A Sentence Cloze Dataset for Chinese Machine Reading Comprehension

Yiming Cui[†‡], **Ting Liu**[†], **Ziqing Yang**[‡], **Zhipeng Chen**[‡], **Wentao Ma**[‡],
Wanxiang Che[†], **Shijin Wang**[‡§], **Guoping Hu**[‡]
[†]Research Center for SCIR, Harbin Institute of Technology, Harbin, China
[‡]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China
[§]iFLYTEK AI Research (Hebei), Langfang, China
[†]{ymcui,tliu,car}@ir.hit.edu.cn
[‡§]{ymcui,zqyang5,zpchen,wtma,sjwang3,gphu}@iflytek.com

Abstract

Owing to the continuous efforts by the Chinese NLP community, more and more Chinese machine reading comprehension datasets become available. To add diversity in this area, in this paper, we propose a new task called Sentence Cloze-style Machine Reading Comprehension (SC-MRC). The proposed task aims to fill the right candidate sentence into the passage that has several blanks. We built a Chinese dataset called CMRC 2019 to evaluate the difficulty of the SC-MRC task. Moreover, to add more difficulties, we also made fake candidates that are similar to the correct ones, which requires the machine to judge their correctness in the context. The proposed dataset contains over 100K blanks (questions) within over 10K passages, which was originated from Chinese narrative stories. To evaluate the dataset, we implement several baseline systems based on the pre-trained models, and the results show that the state-of-the-art model still underperforms human performance by a large margin. We release the dataset and baseline system to further facilitate our community. Resources available through `https://github.com/ymcui/cmrc2019`

1 Introduction

Machine Reading Comprehension (MRC) is a task to comprehend given articles and answer the questions based on them, which is an important ability for artificial intelligence. The recent MRC research was originated from the cloze-style reading comprehension (Hermann et al., 2015; Hill et al., 2015; Cui et al., 2016), which requires to fill in the blank with a word or named entity, and following works on these datasets have laid the foundations of this research (Kadlec et al., 2016; Cui et al., 2017; Dhingra et al., 2017). Later on, SQuAD (Rajpurkar et al., 2016) was proposed, and the answer transformed from a single word to a span, which has become a representative span-extraction dataset and massive neural network approaches (Wang and Jiang, 2016; Xiong et al., 2016; Wang et al., 2017; Hu et al., 2018; Wang et al., 2018; Yu et al., 2018) have been proposed which further accelerated the MRC research.

Besides the MRC in English text, we have also seen rapid progress on Chinese MRC research. Cui et al. (2016) proposed the first Chinese cloze-style reading comprehension dataset: People Daily & Children's Fairy Tale (PD&CFT). Later, Cui et al. (2018) proposed another dataset for CMRC 2017, which is gathered from children's reading books, consisting of both cloze and natural questions. He et al. (2018) proposed a large-scale open-domain Chinese reading comprehension dataset (DuReader), which consists of 200k queries annotated from the user query logs on the search engine. In span-extraction MRC, Cui et al. (2019b) proposed CMRC 2018 dataset for Simplified Chinese, and Shao et al. (2018) proposed DRCD dataset for Traditional Chinese, similar to the popular dataset SQuAD (Rajpurkar et al., 2016). Zheng et al. (2019) proposed a large-scale Chinese idiom cloze dataset.

Though various efforts have been made, most of these datasets stop at token-level or span-level inference, which neglect the importance of long-range reasoning of the context. Moreover, powerful pre-trained models such as BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), RoBERTa (Liu et al.,

