Domain-agnostic Question-Answering with Adversarial Training

Seanie Lee[*1], **Donggyu Kim**[*1], **Jangwon Park**[*2]

[1]42Maru, Seoul, Korea
[2]Samsung Research, Seoul, Korea
{lsnfamily02,donggyukimc}@42maru.ai, jang1.park@samsung.com

Abstract

Adapting models to new domain without fine-tuning is a challenging problem in deep learning. In this paper, we utilize an adversarial training framework for domain generalization in Question Answering (QA) task. Our model consists of a conventional QA model and a discriminator. The training is performed in the adversarial manner, where the two models constantly compete, so that QA model can learn domain-invariant features. We apply this approach in MRQA Shared Task 2019 and show better performance compared to the baseline model.

1 Introduction

Followed by the success of deep learning in various tasks, it becomes important to build a single model covering various domains without further fine-tuning to out-of-domain distribution. Because for real world application, a model is required to generalize to unseen sources of data.

In case of Question Answering (QA) task which is one of the promising areas in NLP, however, models outperforming human on SQuAD (Rajpurkar et al., 2016) cannot generalize well to other datasets. Models rather overfit to a specific dataset and require additional training on other dataset to adapt to new domain (Yogatama et al., 2019).

Thus, in order to build a domain-agnostic QA model which is capable of handling out-of-domain data, it is necessary for model to learn domain-invariant features rather than specific ones. In this paper, we apply adversarial training framework to train a QA model with domain-agnostic representation. As shown in Figure 1, the model is divided into two components, which are the QA model and the domain discriminator. The discrim-

inator predicts domain label of hidden representation from QA model. During the training, the QA model tries to fool the discriminator so that the hidden representation becomes indistinguishable to the discriminator. Meanwhile the discriminator is trained to identify the domain label correctly. As a result, QA model can learn domain-invariant features. Our framework can be applied to any existing QA model because the architecture of QA model stays unchanged.

We train and validate our method on 12 datasets (6 datasets for training and 6 datasets for validation) which are provided by MRQA Shared Task. Each training dataset is considered different domain for adversarial learning in which QA model learns domain-invariant feature representation by competing with discriminator. Our experimental result shows that the proposed method improves performance compared to baseline.

2 Related Works

Pre-trained Language Model Recently, there have been several applications for using pre-trained language models, such as ELMo (Peters et al., 2018), GPT (Radford et al., 2018), or BERT (Devlin et al., 2018) to transfer the knowledge from pre-training to various downstream NLP tasks.

BERT is pretrained with bidirectional encoder (Vaswani et al., 2017) on large corpora. Unlike other auto-regressive language models (unidirectional or concatenation of forward and backward language model), BERT randomly masks some input tokens and predicts the masked tokens based on its context. The masked language model enables bidirectional representation, which leads to significant improvements on a number of NLP tasks, such as sentence classification, POS tagging or question answering.
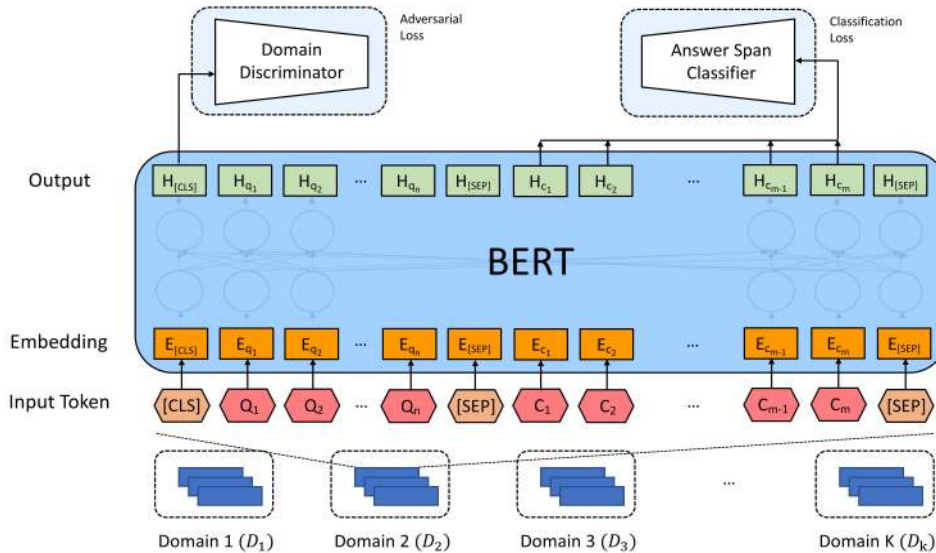
---

