Domain Adversarial Training for Robust QA

Stanford CS224N {Default} Project
Track: {RobustQA}

Abhay Singhal
Department of Computer Science
Stanford University
sabhay@stanford.edu

Navami Jain
Department of Computer Science
Stanford University
navajain@stanford.edu

Shayana Venukanthan
Department of Computer Science
Stanford University
shayanav@stanford.edu

Abstract

The goal of this research is to understand the optimal implementation of domain adversarial training for robust question answering. While past studies have implemented adversarial training for this problem, we postulate that the model may benefit from learning both domain-invariant and domain-specific features. We investigate whether partially rather than perfectly domain-independent features in domain adversarial training produces Q&A models with high accuracy while still remaining robust to domain shifts. We create partially independent models by only feeding part of the Q&A model's representation into the discriminator. We also explored further model improvements leveraging focal loss to address class imbalance issues in our training data and Wasserstein distance as an alternative measure to adversarially train the discriminator. We find that developing features with partial domain independence successfully improves the model's performance on unseen data. Our model achieves EM = 41.95 and F1 = 60.096 scores on the OOD test set.

1 Key Information to include

- TA mentor: Yian Zhang
- We have no external collaborators or mentors and are not using the project for any other class.

2 Introduction

Large language models such as BERT [1] have shown to achieve close to human-like performance and understanding when fine-tuned on large datasets. However, when fine-tuned only on small or low-resource data, they perform significantly worse. Since the distribution of training data is rarely, if ever, seen in the real world, the failure to generalize poses a key problem to these systems achieving human-like performance. Moreover, increasing evidence shows that this failure to generalize understanding to out-domain or low-resource data is likely due models learning brittle, distribution-specific correlations [2].

One of the best proxies for understanding, both for humans and language models, is the ability to answer questions. Hence, we focus on improving the question answering models to generalize like humans. Specifically, we build a cloze Q&A built on DistilBERT [3] that takes as input (context, question) pairs and outputs answers to the question based on the context and adapts to unseen

domains with only a few training examples.

As such, we tackle the issue of brittleness and domain-dependence in the features derived by DistilBERT. To do so, we employ Domain Adversarial Training, which is further described in our Approach, which has been shown to improve out-of-domain generalization [4]. Building on their work, we show that partial domain independence, i.e. a mixture of domain-invariant and domain-specific features as described in our Approach, leads to improvement.

Moreover, we explore Wasserstein distance in lieu of KL divergence used by Lee et al. and find that it improves performance.

Lastly, to address the class imbalance between in-domain and out-of-domain training examples, in lieu of cross-entropy loss, we explore Focal Loss, a loss that varies weighting according to classification error by class, for the discriminator [5].

3 Related Work

This section helps the reader understand the research context of your work, by providing an overview of existing work in the area.

Adapting models to new domains without finetuning is a challenging problem in deep learning. One way to improve a model's out-of-domain performance is through domain adversarial training. In their paper, "Domain-agnostic Question-Answering with Adversarial Training", Lee, et. al., 2019 [4] utilize an adversarial training framework for domain generalization in Question Answering (QA) tasks. In order to produce domain-invariant features, Lee, et. al. implement an adversarial training approach where in addition to the classifier component (the QA model), a discriminator is tasked to identify the domain of the data, and the classification component tries to extract features that maximize the loss of the discriminator by creating indistinguishable hidden representations. If the QA model can project the question and passage into an embedding space where the discriminator cannot tell the difference between embeddings from different domains, they assume the QA model learns domain-invariant feature representation [4].

We reference this paper as our main approach to improve the performance of DistilBERT model on the RobustQA task: in addition to our baseline DistilBERT model for QA, we implemented our own original discriminator for domain identification, and trained both with domain adversarial training.

However, in some datasets in both validation and testing, the performance of the model is degraded by Lee et. al's proposed method [4]. However, the paper offers no potential reasons for this poorer performance, and so we do not know how this could extend to other datasets or language models. The inconsistent improvement seen across domains suggests that completely domain-independent representations may not provide the best results. We hence explore partial domain independence by holding out certain features from the discriminator. As such, while some features can be trained to be domain-invariant, others can be trained to recognize the domain and task and appropriately select domain-specific features that improve prediction.

