., Sun, M.: Topical word embeddings. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. vol. 29, pp. 2418–2424 (2015)
14. Liu, Z., Huang, D., Zhang, J., Huang, K.: Research on attention memory networks as a model for learning natural language inference. In: Proceedings of the Workshop on Structured Prediction for NLP. pp. 18–24 (2016)
15. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1412–1421 (2015)
16. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). vol. 14, pp. 1532–1543 (2014)
17. Saxe, A.M., McClelland, J.L., Ganguli, S.: Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In: International Conference on Learning Representations (2013)
18. Sordoni, A., Bachman, P., Trischler, A., Bengio, Y.: Iterative alternating neural attention for machine reading (2016)
19. Sukhbaatar, S., Szlam, A., Weston, J., Fergus, R.: End-to-end memory networks. In: Advances in neural information processing systems. pp. 2440–2448 (2015)
20. Taylor, W.L.: Cloze procedure: A new tool for measuring readability. Journalism Bulletin 30(4), 415–433 (1953)
21. Trischler, A., Ye, Z., Yuan, X., Suleman, K.: Natural language comprehension with the epireader. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 128–137 (2016)
22. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. In: Advances in Neural Information Processing Systems (2015)
23. Zhuang-Liu, Degen-Huang, K.H., Zhang, J.: DIM reader: Dual interaction model for machine comprehension. In: International Symposium on Natural Language Processing Based on Naturally Annotated Big Data China National Conference on Chinese Computational Linguistics. pp. 387–397 (2017)