D-NET: A Simple Framework for Improving the Generalization of Machine Reading Comprehension

Hongyu Li, Xiyuan Zhang, Yibing Liu, Yiming Zhang,
Quan Wang, Xiangyang Zhou, Jing Liu, Hua Wu, Haifeng Wang
Baidu Inc., Beijing, China
{lihongyu04, zhangxiyuan01, liuyibing01, zhangyiming04,
wangquan05, zhouxiangyang, liujing46, wu_hua, wanghaifeng}@baidu.com

Abstract

In this paper, we introduce a simple system Baidu submitted for MRQA (Machine Reading for Question Answering) 2019 Shared Task that focused on generalization of machine reading comprehension (MRC) models. Our system is built on a framework of pretraining and fine-tuning, namely D-NET. The techniques of pre-trained language models and multi-task learning are explored to improve the generalization of MRC models and we conduct experiments to examine the effectiveness of these strategies. Our system is ranked at top 1 of all the participants in terms of averaged F1 score. Our codes and models will be released at PaddleNLP [1].

1 Introduction

Machine reading comprehension (MRC) requires machines to understand text and answer questions about the text, and it is an important task in natural language processing (NLP). With the increasing availability of large-scale datasets for MRC (Rajpurkar et al., 2016; Bajaj et al., 2016; Dunn et al., 2017; Joshi et al., 2017; He et al., 2018) and the development of deep learning techniques, MRC has achieved remarkable advancements in the last few years (Wang and Jiang, 2016; Seo et al., 2016; Xiong et al., 2016; Wang et al., 2017; Liu et al., 2018; Wang et al., 2018; Yu et al., 2018). Although a number of neural models obtain even human parity performance on several datasets, these models may generalize poorly on other datasets (Talmor and Berant, 2019).

We expect that a truly effective question answering system works well on both the examples drawn from the same distribution as the training data and the ones draw from different distributions. Nevertheless, there has been relatively little work that explores the generalization of MRC models.

This year, MRQA (Machine Reading for Question Answering) 2019 Shared Task tries to test whether the question answering systems can generalize well beyond the datasets on which they are trained. Specifically, participants will submit question answering systems trained on a training set pooled from six existing MRC datasets, and the systems will be evaluated on twelve different test datasets without any additional training examples in the target domain (i.e. generalization).

As shown in Table 1, the major challenge of the shared task is that the train and test datasets differ in the following ways:

- **Questions**: They come from different sources, e.g. crowdsourcing workers, examine writers, search logs, synthetics, etc.
- **Documents**: They involve passages from different sources, e.g. wikipedia, news, movies, textbook, etc.
- **Language Understanding Ability**: They might require different language understanding abilities, e.g. matching, reasoning and arithmetic.

To address the above challenge, we introduce a simple framework of pre-training and fine-tuning, namely D-NET, for improving the generalization of MRC models by exploring the following techniques:

- **Pre-trained Models**: We leverage multiple pre-trained models, e.g. BERT (Devlin et al., 2019), XLNET (Yang et al., 2019) and ERNIE 2.0 (Sun et al., 2019). Since different pre-trained models are trained on various

[1] https://github.com/PaddlePaddle/models/tree/develop/PaddleNLP/Research/MRQA2019-D-NET

Dataset	Question Sources	Document Sources	Language Understanding	Train	Dev	Test
SQuAD	Crowdsourced	Wiki.	Matching	✓	✓	
NewsQA	Crowdsourced	News	Matching	✓	✓	
TriviaQA	Trivia	Web Snippets	Matching	✓	✓	
SearchQA	Trivia	Web Snippets	Matching	✓	✓	
HotpotQA	Crowdsourced	Wiki.	Reasoning	✓	✓	
NaturalQuestions	Query Log	Wiki.	Matching	✓	✓	
BioASQ	Crowdsourced	Biomedical articles	Matching		✓	✓
DROP	Crowdsourced	Wiki.	Arithmetic		✓	✓
DuoRC	Crowdsourced	Movie	Reasoning		✓	✓
RACE	Teachers	Examination	Reasoning		✓	✓
RelationExtraction	Question Template	Wiki.	Matching		✓	✓
TextbookQA	Textbook	Textbook	Reasoning		✓	✓
BioProcess	Biologist	Biology Textbook	Reasoning			✓
ComplexWebQuestions	Synthetic & Rephrasing	Web Snippets	Reasoning			✓
MCTest	Crowdsourced	Story	Reasoning			✓
QAMR	Crowdsourced	Wiki.&News	Matching			✓
QAST	Crowdsourced	Speech Transcriptions	Matching			✓
TREC	Query Log	Web doc.	Matching			✓

Table 1: The datasets of MRQA 2019 Shared Task include 6 training sets and 12 testing sets. The train, dev and test datasets differ in the following ways (1) question sources; (2) document sources; (3) language understanding

corpus with different pre-training tasks (e.g. masked language model, discourse relations, etc.), they may capture different aspects of linguistics. Hence, we expect that the combination of these pre-trained models can improve the generalization capability of MRC models.