---

[Passage]
"森林里有一棵大树，树上有一个鸟窝。**[BLANK1]**，还从来没有看到过鸟宝宝长什么样。小松鼠说："我爬到树上去看过，鸟宝宝光溜溜的，身上一根羽毛也没有。" "我不相信，"小白兔说，"所有的鸟都是有羽毛的。" "鸟宝宝没有羽毛，"小松鼠说，"你不信自己去看。"小白兔不会爬树，它没有办法去看。小白兔说："我请蓝狐狸去看一看，我相信蓝狐狸的话。"小松鼠说："蓝狐狸跟你一样，也不会爬树。"蓝狐狸说："我有魔法树叶，我能变成一只狐狸鸟。" **[BLANK2]**，一下子飞到了树顶上。"蓝狐狸，你看到了吗？"小白兔在树下大声喊。"我看到了，鸟窝里有四只小鸟，他们真是光溜溜的，一根羽毛也没有。"蓝狐狸说。就在这时候，鸟妈妈和鸟爸爸回来了，**[BLANK3]**，立刻大喊大叫："抓强盗啊！抓强盗啊！强盗闯进了我们家里，想偷我们的孩子！" **[BLANK4]**，全都飞了过来。他们扇着翅膀，朝蓝狐狸冲过来，用尖尖的嘴啄他，用爪子抓他。蓝狐狸扑扇翅膀，赶紧飞。鸟儿们排着队伍，紧紧追上来。**[BLANK5]**，它飞得不高，也飞得不快。"救命啊，救命！"蓝狐狸说，"我不是强盗，我是蓝狐狸！"

[Candidates]
0: 蓝狐狸是第一次变成狐狸鸟
1: 森林里所有的鸟听到喊声
2: 他们看到鸟窝里蹲着一只蓝色的大鸟
3: 蓝狐狸真的变成了一只蓝色的大鸟
4: 小动物们只看到过鸟妈妈和鸟爸爸在鸟窝里飞进飞出
5: 小松鼠变成了一只蓝色的大鸟

[Answers]
4, 3, 2, 1, 0

[Passage]
A long time ago, there was a queen. **[BLANK1]** Soon after the child was born, the Queen died. **[BLANK2]** The stepmother didn't like her very much. She made Snow White do the housework all day and all night. A wizard had given this Queen a glass. The glass could speak. It was on the wall in the Queen's room. Every day the Queen looked in the glass to see how beautiful she was. As she looked in the glass, she asked: "Tell me, glass upon the wall, who is most beautiful of all?" And the glass said: "The Queen is most beautiful of all.". Years went by. Snow-white grew up and became a little girl. Every day the Queen looked in the glass and said, "Tell me, glass upon the wall, **[BLANK3]**" And the glass said, "Snow-white is most beautiful of all.". When the Queen heard this, **[BLANK4]**. She said, "Snow-white is not more beautiful than I am. There is no one who is more beautiful than I am.". So she called a hunter and said, "Take Snow-white into the forest and kill her.". The hunter took Snow-white to the forest, but he did not kill her, because she was so beautiful and so lovely. He put Snow White in the forest and went away.

[Candidates]
0: The king married another queen
1: She had a pretty daughter named Snow White
2: The king was also passed away
3: who is most beautiful of all?
4: she was very happy
5: she was very angry

[Answers]
1, 0, 3, 5

Figure 1: Examples of the proposed CMRC 2019 dataset. The candidate with underline means it is a fake candidate (does not belong to any blank). For clarity, we also provide an English example.

2019) have surpassed human performance on various MRC datasets, such as SQuAD (Rajpurkar et al., 2016), SQuAD 2.0 (Rajpurkar et al., 2018), CoQA (Reddy et al., 2019), RACE (Lai et al., 2017), etc.

To further test the machine comprehension ability, In this paper, we propose a new task called Sentence Cloze-style Machine Reading Comprehension (SC-MRC). The proposed task preserves the simplicity of cloze-style reading comprehension but requires sentence-level inference when filling the blanks. Figure 1 shows an example of the proposed dataset. We conclude our contributions in three aspects.

- We propose a new machine reading comprehension task called Sentence Cloze-style Machine Reading Comprehension (SC-MRC), which aims to test the ability of sentence-level inference.