Furthermore, we note the similarity of the DANN (Domain Adversarial Neural Network) approach proposed by Lee et al. to a GAN, wherein the DistilBERT featurizer is analogous to a generator. Following the advancements in GANs, we apply Arjovsky et al's Wasserstein GAN [6] algorithm to a DANN. Specifically, Arjovsky et al. propose replacing the KL divergence used in a GAN with Earth Mover distance, i.e. Wasserstein-1, and show improved training stability, reduced mode collapse, and higher quality outputs. We apply it to this supervised context by replacing KL divergence with Wasserstein-1 distance and also find improved generalization.

Finally, given the classification problem introduced by the DANN, the problem of class imbalance becomes pertinent as out-of-domain examples are vastly underrepresented (by definition). Lin et

al. show Focal Loss [5] addresses the class imbalance problem by applying a modulating term to cross-entropy loss to focus the loss function on hard, misclassified examples. Given class imbalance, this is likely to be rarer classes and hence focal loss weights classification of all classes more equally. Since we face significant class imbalance, we explore the replacement of cross-entropy loss with focal loss for the discriminator.

4 Approach

This section details your approach(es) to the problem. For example, this is where you describe the architecture of your neural network(s), and any other key methods or algorithms.

4.1 Baseline

As described in the default final project handout, the baseline model used was a DistilBERT QA model (named for being a "distilled" version of the BERT model). Preprocessing and chunking are done as described.

4.2 Domain Adversarial Training

We implemented our QA model with domain adversarial training using a very similar approach to Lee, et. al., 2019 [4].

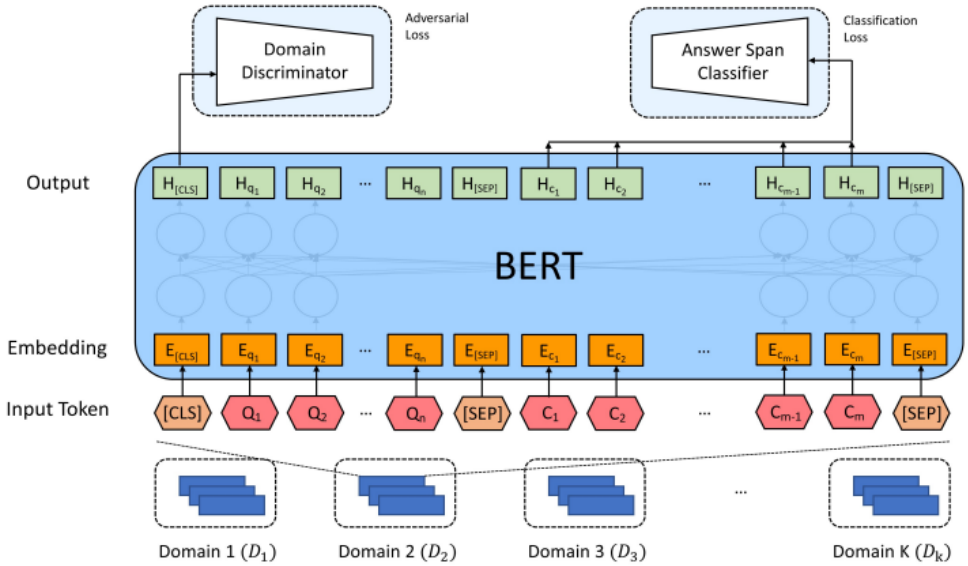


Figure 1: Training procedure for learning domain-invariant feature representations. The discriminator is trained to predict domain of the dataset based on the output [CLS] token. The model classifier predicts the appropriate answer while fooling the discriminator. Taken from Lee et. al 2019. [4]

We added a discriminator D which can be mathematically be expressed as

$$D = W_3 * (ReLU(W_2 * (ReLU(W_1 * h + b_1)) + b_2)) + b_3,$$

where h is the `[CLS]` token from the last hidden layer W_1, W_2, and W_3 are weight matrices and b_1, b_2, and b_3 are bias terms for the linear layers.