- **Multi-task Learning**: Since the pre-training is usually performed on corpus with restricted domains, it is expected that increasing the domain diversity by further pre-training on other corpus may improve the generalization capability. Hence, we incorporate masked language model by using corpus from various domains as an auxiliary task in the fine-tuning phase, along with MRC. The side effect of adding a language modeling objective to MRC is that it can avoid catastrophic forgetting and keep the most useful features learned from pre-training task (Chronopoulou et al., 2019). Additionally, we explore multi-task learning (Liu et al., 2019) by incorporating the supervised dataset from other NLP tasks (e.g. natural language inference and paragraph ranking) to learn better language representation.

Our system is ranked at top 1 of all the participants in terms of averaged F1 score. We also conduct the experiments to examine the effectiveness of multiple pre-trained models and multi-task learning. Our major observations are as follows:

- The pre-trained models are still the most important keys to improve the generalization of MRC models in our experiments. Moreover, the ensembles of MRC models based on different pre-trained models show better generalization on out-of-domain set than the ensembles of MRC models based on the same pre-trained models.

- The auxiliary task of masked language model can help improve the generalization of MRC models.

- We do not observe much improvements from the auxiliary tasks of natural language inference and paragraph ranking.

The remainder of this paper is structured as follows: Section 2 describes the detailed overview of our system. Section 3 shows the experimental settings and results. Finally, we conclude our work in Section 4.

2 System Overview

Figure 1 depicts D-NET, a simple framework of pre-training and fine-tuning to improve the generalization capability of MRC models. There are basically two stages in D-NET: (1) We incorporate multiple pre-trained language models. (2) We fine-tune MRC models with multi-task learning. In this section, we will introduce each stage in details.

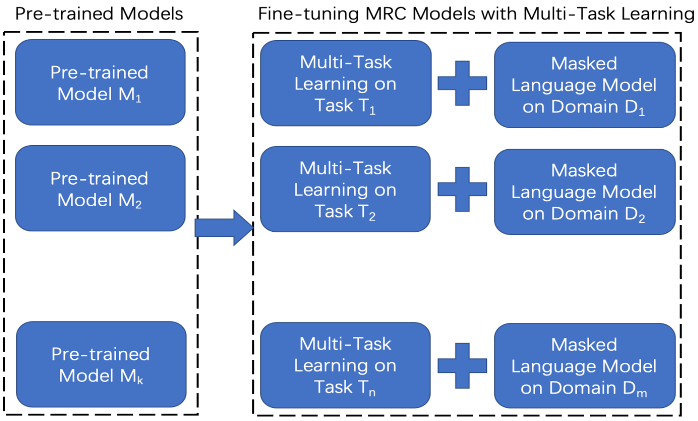
Pre-trained Models
Fine-tuning MRC Models with Multi-Task Learning
Pre-trained Model M1
Pre-trained Model M2
Pre-trained Model Mk
Multi-Task Learning on Task T1
Multi-Task Learning on Task T2
Multi-Task Learning on Task Tn
Masked Language Model on Domain D1
Masked Language Model on Domain D2
Masked Language Model on Domain Dm

Figure 1: D-NET: A framework of pre-training and fine-tuning for MRC.

2.1 Pre-trained Models

Recently pre-trained language models present new state-of-the-art results in MRC. Since different pre-trained models are trained on various corpus with different pre-training tasks, they may capture different aspects of linguistics. Hence, we expect that the combination of these pre-trained models may generalize well on various corpus with different domains. The pre-trained models that are used in our experiments are listed below:

BERT (Devlin et al., 2019) uses multi-layer Transformer encoding blocks as its encoder. The pre-training tasks include masked language model and next sentence prediction, which enable the model to capture bidirectional and global information. In our system, we use the BERT large configuration that contains 24 Transformer encoding blocks, each with 16 self attention heads and 1024 hidden units.