*Equal contribution

Figure 1: Overall training procedure for learning domain-invariant feature representation. Model learns to predict start and end position in the passage and fool discriminator for domain-invariant representation.

Domain Generalization Even though many deep learning models surpass human-level performance on various task, they perform poorly on out-of-domain dataset. To address this problem, domain adaptation and domain generalization are proposed, making models more robust to out-of-domain data. The difference between domain adaptation and domain generalization is that for domain generalization, data from the target domain is not available during training.

Several methods for domain generalization exist. One of them is to train a model for each in-domain dataset. When testing on out-of-domain, select the most correlated in-domain dataset and use that model for inference (Xu et al., 2014). Other works such as (Ghifary et al., 2015; Muandet et al., 2013), model is trained to learn a domain-invariant feature by using multi-view autoencoders and mean map embedding-based techniques.

Other approaches (Khosla et al., 2012; Li et al., 2017) break down parameters of a model into domain-specific and domain-agnostic components during training with in-domain dataset, and use the domain invariant parameters for predicting data from unseen target domain.

Recently, meta-learning has been proposed for domain generalization. Some methods (Li et al., 2018a; Balaji et al., 2018; Li et al., 2019) leverage meta-learning framework for domain generalization.

Adversarial Training The idea of adversarial training is originally proposed in the field of image generation (Goodfellow et al., 2014), known as Generative Adversarial Network (GAN). GAN is also adopted in text generation (Yu et al., 2017) with policy gradient for bypassing non-differentiable operation. The concept of adversarial training is not limited to the task of generation. It can be extended to text classification (Chen et al., 2016; Liu et al., 2017; Chen and Cardie, 2018), and relation extraction (Wu et al., 2017). Likewise, attempts are made to get language-invariant features with adversarial training (Chen et al., 2016; Zhang et al., 2017).

Adversarial training has been used for domain adaptation or domain generalization as well. In Domain-Adversarial Neural Network (DANN) (Ganin et al., 2016), it has two classifiers: one classifies task-specific class labels, and the other classifies whether the data belong to source or target domain. Recently, One approach (Li et al., 2018b) extends adversarial autoencoder by minimizing maximum mean discrepancy among different domains for domain-invariant feature representation.

3 Proposed Methodology

We assume that there exists domain invariant feature representation such that QA model generalize well to predict answer on unseen out-of-domain. In order to adapt to out-of-domain, adversarial learning procedure is leveraged for learning

Datasets	Samples	Avg.Q.len	Avg.P.len	Source
BioASQ (BA)	1,504	16.4	353.9	Bio-medical literature
DROP (DP)	1,503	12.0	268.4	Wiki + National Football League (NFL) game summaries and history articles
DuoRC (DR)	1,501	9.8	798.9	Wiki + IMDb
RACE (RA)	674	12.4	381.0	English exams for Chinese middle and high school
RelationExtraction (RE)	2,948	11.6	38.0	Wiki (WikiReading dataset)
TextbookQA (TQ)	1,503	12.1	751.0	1k lessons and 26k multi-modal questions, from middle school science curriculum

Table 1: Statistics of out-of-domain validation dataset. **Q** and **P** stands for question and passage, respectively. Length is calculated based on word-level token.

domain-invariant representation. We present our proposed method in detail in the following sections.