- We release a challenging Chinese dataset CMRC 2019, which consists of 100K blanks, to evaluate the SC-MRC task.[1]

- Experiments on several state-of-the-art Chinese pre-trained language models show that there is still much room for them to surpass human performance, indicating the proposed data is challenging.

2 The Proposed Dataset

2.1 Task Definition

Generally, the reading comprehension task can be described as a triple $\langle \mathcal{P}, \mathcal{Q}, \mathcal{A} \rangle$, where $\mathcal{P}$ represents **Passage**, $\mathcal{Q}$ represents **Question**, and the $\mathcal{A}$ represents **Answer**. Specifically, for sentence cloze-style reading comprehension task, we select several sentences in the passages and replace with special marks (for example, `[BLANK]`), forming an incomplete passage. The sentences are identified using LTP (Che et al., 2010), and we further split the sentence with comma and period mark, as some of the sentences are too long. The selected sentences form a candidate list, and the machine should fill in the blanks with these candidate sentences to form a complete passage. Note that, to add more difficulties, we could also add the fake candidates, which do not belong to any blanks in the passage.

2.2 Passage Selection

The raw material of the proposed dataset is from children's books, containing fairy tales and narratives, which is the proper genre for testing the sentence-level inference ability, requiring the correct sentence order of the stories. During the passage selection, we restrict the character-level passage length in the

---

[1]The data was used in the shared task of CMRC 2019 workshop, as thus, we directly name this dataset as CMRC 2019.

range of 500 to 750. If the passage is too short, then there will be only few blanks in the passage. If the passage is too long, it will be harder for the model to process. After the passage selection, we got 10k passages and split them into three parts for generating the training, development, and test set.

	Genre	Query Type	Answer Type	Doc #	Query #
PD&CFT (Cui et al., 2016)	News, Story	Cloze	Word	28K	100K
WebQA (Li et al., 2016)	Web	NQ	Entity	-	42K
CMRC 2017 (Cui et al., 2018)	News	Cloze&NQ	Word	-	364K
DuReader (He et al., 2018)	Web	NQ	Free Form	1M	200K
CMRC 2018 (Cui et al., 2019b)	Wiki	NQ	Passage Span	-	18K
DRCD (Shao et al., 2018)	Wiki	NQ	Passage Span	-	34K
C^3 (Sun et al., 2019)	Mixed	NQ	Choices	14K	24K
ChID (Zheng et al., 2019)	News, Novels, etc.	Cloze	Chinese Idioms	580K	729K
CMRC 2019	Story	Cloze	Sentence	10K	100K

Table 1: Comparisons of Chinese MRC datasets. NQ represents natural questions.

2.3 Cloze Generation

Sentence cloze task does not require human annotation as it only requires the selection of the blanks, and they will naturally become the answers. However, to ensure a high-quality cloze generation, the following rules are applied.

- The first sentence is always skipped, which usually contains important topic information.

- Select the sentence based on the comma or period mark, resulting in the range of 10 to 30 characters. Note that we eliminate the comma or period at the end of the candidate sentence.

- If a part of a long sentence is selected, we do not choose other parts to avoid too many consecutive blanks.

2.4 Fake Candidates

In order to bring difficulties in this task and better test the ability of machine reading comprehension, we propose to add fake candidates to confuse the system. In this way, the machine should not only generate the correct order of the candidate sentences but also should identify the fake candidates that do not belong to any passage blanks. A good fake candidate should have the following characteristics.

- The topic of the fake candidate should be the same as the passage.

- If there are named entities in the fake candidates, it should also appear in the passage.

- It could NOT be a machine-generated sentence, or it would be very easy for the machine to pick the fake one out.

A natural way to generate fake candidates is to adopt human annotation, while it is rather time-consuming. In order to minimize the cost by human annotation, in this paper, we propose a novel approach to generate fake candidates that is qualified for the requirements above.

