Consensus Attention-based Neural Networks for Chinese Reading Comprehension

Yiming Cui[†*], **Ting Liu**[‡], **Zhipeng Chen**[†], **Shijin Wang**[†] and **Guoping Hu**[†]
[†]iFLYTEK Research, Beijing, China
[‡]Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China
[†]{ymcui,zpchen,sjwang3,gphu}@iflytek.com
[‡]tliu@ir.hit.edu.cn

Abstract

Reading comprehension has embraced a booming in recent NLP research. Several institutes have released the Cloze-style reading comprehension data, and these have greatly accelerated the research of machine comprehension. In this work, we firstly present Chinese reading comprehension datasets, which consist of People Daily news dataset and Children's Fairy Tale (CFT) dataset. Also, we propose a consensus attention-based neural network architecture to tackle the Cloze-style reading comprehension problem, which aims to induce a consensus attention over every words in the query. Experimental results show that the proposed neural network significantly outperforms the state-of-the-art baselines in several public datasets. Furthermore, we setup a baseline for Chinese reading comprehension task, and hopefully this would speed up the process for future research.

1 Introduction

The ultimate goal of machine intelligence is to read and comprehend human languages. Among various machine comprehension tasks, in recent research, the Cloze-style reading comprehension task has attracted lots of researchers. The Cloze-style reading comprehension problem (Taylor, 1953) 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:

$$\langle \mathcal{D}, \mathcal{Q}, \mathcal{A} \rangle$$

where $\mathcal{D}$ is the document, $\mathcal{Q}$ is the query and $\mathcal{A}$ is the answer to the query.

By adopting attention-based neural network approaches (Bahdanau et al., 2014), the machine is able to learn the relationships between document, query and answer. But, as is known to all, the neural network based approaches need large-scale training data to train a reliable model for predictions. Hermann et al. (2015) published the CNN/Daily Mail news corpus for Cloze-style reading comprehensions, where the content is formed by the news articles and its summarization. Also, Hill et al. (2015) released the Children's Book Test (CBT) corpus for further research, where the training samples are generated through automatic approaches. As we can see that, automatically generating large-scale training data for neural network training is essential for reading comprehension. Furthermore, more difficult problems, such as reasoning or summarization of context, need much more data to learn the higher-level interactions.

Though we have seen many improvements on these public datasets, some researchers suggested that these dataset requires less high-level inference than expected (Chen et al., 2016). Furthermore, the public datasets are all automatically generated, which indicate that the pattern in training and testing phase are nearly the same, and this will be easier for the machine to learn these patterns.

In this paper, we will release Chinese reading comprehension datasets, including People Daily news datasets and Children's Fairy Tale datasets. As a highlight in our datasets, there is a human evaluated

[*]This work was done by the Joint Laboratory of HIT and iFLYTEK (HFL).

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1777–1786, Osaka, Japan, December 11-17 2016.

dataset for testing purpose. And this will be harder for the machine to answer these questions than the automatically generated questions, because the human evaluated dataset is further processed, and may not be accordance with the pattern of automatic questions. More detailed analysis will be given in the following sections. The main contributions of this paper are as follows:

- To our knowledge, this is the first released Chinese reading comprehension datasets and human evaluated test sets, which will benefit the research communities in reading comprehension.
- Also, we propose a refined neural network that aims to utilize full representations of query to deal with the Cloze-style reading comprehension task, and our model outperform various state-of-the-art baseline systems in public datasets.

The rest of the paper will be organized as follows. In Section 2, we will briefly introduce the existing Cloze-style datasets, and describe our Chinese reading comprehension datasets in detail. In Section 3, we will show our refined neural network architecture for Cloze-style reading comprehension. The experimental results on public datasets as well as our Chinese reading comprehension datasets will be given in Section 4. Related work will be described in Section 5, and we make a brief conclusion of our work at the end of this paper.

2 Chinese Reading Comprehension Datasets

We first begin with a brief introduction of the existing Cloze-style reading comprehension datasets, and then introduce our Chinese reading comprehension datasets: People Daily and Children's Fairy Tale.