D trains to minimize the multi-class cross-entropy loss, L_D, of domain category prediction given h, the first hidden representation of both question and passage (which can also be thought of as the extracted features). We then added L_{adv}, which we first set as the Kullback-Leibler (KL) divergence between uniform distribution over K classes and the discriminator's prediction, to the QA model's

loss so that the QA model trains to confuse the discriminator and hopefully develops domain-invariant features. KL divergence, like cross-entropy loss, is a tool to measure the difference between two distributions. The final loss for QA model is $L_{QA} + \lambda L_{adv}$ where λ is a hyper-parameter controlling the importance of the adversarial loss. In our experiments, we optimize both the QA model and discriminator at the same time. We experiment to find the optimal λ with L_{adv} using KL divergence and L_D using cross-entropy loss.

4.3 Partial Domain Independence

Based on results from existing literature, including Lee, et. al., 2019 [4] and Zhu, 2021 [7], we postulate that the model may benefit from learning both domain-invariant and domain-specific features. To investigate whether partially rather than perfectly domain-independent features in domain adversarial training produces Q&A models with high accuracy while still remaining robust to domain shifts, we create partially independent models by only feeding part of the Q&A model's representation h into the discriminator. We control this with hyperparameter β, which dictates the proportion of features of h that are trained to be domain-invariant, while $1 - \beta$ of h's features remain domain-specific. Our entire ideation and implementation of partial domain independence is original. We experiment to find optimal β, assuming we have optimal λ.

4.4 Focal Loss

Our training datasets have a large class imbalance between in-domain and out-of-domain classes, as we have 50,000 training samples for each in-domain dataset and only 127 training samples for each out-of-domain dataset. In order to address this imbalance, we refer to Lin, et. al., 2018's [5] work and replace our QA model's cross-entropy loss with Lin, et. al.'s focal loss implementation [8]. Starting from cross entropy (CE) loss for binary classification:

$$\text{CE}(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{otherwise.} \end{cases}$$

for notational convenience, they define p_t:

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise,} \end{cases}$$

and rewrite CE(p, y) = CE(p_t) $= \log(p_t)$. They then introduce a weighting factor α_t and a modulating factor $(1 - p_t)^\gamma$ to the cross entropy loss to get the following equation for focal loss (FE):

$$FE(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t).$$

Focal loss allows the loss function to apply more focus on difficult, misclassified examples. We build off of Lin, et. al.'s results to find optimal α and γ, assuming we have optimal λ and β.

4.5 Wasserstein Distance

We explored model improvements by replacing the Kullback-Leiber (KL) divergence in the L_{adv} term of our loss expression with a Wasserstein distance measure to adversarially train the discriminator function. At a high level, the Wasserstein distance is an earth-mover distance metric between two probability distributions, defined as:

$$\mathbb{W}(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}[||x - y||]$$

where $\Pi(P_r, P_g)$ is the set of all joint distributions over x and y such that the marginal distributions are equal to P_r and P_g.

In this case, the predicted domain from the discriminator is representative of the source domain and a uniform distribution is the target domain. We leveraged an existing implementation [9] of calculating

Wasserstein distance and integrated it into our loss function. Similar to how it operates in W-GANs, we expected Wasserstein distance to provide more information in the training of the discriminator and 'generator'. It also had the advantage of being able to operate on the logits directly.

5 Experiments

5.1 Data

We use three in-domain datasets: the Stanford Question Answering Dataset (SQuAD) [10], Natural Questions [11], and NewsQA [12], all pre-processed in the same format as SQuAD. In addition, we use three out-of-domain datasets: DuoRC [13], RACE [14], and RelationExtraction [15].

Dataset	Question Source	Passage Source	Train	dev	Test
		`in-domain datasets`			
SQuAD [5]	Crowdsourced	Wikipedia	50000	10,507	-
NewsQA [7]	Crowdsourced	News articles	50000	4,212	-
Natural Questions [6]	Search logs	Wikipedia	50000	12,836	-
		`oo-domain datasets`			
DuoRC [9]	Crowdsourced	Movie reviews	127	126	1248
RACE [10]	Teachers	Examinations	127	128	419
RelationExtraction [11]	Synthetic	Wikipedia	127	128	2693

Please refer to the default project handout for the Robust QA track [2] for further detail.