3.1 Problem Definition

We formulate the task as follows: given the K in-domain datasets $\mathcal{D}_i$, consisting of triplets of passage $\mathbf{c}$, question $\mathbf{q}$, and answer $\mathbf{y}$, where $\mathcal{D}_i = \{\mathbf{c}_i^{(k)}, \mathbf{q}_i^{(k)}, \mathbf{y}_i^{(k)}\}_{i=1}^{N_k}$. The model learned from $\{\mathcal{D}_i\}_{i=1}^{K}$ predicts answer $\mathbf{y}_j^l$ from $\mathbf{c}_j^l, \mathbf{q}_j^l$ for each L out-of-domain datasets $\{\mathcal{D}_j\}_{j=1}^{L}$.

3.2 Prediction Model

Our method can be applied to any QA models which learn representation in the joint embedding space of passage and question. In this paper, we use BERT for QA because it is pre-trained on a large corpus and known to be generalized on several different tasks. As for standard QA task, the model is trained to minimize negative log-likelihood of answer $\mathbf{y}$ for all the given in-domain datasets, where $N, \mathbf{y}_{i,s}$, and $\mathbf{y}_{i,e}$ are respectively the total number of in-domain data, the start position and the end position of answer in the passage.

$$\mathcal{L}_{QA} = -\frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \big[\log P_\theta(\mathbf{y}_{i,s}^{(k)}|\mathbf{x}_i^{(k)}, \mathbf{q}_i^{(k)}) + \\ \log P_\theta(\mathbf{y}_{i,e}^{(k)}|\mathbf{x}_i^{(k)}, \mathbf{q}_i^{(k)}) \big] \quad (1)$$

3.3 Adversarial Training

Minimizing the cross-entropy as in equation (1) does not ensure that the model will generalize on unseen domain. Rather it tends to overfit to certain datasets. Inspired by GAN (Goodfellow et al., 2014), we propose a simple yet effective method

to regularize the model such that it learns domain-invariant features.

In the adversarial training procedure, QA model learns to make the discriminator to be uncertain about its prediction. On the other hand, the discriminator is trained to classify the joint embedding of question and passage from QA model into the given K domains. If the QA model can project question and passage into an embedding space where the discriminator cannot tell the difference between embeddings from different K domains, we assume the QA model learns domain-invariant feature representation.

We formulate the adversarial training as follows. A discriminator $\boldsymbol{D}$ is trained to minimize the cross-entropy loss as of equation (2), where l is domain category and $\mathbf{h} \in \mathbb{R}^d$ is the hidden representation of both question and passage. In our experiment, we use [CLS] token representation from BERT for $\mathbf{h}$.

$$\mathcal{L}_{\boldsymbol{D}} = -\frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \log P_\phi(l_i^{(k)}|\mathbf{h}_i^{(k)}) \quad (2)$$

For the QA model, it tries to maximize the entropy of $P_\phi(l_i^{(k)}|h_i^{(k)})$. In other words, it minimizes Kullback-Leibler (KL) divergence between uniform distribution over K classes denoted as $\mathcal{U}(l)$ and the discriminator's prediction as in equation (3). Then the final loss for QA model is $\mathcal{L}_{QA} + \lambda \mathcal{L}_{adv}$ where λ is a hyper-parameter for controlling the importance of the adversarial loss. In our experiments, we alternate between optimiz-

Model	BA		DP		DR		RA		RE		TQ		Avg	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Bert-base	**46.44**	**60.81**	28.31	37.88	42.78	53.32	**28.23**	39.51	**73.33**	**83.89**	44.30	52.03	43.90	54.57
Bert-base-adv	43.35	60.04	**30.51**	**40.01**	**45.97**	**57.89**	26.50	**39.73**	72.67	83.53	**45.62**	**55.67**	**44.10**	**56.15**