Note that we use this pre-trained model for experimental purpose, and it is not included in the final submission. In our experiments, we initialize the parameters of the encoding layers from the checkpoint [2] of the model (Alberti et al., 2019) namely BERT + N-Gram Masking + Synthetic Self-Training. The model is initialized from Whole Word Masking BERT (BERT$_{wwm}$), further fine-tuned on the SQuAD 2.0 task with synthetic generated question answering corpora. In our experiments, we find that this model performs consistently better than the original BERT$_{large}$ and BERT$_{wwm}$ without synthetic data augmentation, as officially released by Google [3].

XLNET (Yang et al., 2019) uses a novel pre-training task, i.e. permutation language modeling, by introducing two-stream self attention. Besides BooksCorpus and Wikipedia, on which the BERT is trained, XLNET uses more corpus in its pre-training, including Giga5, ClueWeb and Common Crawl. In our system, we use the 'large' configuration that contains 24 layers, each with 16 self attention heads and 1024 hidden units.

We initialize the parameters of XLNET encoding layers using the version that is released by the authors [4]. In our experiments, we find that XLNET shows superior performance on the datasets that require reasoning and arithmetic, e.g. DROP and RACE.

ERNIE 2.0 (Sun et al., 2019) is a continual pre-training framework for language understanding in which pre-training tasks can be incrementally built and learned through multi-task learning. It designs multiple pre-training tasks, including named entity prediction, discourse relation recognition, sentence order prediction, to learn language representations.

ERNIE uses the same Transformer encoder as BERT. In our system, we use the 'large' configuration that contains 24 Transformer encoding blocks, each with 16 self attention heads and 1024 hidden units. We initialize the parameters of ERNIE encoding layer using the official released

[2] The checkpoint can be downloaded from `https://worksheets.codalab.org/worksheets/0xd7b08560b5b24bd1874b9429d58e2df1`

[3] `https://github.com/google-research/bert`

[4] `https://github.com/zihangdai/xlnet/`

Model ID		M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}
Pre-trained Model	BERT	✓	✓	✓	✓	✓	✓					
	XLNET							✓	✓	✓	✓	
	ERNIE											✓
Masked LM	In-domain		✓	✓		✓	✓					
	Search Snippets		✓	✓	✓	✓	✓					
	Y!A				✓							
Supervised Task	MNLI			✓								
	ParaRank						✓					
Hyper Parameters	Max Seq Len	512	512	512	512	512	512	640	640	640	768	512
	Batch Size	48	48	48	48	32	48	128	24	24	24	64
	λ_{MLM}		2.0	0.4	0.4	0.4	0.4					
	λ_{MNLI}			0.4								
	λ_{PR}						0.8					

Table 2: The configurations and hyper-parameters of the eleven models used in our experiments. The configurations include the pre-trained models, the corpus for the masked language model task, the types of supervised NLP tasks. The hyper-parameters include the max sequence length, batch size and the mix ratio λ used the auxiliary tasks in multi-task learning.

version [5].

2.2 Fine-tuning MRC Models with Multi-Task Learning

To fine-tune MRC models, we simply use a linear output layer for each pre-trained model, followed by a standard softmax operation, to predict answer boundaries. We further introduce multi-tasking learning in the fine-tuning stage to learn more general language representations. Specifically, we have the following auxiliary tasks:

Masked Language Model Since the pre-training is usually preformed on the corpus with restricted domains, it is expected that further pre-training on more diverse domains may improve the generalization capability. Hence, we add an auxiliary task, masked language model (Chronopoulou et al., 2019), in the fine-tuning stage, along with the MRC task. Moreover, we use three corpus with different domains as the input for masked language model: (1) the passages in MRQA in-domain datasets that include wikipedia, news and search snippets; (2) the search snippets from Bing [6]. (3) the science questions in Yahoo! Answers.[7]. The side effect of adding a language modeling objective to MRC is that it can avoid catastrophic forgetting and keep the most useful features learned from pre-training task (Chronopoulou et al., 2019).

Supervised Tasks Motivated by (Liu et al., 2019), we explore multi-task learning by incorporating the supervised datasets from other NLP tasks to learn more general language representation.

Specifically, we incorporate natural language inference and paragraph ranking as auxiliary tasks to MRC. (1) Previous work (Clark et al., 2019; Liu et al., 2019) show that MNLI (Williams et al., 2017) (a popular natural language inference dataset) can help improve the performance of the major task in a multi-task setting. In our system, we also leverage MNLI as an auxiliary task. (2) Previous work (Tan et al., 2017; Wang et al., 2018) examine the effectiveness of the joint learning of MRC and paragraph ranking. In our system, we also leverage paragraph ranking as an auxiliary task. We generate the datasets of paragraph ranking from MRQA in-domain datasets. The generated data and the details of data generation will be released at PaddleNLP.