5.2 Evaluation method

The primary evaluation metrics we used for the milestone were EM and F1 score. EM stands for 'Exact Match' and, as the term alludes, measures whether a model's prediction exactly matches the characters of the true answer. The reported EM is the average over individual example scores. F1 is a more common metric for classification problems–the harmonic mean of precision and recall. In contrast to exact match (an all-or-nothing approach), the precision and recall are calculated based on the number of shared words in the prediction compared to the ground truth.

5.3 Experimental details

We began by training the provided baseline model using the default hyperparameters for the classification component, namely with 3 epochs, a constant learning rate of 3e-05, and batch size of 16. Learning rate and batch size were kept constant throughout our experiments.

Following this, we added a discriminator component that takes the first hidden layer of the model and predicts domain class. The discriminator was trained on a multi-class cross entropy loss function. Domains of datasets were one-hot encoded based on their type (Wikipedia, News Articles, Movie Reviews, Examinations). We assess three lambda values: $\lambda = .01, .005, .05$, where λ is the hyperparameter controlling the importance of the discriminator's adversarial loss.

We initially used stochastic gradient descent (SGD) as the optimizer of our discriminator, but changed the optimizer to Adam.

Then, we implemented partial domain independence by feeding only part of the Q&A model's representation h into the discriminator. We assess beta values: $\beta = 0.85, 0.90, 0.95, 0.99$, where β is the hyperparameter controlling what proportion of h's features are trained to be domain-invariant.

Then, we began experimenting with focal loss, Wasserstein distance, and our samplers both separately and in combination. We assess various combinations of gamma and alpha values and sampler types:

$\gamma = 1.0, 2.0, 3.0$, $\alpha = 0, 0.25$, sampler = random (default) or weighted. We also experimented with lambda again with values: $\lambda = 0.01, 0.05$.

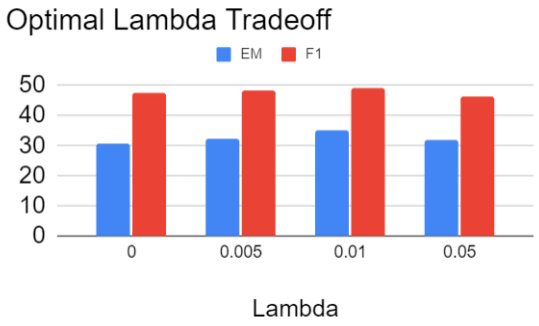
Training and evaluation time was relatively consistent across models. Training typically took between 2-3 hours (for three epochs) and evaluation took less than 5 minutes.

5.4 Results

Unless specified otherwise, the provided EM and F1 scores are for the dev set.

5.4.1 Adversarial Training: Lambda, λ

Model	lambda	EM	F1
Baseline	n/a	30.63	47.72
Adv. training with SGD	0.01	31.152	46.896
Adv. training with Adam	0.01	35.079	49.321
Adv. training with Adam	0.05	31.94	46.40
Adv. training with Adam	0.005	32.46	48.35



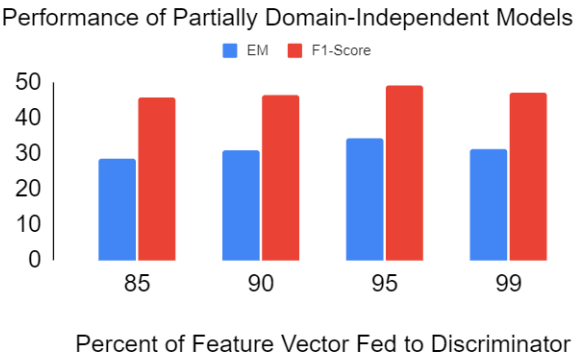
We see an increase in performance on the dev set with the introduction of adversarial training. Interestingly, we see that there's a balance to be maintained for our lambda value, in that giving too great of a significance to the adversarial training loss component of our final loss reduces overall performance on out-of-domain datasets (lambda = 0.05), as does giving too little (lambda = 0.005). We see that the optimal lambda for adversarial training with the Adam optimizer is $\lambda = 0.01$.