2.1 Existing Cloze-style Datasets

Typically, there are two main genres of the Cloze-style datasets publicly available, which all stem from the English reading materials.

CNN/Daily Mail.[1] The news articles often come with a short summary of the whole report. In the spirit of this, Hermann et al. (2015) constructed a large dataset with web-crawled CNN and Daily Mail news data. Firstly, they regard the main body of the news article as the *Document*, and the *Query* is formed through the summary of the article, where one entity word is replaced by a placeholder to indicate the missing word. And finally, the replaced entity word will be the *Answer* of the *Query*. Also, they have proposed the *anonymize* the named entity tokens in the data, and re-shuffle the entity tokens for every sample in order to exploit general relationships between anonymized named entities, rather than the common knowledge. But as Chen et al. (2016)'s studies on these datasets showed that the anonymization is less useful than expected.

Children's Book Test. [2] There was also a dataset called the Children's Book Test (CBT) released by Hill et al. (2015), which is built from the children's book story. Different from the previously published CNN/Daily Mail datasets, they formed the *Document* with 20 consecutive sentences in the book, and regard the 21st sentence as the *Query*, where one word is blanked with a placeholder. The missing word are chosen from named entities (NE), common nouns (CN), verbs and prepositions. As the verbs and prepositions are less dependent with the document, most of the studies are focusing on the NE and CN datasets.

2.2 People Daily and Children's Fairy Tale Datasets

In this part, we will introduce our Chinese reading comprehension datasets in detail[3]. Though many solid works on previously described public datasets, there is no studies on Chinese reading comprehension datasets. What makes our datasets different from previous works are listed as below.

- As far as we know, the proposed dataset is the first Chinese Cloze-style reading comprehension datasets, which will add language diversity in the community.

[1]The pre-processed CNN and Daily Mail datasets are available at `http://cs.nyu.edu/~kcho/DMQA/`

[2]The CBT datasets are available at `http://www.thespermwhale.com/jaseweston/babi/CBTest.tgz`

[3]Our datasets are available at `http://hfl.iflytek.com/chinese-rc/`.

Document	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 ” **XXXXX** ， 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 。	1 ||| People Daily (Jan 1). According to report of “New York Times”, the Wall Street stock market continued to rise as the global stock market in the last day of 2013, ending with the highest record or near record of this year. 2 ||| “New York times” reported that the S&P 500 index rose 29.6% this year, which is the largest increase since 1997. 3 ||| Dow Jones industrial average index rose 26.5%, which is the largest increase since 1996. 4 ||| NASDAQ rose 38.3%. 5 ||| In terms of December 31, due to the prospects in employment and possible acceleration of economy next year, there is a rising confidence in consumers. 6 ||| As reported by Business Association report, consumer confidence rose to 78.1 in December, significantly higher than 72 in November. 7 ||| Also as “Wall Street journal” reported that 2013 is the best U.S. stock market since 1995. 8 ||| In this year, to chase the “silly money” is the most wise way to invest in U.S. stock. 9 ||| The so-called “silly money” is that, to buy and hold the common combination of U.S. stock. 10 ||| This strategy is better than other complex investment methods, such as hedge funds and the methods adopted by other professional investors.
Query	所谓 的 “ 傻钱 ” **XXXXX** ， 其实 就 是 买 入 并 持有 美国 股票 这样 的 普通 组合 。	The so-called “silly money” **XXXXX** is that, to buy and hold the common combination of U.S. stock.
Answer	策略	strategy

Figure 1: Example training sample in People Daily datasets (the English translation is given in the right box). The ”**XXXXX**“ represents the missing word. In this example, the document consists of 10 sentences, and the 9th sentence is chosen as the query.

- We provide a large-scale Chinese reading comprehension data in news domain, as well as its validation and test data as the in-domain test.
- Further, we release two out-of-domain test sets, and it deserves to highlight that one of the test sets is made by the humans, which makes it harder to answer than the automatically generated test set.