3 Experiments and Results

3.1 Experimental Settings

In our experiments, we train eleven single models (M_0-M_{10}) under the framework of D-NET. Table 2 lists the detailed configurations and the hyper-parameters of these models. In the settings of multi-task leaning, we randomly sample

[5] https://github.com/PaddlePaddle/ERNIE
[6] http://www.msmarco.org/dataset.aspx
[7] http://goo.gl/JyCnZq

Systems	Dev In-domain F1	Dev Out-of-domain F1	Test F1
Official baseline	77.87	58.67	61.76
1 XLNET (M_6) + 1 ERNIE (M_{10}) (*submitted*)	**84.15**	**69.67**	**72.50**
4 BERTs (M_1-M_4)	84.25	68.33	-
4 XLNETs (M_6-M_9)	84.45	69.56	-
1 XLNET (M_6) + 1 BERT*	84.30	69.99	-
1 XLNET (M_6) + 1 ERNIE (M_{10}) + 1 BERT*	**84.82**	**70.42**	-

Table 3: System performance on the development and test set. Our submitted version for the shared task is marked as 'submitted'. Please refer to Table 2 with corresponding model ID for details about the model configurations.
* We use the technique of knowledge distillation to learn a single BERT-based model from a teacher that is an ensemble of 4 BERTs(M_1-M_4).

batches from different tasks with 'mix ratio' $1 : \lambda_{\text{MLM}} : \lambda_{\text{MNLI}} : \lambda_{\text{PR}}$.

When fine-tuning all pre-trained models, we use Adam optimizer with learning rate of 3×10^{-5}, learning rate warmup over the first 10% steps, and linear decay of the learning rate [8]. All the models are fine-tuned for two epochs. The experiments are conducted with PaddlePaddle framework on NVIDA TESLA V100 GPUs (with 32G RAM).

3.2 Experimental Results

3.2.1 The Main Results and the Effects of Pre-trained Models

Table 3 shows the main results and the results for the effects of pre-trained models. From Table 3, we have the following observations:

(1) Our submitted system significantly outperforms the official baseline by about 10 F1 score, and it is ranked at top 1 of all the participants in terms of averaged F1 score [9]. The technique of model ensemble can improve the generalization of MRC models. In the shared task, the participants are required to submit a question answering system which is able to run on a single GPU [10] with certain latency limit. Hence, we choose to submit a system that combines only one XLNET-based model with one ERNIE-based model.

(2) The pre-trained models are still the most important keys to improve the generalization of MRC models in our experiments. For example, pure XLNET-based models perform consistently better than BERT-based models with multi-task learning. Moreover, the ensembles of MRC models based on different pre-trained models show better generalization on out-of-domain set than the ensembles of MRC models based on the same pre-trained models. For example, the ensemble of one BERT-based model and one XLNET-based model has better generalization than the ensemble of one BERT-based models and the ensemble of four XLNET-based models. By incorporating one BERT-based model to our submitted system, the generalization capability of the system is further improved. One possible reason behind this observation is that different pre-trained models are trained on different corpus by designing different pre-training tasks (e.g. masked language model, discourse relations, etc.), and they may capture different aspects of linguistics.

3.2.2 The Effects of Multi-Task Learning

We conduct the experiments to examine the effects of multi-task learning on BERT. Table 4 shows the experimental results:

(1) From the first two rows in Table 4, we can observe that the auxiliary task of masked language model can improve the performance on both in-domain and out-of-domain development set, especially on the out-of-domain set. This means the task of masked language model can help improve the generalization of MRC models on out-of-domain data.

(2) From the last two rows in Table 4, we do not observe that the auxiliary tasks of natural language inference and paragraph ranking bring further benefits in terms of generalization. Although paragraph ranking brings better performance on the in-domain development set, it performs worse on the out-of-domain development set. This ob-

[8] When fine-tuning XLNET, we use layer-wise learning rate decay.

[9] Please refer to the official evaluation results on test set for the details: https://docs.google.com/spreadsheets/d/1vE-uK4aUKqSnTyflwCrE9R9XP_J2Is2uN72tcGPKeSM

[10] NVIDIA TITAN Xp

Models	Dev In-domain F1	Dev Out-of-domain F1
BERT (M_0)	82.40	66.35
BERT + MLM (M_1)	83.19	**67.45**
BERT + MLM, + MNLI (M_2)	83.15	66.92
BERT + MLM, + ParaRank (M_5)	**83.51**	66.83

Table 4: The experimental results on examining the effects of multi-task learning. Please refer to Table 2 with corresponding model ID for details about the model configurations.

servation is different from the previous work (Tan et al., 2017; Wang et al., 2018; Clark et al., 2019; Liu et al., 2019) that multi-task learning can improve the system performance. One possible reason might be the size of MRQA training data is large. Hence, the auxiliary tasks do not bring further advantages in terms of learning more robust language representations from more supervised data.