Model	BP		CQ		MC		MR		ST		TR		Avg	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Bert-base	38.36	57.38	47.40	55.29	54.16	66.12	47.83	64.81	**58.64**	**77.02**	36.73	53.96	47.19	62.43
Bert-base-adv	**42.92**	**61.09**	**48.13**	**56.50**	**55.83**	**69.30**	**52.82**	**68.78**	52.73	75.63	**39.08**	**56.79**	**48.59**	**64.68**

Table 2: Model performance on validation and test set. Above is the validation set and below is the test set.

ing QA model and discriminator.

$$\mathcal{L}_{adv} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N_k} KL(\mathcal{U}(l) \parallel P_\phi(l_i^{(k)}|\mathbf{h}_i^{(k)})) \quad (3)$$

4 Experiments and Result

4.1 Dataset

We validate our adversarial model for MRQA Shared Task with 6 different out-of-domain datasets, which are BioASQ (BA) (Tsatsaronis et al., 2012), DROP (DP) (Dua et al., 2019), DuoRC (DR) (Saha et al., 2018), RACE (RA) (Lai et al., 2017), RelationExtraction (RE) (Levy et al., 2017), and TextbookQA (TQ) (Kembhavi et al., 2017). Table 1 shows the statistics and description of these datasets. Each dataset has about 1k samples. However, the number of samples from each dataset varies. Thus, we use stratified sampling in order to make class-balanced stochastic mini-batch having certain amount of samples from all domains. We use maximum sequence length of 64 and 384 for question and passage respectively. But some examples are longer than 384. Therefore each passage is split into several chunks with a window size of 128. We discard samples without answers because all questions are considered to be answerable from given context in MRQA shared task.

Note that the final evaluation shown in the Table 2 is conducted by MRQA organizers with additional 6 out-of-domain undisclosed private test datasets, which are BioProcess (BP) (Scaria et al., 2013), ComplexWebQuestion (CQ) (Talmor and Berant, 2018), MCTest (MC) (Richardson et al., 2013), QAMR (MR) (Michael et al., 2017), QAST (ST) (Jitkrittum et al., 2009) and TREC (TR) (Voorhees, 2001).

4.2 Implementation Details

We implement our model based on the Hugging-Face's open-source BERT implementation[1] in Pytorch (Paszke et al., 2017). The performance of the baseline in our experiment differs from the official baseline of MRQA, which is based on AllenNLP (Gardner et al., 2018). We follow the hyperparameters as BERT for our model. In detail, we use "bert-base-uncased" with a learning rate 3e-5 and a batch size of 64. Additionally, our model requires one more hyperparameter λ, which indicates the importance of adversarial loss as described in the equation (3). We find out that the value of 1e-2 for λ gives the best result in our experiments. The baseline and adversarial model are trained on V100 GPU for about 5 GPU hours. For training, we use 6 in-domain datasets, which are SQuAD, TriviaQA (Joshi et al., 2017), Natural Questions (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), SearchQA (Dunn et al., 2017), and NewsQA (Trischler et al., 2016) provided by MRQA. We select the best performing model on validation set, where models are trained for 1 or 2 epochs. The codes for our model are available at https://github.com/seanie12/mrqa.

4.3 Performance Comparison

Table 2 shows the performance evaluation results of models on out-of-domain datasets. In the table, the model trained with our adversarial learning is named with '-adv'. The top of the table is the result of validation datasets while the bottom is the result of test datasets. As shown in the table, overall, the model with adversarial learning has better performance compared to the baseline in terms of both EM and F1 measures.

For validation datasets, the average F1 score of our model is about 1.5 point higher than the baseline. In detail, our model outperforms the baseline

---
[1]https://github.com/huggingface/transformers

in DP, DR, RC, and RA dataset by large margin. But the adversarial learning degrades performance in BA and RE. We can see the same aspect in terms of EM score. Similar to the result of validation datasets, our model shows better performance in terms of EM (Exact Match) and F1 on the most of test datasets except for ST. Overall, our model has superior performance with considerable margin of over 2 point in F1.