People Daily. We roughly collected 60K news articles from the People Daily website[4]. Following Liu et al. (2016), we process the news articles into the triple form $\langle \mathcal{D}, \mathcal{Q}, \mathcal{A} \rangle$. The detailed procedures are as follows.

- Given a certain document $\mathcal{D}$, which is composed by a set of sentences $\mathcal{D} = \{s_1, s_2, ..., s_n\}$, we randomly choose an answer word $\mathcal{A}$ in the document. Note that, we restrict the answer word $\mathcal{A}$ to be a noun, as well as the answer word should appear at least twice in the document. The part-of-speech and sentence segmentation is identified using LTP Toolkit (Che et al., 2010). We do not distinguish the named entities and common nouns as Hill et al. (2015) did.
- Second, after the answer word $\mathcal{A}$ is chosen, the sentence that contains $\mathcal{A}$ is defined as the query $\mathcal{Q}$, in which the answer word $\mathcal{A}$ is replaced by a specific placeholder $\langle X \rangle$.
- Third, given the query $\mathcal{Q}$ and document $\mathcal{D}$, the target of the prediction is to recover the answer $\mathcal{A}$.

In this way, we can generate tremendous triples of $\langle \mathcal{D}, \mathcal{Q}, \mathcal{A} \rangle$ for training the proposed neural network, without any assumptions on the nature of the original corpus. Note that, unlike the previous work, using the method mentioned above, the document can be re-used for different queries, which makes it more general to generate large-scale training data for neural network training. Figure 1 shows an example of People Daily datasets.

Children's Fairy Tale. Except for the validation and test set of People Daily news data, we also present two out-of-domain test sets as well. The two out-of-domain test sets are made from the Children's Fairy Tale (CFT), which is fairly different from the news genre. The reason why we set out-of-domain test sets is that, the children's fairy tale mainly consists of the stories of animals or virtualized characters, and

[4] http://www.people.com.cn

	People Daily			Children's Fairy Tale	
	Train	Valid	Test	Test-auto	Test-human
# Query	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	248,160			N/A	

Table 1: Statistics of People Daily datasets and Children's Fairy Tale datasets.

this prevents us from utilizing the gender information and world knowledge in the training data, which is important when solving several types of questions, such as coreference resolutions etc.

In CFT dataset, one test set is automatically generated using the algorithms described above, and the other one is made by the human, which suggest that the latter is harder than the former one. Because the automatically generated test sets are aware of the co-occurence or fixed collocation of words, and thus when the pattern around the query blank exactly appeared in the document, it is much easier for the machine to identify the correct answer. While in building human evaluation test set, we have eliminated these types of samples, which makes it harder for the machine to comprehend. Intuitively, the human evaluation test set is harder than any other previously published Cloze-style test sets.

The statistics of People Daily news datasets as well as Children's Fairy Tale datasets are listed in the Table 1.

3 Consensus Attention Sum Reader

In this section, we will introduce our attention-based neural network model for Cloze-style reading comprehension task, namely Consensus Attention Sum Reader (CAS Reader). Our model is primarily motivated by Kadlec et al. (2016), which aims 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 every time slices of query, and make a *consensus* attention among different steps.

Formally, when given a set of training triple $\langle \mathcal{D}, \mathcal{Q}, \mathcal{A} \rangle$, we will construct our network in the following way. We first convert one-hot representation of the document $\mathcal{D}$ and query $\mathcal{Q}$ into continuous representations with a shared embedding matrix W_e. As the query is typically shorter than the document, by sharing the embedding weights, the query representation can be benefited from the embedding learning in the document side, which is better than separating embedding matrices individually.