3.2.3 Summary

In a summary, we have the following major observations about generalization in our experiments: (1) The pre-trained models are still the most important keys to improve the generalization of MRC models in our experiments. The ensemble of MRC models based on different pre-trained models can improve the generalization of MRC models. (2) The auxiliary task of masked language model can help improve the generalization of MRC models. (3) We do not observe much improvements from the auxiliary tasks of natural language inference and paragraph ranking.

3.3 Analysis

In this section, we try to examine that what properties may affect the generalization capability of the submitted system. Specifically, we analyze the performance of the submitted system on different subsets of the testing set. Since the testing set differs from the training set in terms of document sources (see Table 1), we divide the testing set into two subsets: (1) Wiki & Web & News and (2) Other. Please refer to Table 5 for the detailed partition. The document source of the first subset is similar to the training set and we expect that the system works better on the first subset. However, we observe from Table 5 that the system performs similarly on two subsets. The difference on document sources does not bring too much difference on generalization.

We also divide the testing set into three subsets by the requirement of language understanding ability: (1) Matching, (2) Reasoning and (3) Arithmetic. Please refer to Table 6 for the detailed partition. Since most of the questions in the training set (except HotpotQA) require only matching but less reasoning, we expect that the system performs better on the first subset. From Table 6, we observe that the system performs much worse on the the subsets of Reasoning and Arithmetic. Another reason might be that the current models are not well designed for reasoning or arithmetic. Hence, they perform worse on these subsets.

Doc Source	Avg. F1
Wiki & Web & News	72.36
Other	72.60

Table 5: The performance of the submitted system on two subsets that contain different document sources. The two subsets are as follows: (1) Wiki & Web & News: DROP, RelationExtraction, ComplexWebQuestions, QAMR, TREC and (2) Other: BioASQ, DuoRC, RACE, Textbook, BioProcess, MCTest.

Language Understanding	Avg. F1
Matching	79.22
Reasoning	68.73
Arithmetic	61.53

Table 6: The performance of the submitted system on three subsets that require different language understanding ability. The three subsets are as follows: (1) Matching: BioASQ, RelationExtraction, QAMR, QAST, TREC; (2) Reasoning: DuoRC, RACE, Textbook, BioProcess, ComplexWebQuestions, MCTest and (3) Arithmetic: DROP.

4 Conclusions

In this paper, we describe a simple baseline system that Baidu submitted for the MRQA 2019 Shared Task. Our system is built on a framework of pre-training and fine-tuning, namely D-NET. D-NET employs the techniques of pre-trained language models and multi-task learning to improve the generalization of MRC models and we conduct the experiments to examine the effectiveness of these strategies.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic qa corpora generation with roundtrip consistency. In *arXiv preprint arXiv:1906.05416*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. In *arXiv preprint arXiv:1611.09268*.

Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. 2019. An embarrassingly simple approach for transfer learning from pretrained language models. 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 2089–2095.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *arXiv preprint arXiv:1905.10044*.

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.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. In *arXiv preprint arXiv:1704.05179*.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2018. Dureader: a chinese machine reading comprehension dataset from real-world applications. In *In Proceedings of Machine Reading for Question Answering (MRQA) Workshop at ACL. 2018.*, page 37.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *arXiv preprint arXiv:1705.03551*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2018. Stochastic answer networks for machine reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1694–1704.

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.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. In *arXiv preprint arXiv:1611.01603*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019. Ernie 2.0: A continual pre-training framework for language understanding. In *arXiv preprint arXiv:1907.12412*.

Alon Talmor and Jonathan Berant. 2019. Multiqa: An empirical investigation of generalization and transfer in reading comprehension. In *arXiv preprint arXiv:1905.13453*.

Chuanqi Tan, Furu Wei, Nan Yang, Bowen Du, Weifeng Lv, and Ming Zhou. 2017. S-net: From answer extraction to answer generation for machine reading comprehension. In *arXiv preprint arXiv:1706.04815*.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. In *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.

Yizhong Wang, Kai Liu, Jing Liu, Wei He, Yajuan Lyu, Hua Wu, Sujian Li, and Haifeng Wang. 2018. Multi-passage machine reading comprehension with cross-passage answer verification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1918–1927.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. In *arXiv preprint arXiv:1704.05426*.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. In *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. In *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. In *arXiv preprint arXiv:1804.09541*.