5 Discussion

In this section, we discuss some trials that have failed to improve the performance but might be helpful for future works.

5.1 Span Refinement

QA sample consists of a question, a passage, and an answer span. There could exist multiple answer spans because more than one phrase in the passage can be matched with the answer text. For simplicity, only the first occurrence of answer text is used for training in most of the baseline codes. However, considering context and semantic of the given question and answer, a certain phrase in the passage is more likely to be plausible answer span relevant to the question. In order to find the most plausible answer span, a question and sentences in the passage are encoded into fixed-size vectors with universal sentence encoder (Cer et al., 2018). We choose the span in a sentence, which is the most similar to the question in terms of cosine similarity, as golden span. In our experiment, this approach boosts up the performance of some datasets but degrades the performance a lot in the other datasets.

5.2 Meta Learning

We apply meta learning to domain generalization (Li et al., 2018a, 2019; Balaji et al., 2018) to simulate train/test domain shift. For every epoch, one dataset is randomly selected as virtual test domain. As described in (Finn et al., 2017), QA model is trained to maximize meta objective, which leads to improve the performance in train domain, but also in test domain. But this requires to compute Hessian-vector products, which slows down the training. This is even worse for BERT because there are 110M parameters to fine-tune. Moreover, contrary to the previous works, the meta learning for domain generalization does not help improve the performance.

6 Conclusion

We leverage adversarial learning to learn domain-invariant features. In our experiments, the proposed method consistently improves the performance of baseline and it is applicable to any QA model. In future work, we will try adversarial learning for pre-training model with diverse set of domains.

Acknowledgments

References

Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. 2018. Metareg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems*, pages 998–1008.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Xilun Chen, Ben Athiwaratkun, Yu Sun, Kilian Q. Weinberger, and Claire Cardie. 2016. Adversarial deep averaging networks for cross-lingual sentiment classification. *CoRR*, abs/1606.01614.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1226–1240, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

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. *arXiv preprint arXiv:1704.05179*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. 2015. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Wittawat Jitkrittum, Choochart Haruechaiyasak, and Thanaruk Theeramunkong. 2009. QAST: Question answering system for ThaiWikipedia. In *Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions (KRAQ 2009)*, pages 11–14, Suntec, Singapore. Association for Computational Linguistics.

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

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4999–5007.

Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. 2012. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. *arXiv preprint arXiv:1706.04115*.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5542–5550.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2018a. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. 2018b. Domain generalization with adversarial feature learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yiying Li, Yongxin Yang, Wei Zhou, and Timothy M Hospedales. 2019. Feature-critic networks for heterogeneous domain generalization. *arXiv preprint arXiv:1901.11448*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *CoRR*, abs/1704.05742.

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. 2017. Crowdsourcing question-answer meaning representations. *arXiv preprint arXiv:1711.05885*.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18.

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.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language understanding paper.pdf*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203.

Amrita Saha, Rahul Aralikatte, Mitesh M Khapra, and Karthik Sankaranarayanan. 2018. Duorc: Towards complex language understanding with paraphrased reading comprehension. *arXiv preprint arXiv:1804.07927*.

Aju Thalappillil Scaria, Jonathan Berant, Mengqiu Wang, Peter Clark, Justin Lewis, Brittany Harding, and Christopher D Manning. 2013. Learning biological processes with global constraints. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1710–1720.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. *arXiv preprint arXiv:1803.06643*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. 2012. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *2012 AAAI Fall Symposium Series*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ellen M Voorhees. 2001. The trec question answering track. *Natural Language Engineering*, 7(4):361–378.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783, Copenhagen, Denmark. Association for Computational Linguistics.

Zheng Xu, Wen Li, Li Niu, and Dong Xu. 2014. Exploiting low-rank structure from latent domains for domain generalization. In *European Conference on Computer Vision*, pages 628–643. Springer.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada. Association for Computational Linguistics.