Then we use two different bi-directional RNNs to get the contextual representations of document and query, which can capture the contextual information both in history and future. In our implementation, we use the bi-directional Gated Recurrent Unit (GRU) for modeling. (Cho et al., 2014)

$$e(x) = W_e * x, \ where \ \ x \in \mathcal{D}, \mathcal{Q} \tag{1}$$

$$\overrightarrow{h_s} = \overrightarrow{GRU}(e(x)) \tag{2}$$

$$\overleftarrow{h_s} = \overleftarrow{GRU}(e(x)) \tag{3}$$

$$h_s = [\overrightarrow{h_s}; \overleftarrow{h_s}] \tag{4}$$

We take h_{doc} and h_{query} to represent the contextual representations of document and query, both of which are in 3-dimension tensor shape. After that, we directly make a dot product of h_{doc} and $h_{query}(t)$ to get the "importance" of each document word, in respect to the query word at time t. And then,

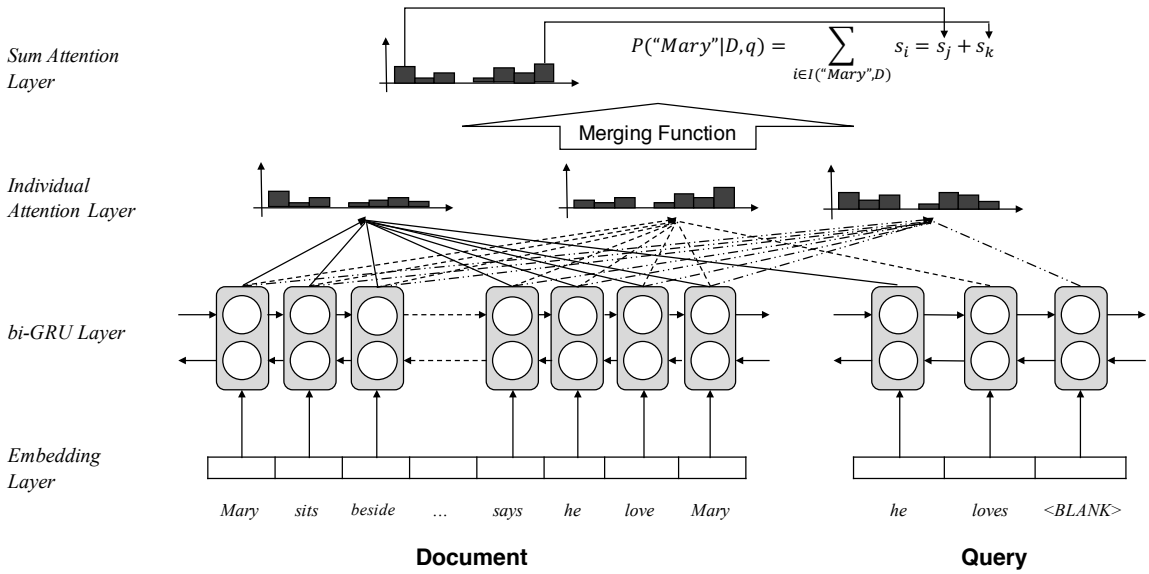

Figure 2: Architecture of the proposed Consensus Attention Sum Reader (CAS Reader).

we use the softmax function to get a probability distribution α over the document h_{doc}, also known as "attention".

$$\alpha(t) = softmax(h_{doc} \odot h_{query}(t)) \tag{5}$$

In this way, for every time step t in the query, we can get a probability distribution over the document, denoted as $\alpha(t)$, where $\alpha(t) = [\alpha(t)_1, \alpha(t)_1, ..., \alpha(t)_n]$, $\alpha(t)_i$ means the attention value of ith word in the document at time t, and n is the length of the document. To get a *consensus attention* over these individual attentions, we explicitly define a merging function f over $\alpha(1)...\alpha(m)$. We denote this as

$$s = f(\alpha(1), ..., \alpha(m)) \tag{6}$$

where s is the final attention over the document, m is the length of the query. In this paper, we define the merging function f as one of three heuristics, shown in equations below.

$$s \propto \begin{cases} softmax(\sum_{t=1}^{m} \alpha(t)), & \text{if } mode = sum; \\ softmax(\frac{1}{m} \sum_{t=1}^{m} \alpha(t)), & \text{if } mode = avg; \\ softmax(\max_{t=1...m} \alpha(t)), & \text{if } mode = max. \end{cases} \tag{7}$$

Finally, we map the attention result s to the vocabulary space V, and sum the attention value which occurs in different place of the document but shares the same word, as Kadlec et al. (2016) do.

$$P(w|\mathcal{D}, \mathcal{Q}) = \sum_{i \in I(w,\mathcal{D})} s_i, \;\; w \in V \tag{8}$$

where $I(w, \mathcal{D})$ indicate the position that word w appear in the document $\mathcal{D}$. Figure 2 shows the proposed neural network architecture.