Typically, a complete story is rather long that we must truncate for easy processing by the machines. In this context, we could directly pick the sentences outside the truncated passage within the same story. As these sentences are still from the same story, the topic and name entities are in accordance with the main passage. Also, it is a part of the original story, which is a natural sentence rather than a machine-generated sentence. Using the strategies above, we could generate many fake candidates and mix them with the correct candidates to form the final candidate sentences.

2.5 Statistics

The general statistics of the final data are given in Table 2, and comparisons with other Chinese MRC datasets are shown in Table 1. As we can see, the proposed dataset mitigates the absence of sentence-level inferential reading comprehension dataset. Note that the training set does not contain any fake

	Train	Dev	Test
Context #	9,638	300	500
Blank #	100,009	3,053	5,118
Max Context Tokens #	731	717	717
Avg Context Tokens #	642	632	633
Max Candidate #	15	15	15
Avg Candidate #	10.4	13.3	13.4
Max True Candidate #	15	14	14
Avg True Candidate #	10.4	10.2	10.2
Max Candidate Tokens #	29	29	29
Avg Candidate Tokens #	13.7	14.1	14.2

Table 2: Statistics of CMRC 2019.

candidates, as we want to test the generalization of the machine reading comprehension system without training on both real and fake candidates.

3 Baseline System

In this paper, we mainly adopt BERT and its related variants for our baseline systems.

- **Input Sequence:** Given a passage p and its n answer options $\{a_1, a_2, \ldots, a_n\}$, we first replace the blanks in p with the special tokens [unused*Num*] from the vocabulary to fit the input format of BERT, where *Num* ranges from 0 to number of blanks -1. Then for each a_i in the answer options, we concatenate a_i and p with the token [SEP] as the input sequence.

- **Main Model:** The input sequence of length l is fed into BERT to get the hidden representations $H \in \mathbb{R}^{l \times d}$. The dot product of H with trainable parameters $w \in \mathbb{R}^d$ gives the logits $t = H \cdot w$, where $t \in \mathbb{R}^l$. Finally, the probabilities of the blanks for the current option is calculated by a softmax over the logits with only positions of blanks unmasked. The training objective is to minimize the cross-entropy between the predicted probabilities and the ground-truth positions.

- **Decoding:** The model outputs the predictions for answer options in terms of the probabilities of blanks they can be filled into. They need to be transformed into the predictions for blanks in terms of the answers they choose. A simple method we used is, among all the answer options for a passage, taking the option that gives the highest probability to a blank as the prediction for that blank (each option is allowed to be the prediction of multiple blanks).

4 Experiments

4.1 Evaluation Metrics

We adopt two metrics to evaluate the systems on our datasets, namely Question-level Accuracy (QAC) and Passage-level Accuracy (PAC). QAC is calculated by the ratio between the correct predictions and total blanks. Similarly, PAC is to measure how many passages have been correctly answered. We only count the passages that all blanks have been correctly predicted.

$$\mathbf{QAC} = \frac{\text{\# correct predictions}}{\text{\# total blanks in dataset}} \times 100\% \; ; \; \mathbf{PAC} = \frac{\text{\# correct passages}}{\text{\# total passages in dataset}} \times 100\% \quad (1)$$

4.2 Experimental Setups

We adopt Chinese BERT-base, BERT-multilingual, Chinese BERT-wwm and RoBERTa with whole word masking (Cui et al., 2019a; Cui et al., 2020) as backbones. Note that both genres share the same vocabulary of WordPiece (Wu et al., 2016) tokens as the same in Chinese BERT[2], which have 21,128 words. All models are trained with 3 epochs on Tesla V100, with an initial learning rate of 3e-5, a maximum sequence length of 512, and a batch size of 24. The implementation was done on PyTorch (Paszke et al., 2017) with Transformers library (Wolf et al., 2019).