4 Experiments

4.1 Experimental Setups

Training details of neural network models are illustrated as follows.

	Embed. # units	Hidden # units	Dropout
CNN News	384	256	None
CBTest NE	384	384	None
CBTest CN	384	384	None
People Daily & CFT	256	256	0.1

Table 2: Other neural network setups for each task. Note that, the dropout is only applied to the output of the GRUs.

	CNN News			CBT NE			CBT CN		
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
# Query	380,298	3,924	3,198	108,719	2,000	2,500	120,769	2,000	2,500
Max # candidates	527	187	396	10	10	10	10	10	10
Avg # candidates	26	26	25	10	10	10	10	10	10
Avg # tokens	762	763	716	433	412	424	470	448	461
Vocabulary	118,497			53,063			53,185		

Table 3: Statistics of public Cloze-style reading comprehension datasets: CNN news data and CBTest NE(Named Entites) / CN(Common Nouns).

- Embedding Layer: We use randomly initialized embedding matrix with uniformed distribution in the interval $[-0.1, 0.1]$. Note that, no pre-trained word embeddings are used in our experiments.
- Hidden Layer: We initialized the GRU units with random orthogonal matrices (Saxe et al., 2013). As GRU still suffers from the gradient exploding problem, we set gradient clipping threshold to 10 in our experiments (Pascanu et al., 2013) .
- Vocabulary: For training efficiency and generalization, in People Daily and CFT datasets, we truncate the full vocabulary (about 200K) and set a shortlist of 100K. All unknown words are mapped to 10 different specific symbols using the method proposed by Liu et al. (2016). There is no vocabulary truncation in CNN and CBTest dataset.
- Optimization: We used the ADAM update rule (Kingma and Ba, 2014) with an initial learning rate $lr = 0.0005$, and used negative log-likelihood as the training objective function. The batch size is set to 32.

Other neural network setups, such as dimensions of embedding layer and hidden layer, and dropout (Srivastava et al., 2014) for each task, are listed in Table 2. We trained model for several epochs and choose the best model according to the performance of validation set. All models are trained on Tesla K40 GPU. Our model is implemented with Theano (Theano Development Team, 2016) and Keras (Chollet, 2015).

4.2 Results on Public Datasets

To verify the effectiveness of our proposed model, we first tested our model on public datasets. Our evaluation is carried out on CNN news datasets (Hermann et al., 2015) and CBTest NE/CN datasets (Hill et al., 2015), and the statistics of these datasets are listed in Table 3. No pre-processing is done with these datasets. The experimental results are given in Table 4. We evaluate the model in terms of its accuracy. Due to the time limitations, we did not evaluate our model in ensemble.

CNN News. The performance on CNN news datasets shows that our model is on par with the Attention Sum Reader, with 0.4% decrease in validation and 0.5% improvements in the test set. But we failed to outperform the Stanford AR model. While the Stanford AR utilized GloVe embeddings (Pennington et

	CNN News		CBTest NE		CBTest CN	
	Valid	Test	Valid	Test	Valid	Test
Deep LSTM Reader†	55.0	57.0	-	-	-	-
Attentive Reader†	61.6	63.0	-	-	-	-
Impatient Reader†	61.8	63.8	-	-	-	-
Human (context+query)‡	-	-	-	*81.6*	-	*81.6*
LSTMs (context+query)‡	-	-	51.2	41.8	62.6	56.0
MemNN (window + self-sup.)‡	63.4	66.8	70.4	66.6	64.2	63.0
Stanford AR♮	*72.4*	*72.4*	-	-	-	-
AS Reader♯	**68.6**	69.5	73.8	68.6	**68.8**	63.4
CAS Reader (mode: avg)	68.2	**70.0**	**74.2**	**69.2**	68.2	**65.7**