---

[2]https://github.com/google-research/bert/blob/master/multilingual.md

4.3   Results

The baseline results are shown in Table 3. The Chinese BERT-base model could give a QAC of 71.2 and 71.0 on the development and test set, respectively. However, with respect to the PAC metric, it only gives an accuracy of below 10, which suggests that there is plenty of room for optimizing the sentence cloze procedure to consider not only the single cloze but also the coherence of the whole passage. BERT with whole word masking strategy substantially outperform original BERT implementation, and using the large model could also give a significant boost on both QAC and PAC metrics.

To evaluate human performance, we invited qualified annotators (Chinese-majored master students) to solve the sentence clozes of 100 passages in the development and test set (randomly sampled), respectively, resulting in 1,016 and 1,027 blanks for each. For each set, three annotators are involved. Then we calculate the average QAC and PAC to roughly estimate the human performance on this dataset.

Comparing the RoBERTa-wwm-ext-large with the human performance, though there is only a gap of 13.3 on QAC, there is a significant gap on PAC, which also suggests that more attention should be drawn on the accuracy of the passage as a whole. We also include the top systems in our evaluation campaign, which used various approaches for improving the final performance, including pseudo-training data generation, data augmentation, ensemble, etc. However, comparing these models with human performance, we can see that there is still much room for improvement, indicating that our dataset is challenging.

System	Dev		Test	
	QAC	PAC	QAC	PAC
Human Performance	*95.9*	*81.0*	*95.3*	*75.0*
Random Selection	7.6	0.0	7.5	0.0
Top Submissions from CMRC 2019				
bert_scp_spm[†]	90.9	60.0	90.8	57.6
mojito[†]	88.2	48.0	86.0	41.8
DA-BERT[†]	86.3	34.3	84.4	27.6
Baseline Systems				
BERT	71.2	10.0	71.0	8.8
BERT-multilingual	66.8	6.67	66.0	5.0
BERT-wwm	72.4	9.3	71.4	7.6
BERT-wwm-ext	75.0	12.7	73.7	9.2
RoBERTa-wwm-ext	75.9	11.0	75.8	12.4
RoBERTa-wwm-ext-large	82.6	23.3	81.7	23.0

Table 3: Experimental results on CMRC 2019. The ensemble systems (unpublished) are marked with †.

5   Conclusion

We proposed a new task called Sentence Cloze-style Machine Reading Comprehension (SC-MRC) and released a Chinese dataset for this purpose. The proposed dataset contains both real and fake candidate sentences for filling the clozes, which not only requires the machine to choose the correct sentence but also distinguishes the real sentence from fake sentences. We built up baseline models based on the popular pre-trained language models, and the results show that the state-of-the-art models still underperform the human performance, especially on PAC evaluation metric. We hope the release of this dataset could bring language diversity in machine reading comprehension tasks and accelerate further investigation on solving the questions that need comprehensive reasoning over multiple clues.

Acknowledgments

References

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.

Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2016. Consensus attention-based neural networks for chinese reading comprehension. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1777–1786. The COLING 2016 Organizing Committee.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-over-attention neural networks for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 593–602. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. Dataset for the first evaluation on chinese machine reading comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2721–2725, Paris, France, may. European Language Resources Association (ELRA).

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019a. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019b. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5886–5891, Hong Kong, China, November. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. *Findings of EMNLP 2020*.

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

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1832–1846. Association for Computational Linguistics.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia, July. Association for Computational Linguistics.

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*.

Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2018. Reinforced mnemonic reader for machine reading comprehension. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4099–4106. International Joint Conferences on Artificial Intelligence Organization, 7.

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 908–918. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 796–805. Association for Computational Linguistics.

Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. 2016. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. *ArXiv*, abs/1607.06275.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. Drcd: a chinese machine reading comprehension dataset. *arXiv preprint arXiv:1806.00920*.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019. Probing prior knowledge needed in challenging chinese machine reading comprehension. *ArXiv*, abs/1904.09679.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198. Association for Computational Linguistics.

Wei Wang, Ming Yan, and Chen Wu. 2018. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1705–1714. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese IDiom dataset for cloze test. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 778–787, Florence, Italy, July. Association for Computational Linguistics.