Table 4: Results on the CNN news, CBTest NE (named entity) and CN (common noun) datasets. Results marked with † are taken from (Hermann et al., 2015), and ‡ are taken from (Hill et al., 2015), and ♮ are taken from (Chen et al., 2016), and ♯ are taken from (Kadlec et al., 2016)

	People Daily		Children's Fairy Tale	
	Valid	Test	Test-auto	Test-human
AS Reader	64.1	67.2	40.9	33.1
CAS Reader (mode: avg)	**65.2**	**68.1**	41.3	**35.0**
CAS Reader (mode: sum)	64.7	66.8	**43.0**	34.7
CAS Reader (mode: max)	63.3	65.4	38.3	32.0

Table 5: Results on People Daily datasets and Children's Fairy Tale (CFT) datasets.

al., 2014), and only normalized the probabilities over the named entities in the document, rather than all the words, and this could make a difference in the results. But in our model, we do not optimize for a certain type of dataset, which make it more general.

CBTest NE/CN. In CBTest NE dataset, our model gives slight improvements over AS Reader, where 0.4% improvements in the validation set and 0.6% improvements in the test set. In CBTest CN, though there is a slight drop in the validation set with 0.6% declines, there is a boost in the test set with an absolute improvements 2.3%, which suggest our model is effective, and it is beneficial to consider every slices of the query when answering.

4.3 Results on Chinese Reading Comprehension Datasets

The results on Chinese reading comprehension datasets are listed in Table 5. As we can see that, the proposed CAS Reader significantly outperform the AS Reader in all types of test set, with a maximum improvements 2.1% on the CFT test-auto dataset. The results indicate that making a consensus attention over multiple time steps are better than just relying on single attention (as AS Reader did). This is similar to the use of "model ensemble", which is also a consensus voting result by different models.

We also evaluated different merging functions. From the results, we can see that the avg and sum methods significantly outperform the max heuristics, and the max heuristics failed to outperform the AS Reader. A possible reason can be explained that the max operation is very sensitive to the noise. If a non-answer word is given to a high probability in one time step of the query, the avg and sum could easily diminish this noise by averaging/summing over other time steps. But once there is a higher value given to a non-answer word in max situation, the noise can not be removed, and will preserve till the end of final attentions, which will influence the predictions a lot.

Also, we have noticed that, though we have achieved over 65% in accuracy among People Daily datasets, there is a significant drop in the two CFT test sets. Furthermore, the the human evaluated test set meets a sharp decline over 8% accuracy to the automatically generated test set. The analyses can be

concluded as

- As we regard the CFT datasets as the out-of-domain tests, there is a gap between the training data and CFT test data, which poses declines in these test sets. Such problems can be remedied by introducing the similar genre of training data.

- Regardless of the absolute accuracies in CFT datasets, the human test set is much harder for the machine to read and comprehend as we discussed before. Through these results, we can see that there is a big gap between the automatically generated queries and the human-selected questions.

Note that, in our human-evaluated test set, the query is also formulated from the *original* sentence in the document, which suggest that if we use more general form of queries, there should be another rise in the comprehension difficulties. For example, instead of asking "I went to the **XXXXX** this morning .", we change into a general question form of "Where did I go this morning ?", which makes it harder for the machine to comprehend, because there is a gap between the general question form and the training data.

5 Related Work

Many NN-based reading comprehension models have been proposed, and all of them are attention-based models, which indicate that attention mechanism is essential in machine comprehensions.

Hermann et al. (2015) have proposed a methodology for obtaining a large quantities of $\langle \mathcal{D}, \mathcal{Q}, \mathcal{A} \rangle$ triples. By using this method, a large number of training data can be obtained without much human intervention, and make it possible to train a reliable neural network to study the inner relationships inside of these triples. They used attention-based neural networks for this task. Evaluation on CNN/DailyMail datasets showed that their approach is effective than traditional baselines.

Hill et al. (2015) also proposed a similar approach for large scale training data collections for children's book reading comprehension task. By using window-based memory network and self-supervision heuristics, they have surpass all other methods in predicting named entities(NE) and common nouns(CN) on both the CBT and the CNN QA benchmark.

Our CAS Reader is closely related to the work by Kadlec et al. (2016). They proposed to use a simple model that using the attention result to directly pick the answer from the document, rather than computing the weighted sum representation of document using attention weights like the previous works. The proposed model is typically motivated by Pointer Network (Vinyals et al., 2015). This model aims to solve one particular task, where the answer is only a single word and should appear in the document at least once. Experimental results show that their model outperforms previously proposed models by a large margin in public datasets (both CBTest NE/CN and CNN/DailyMail datasets).

Liu et al. (2016) proposed an effective way to generate and exploit large-scale pseudo training data for zero pronoun resolution task. The main idea behind their approach is to automatically generate large-scale pseudo training data and then using the neural network model to resolve zero pronouns. They also propose a two-step training: a pre-training phase and an adaptation phase, and this can be also applied to other tasks as well. The experimental results on OntoNotes 5.0 corpus is encouraging and the proposed approach significantly outperforms the state-of-the-art methods.

In our work, we proposed an entirely new Chinese reading comprehension dataset, which add the diversity to the existing Cloze-style reading comprehension datasets. Moreover, we propose a refined neural network model, called Consensus Attention-based Sum Reader. Though many impressive progress has been made in these public datasets, we believe that the current machine comprehensions are still in the pre-mature stage. As we have discussed in the previous section, to answer a *pseudo query* to the document is not enough for machine comprehension. The general question form can be seen as a comprehensive processing of our human brains. Though our human-evaluated test set is still somewhat easy for machine to comprehend (but harder than the automatically generated test set), releasing such dataset will let us move a step forward to the real-world questions, and becomes a good bridge between automatic questions and real-world questions.

6 Conclusion

In this paper, we introduce the first Chinese reading comprehension datasets: People Daily and Children's Fairy Tale. Furthermore, we also propose a neural network model to handle the Cloze-style reading comprehension problems. Our model is able to take all question words into accounts, when computing the attentions over the document. Among many public datasets, our model could give significant improvements over various state-of-the-art baselines. And also we set up a baseline for our Chinese reading comprehension datasets, that we hopefully make it as a starter in future studies.

The future work will be carried out in the following aspects. First, we would like to work on another human-evaluated dataset, which will contain the real-world questions and is far more difficult than the existing datasets publicly available. Second, we are going to investigate hybrid reading comprehension models to tackle the problems that rely on comprehensive induction of several sentences.

Acknowledgements

We would like to thank the anonymous reviewers for their thorough reviewing and proposing thoughtful comments to improve our paper. This work was supported by the National 863 Leading Technology Research Project via grant 2015AA015407, Key Projects of National Natural Science Foundation of China via grant 61632011, and National Natural Science Youth Foundation of China via grant 61502120.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. Ltp: A chinese language technology platform. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 13–16. Association for Computational Linguistics.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *Association for Computational Linguistics (ACL)*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics.

François Chollet. 2015. Keras. `https://github.com/fchollet/keras`.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1684–1692.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. *arXiv preprint arXiv:1603.01547*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ting Liu, Yiming Cui, Qingyu Yin, Shijin Wang, Weinan Zhang, and Guoping Hu. 2016. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. *arXiv preprint arXiv:1606.01603*.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. *ICML (3)*, 28:1310–1318.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Andrew M Saxe, James L McClelland, and Surya Ganguli. 2013. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Wilson L Taylor. 1953. Cloze procedure: a new tool for measuring readability. *Journalism and Mass Communication Quarterly*, 30(4):415.

Theano Development Team. 2016. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700.