For all the following sections and results, the model undergoes domain adversarial training with the Adam optimizer.

5.4.2 Partial Domain-Invariance: Beta, β

For this section, $\lambda = 0.01$

beta	EM	F1
0.85	28.53	45.90
0.90	30.89	46.55
0.95	34.29	49.28
0.99	31.41	47.23



Among the experiments conducted to finetune our beta hyperparameter, we see optimal performance for beta = 0.95 with EM = 34.89 and F1 = 49.28. Unexpectedly, the experiments run with the beta hyperparameter produce slightly lower scores than the previous results without (equivalent to beta=1). However, the fact that our scores are significantly lower for beta = 0.99 than they are for beta=0.95 indicate that our approach with partial domain invariance is still successful.

5.4.3 Focal Loss, Wasserstein, Sampler, Lambda, Beta

QA Model's Loss	Adv. Train Loss	sampler	lambda	beta	gamma	alpha	EM	F1
Focal Loss	KL Divergence	Random	0.01	0.90	0.3	0	30.63	45.69
Focal Loss	KL Divergence	Random	0.01	0.90	1.0	0	30.63	47.49
Focal Loss	KL Divergence	Random	0.01	0.90	2.0	0	31.68	47.33
Focal Loss	KL Divergence	Random	0.01	0.95	2.0	0.25	33.77	48.92
Focal Loss	KL Divergence	Random	0.01	0.90	3.0	0	31.94	47.01
CE Loss	Wasserstein	Random	0.05	1.00	n/a	n/a	32.72	49.24
CE Loss	Wasserstein	Weighted	0.01	0.95	n/a	n/a	31.94	48.49
CE Loss	KL Divergence	Weighted	0.01	0.95	n/a	n/a	29.32	43.78
Focal Loss	**Wasserstein**	**Random**	**0.01**	**0.95**	**2.0**	**0.25**	**35.08**	**51.16**

Our model that combined focal loss and Wasserstein distance with hyperparameters lambda = 0.01, beta = 0.95, gamma = 2.0, alpha = 0.25, and random sampling performed best with scores EM = 35.08 and F1 = 51.16 on the out-of-domain validation set. The early results for models with only focal loss or only Wasserstein were somewhat disappointing considering that focal loss was meant to specifically address class imbalance and Wasserstein was used for its theoretical advantages in domain adaptation: its gradient property and promising generalization bound [16], yet we saw no significant increase in performance. However, we see that combining the two produced improved scores on the dev set. This model achieved scores **EM = 41.789** and **F1 = 60.069** on the out-of-domain test set.

6 Analysis

We can see significant improvement in performance between the DistilBERT baseline and our model with domain adversarial training. For example, let's observe the following example from the RACE set.

Context: "...The teaching arrangement filled me with fear. I was to divide the class of twenty-four boys, aged from seven to thirteen, into three groups and teach them all subjects–including art, football, cricket and so on–in turn at three different levels. Actually, I was depressed at the thought of teaching algebra and geometry–two subjects in which I had been rather weak at school..."
Question: Which subjects was the writer poor at?
Correct Answer: algebra and geometry
DistilBERT: art, football, cricket and so on
Adversarial Model: algebra and geometry

Question: Where must they go to attach the ArcNet?
Correct Answer: Cape Canaveral
DistilBERT: no K
Adversarial Model: Cape Canaveral

Context: "The authorities discover the scientists' project, and arrest them. Miles escapes by disguising himself as a robot, and goes to work as a butler in the house of socialite Luna Schlosser (Diane Keaton). When Luna decides to have his head replaced with something more äesthetically pleasing,M̈iles reveals his true identity to her, whereupon Luna threatens to give Miles to the authorities. In response, he kidnaps her and goes on the run, searching for the Aries Project.and Luna fall in love, but Miles is captured and brainwashed into becoming a complacent member of the society, while Luna joins the rebellion. The rebels kidnap Miles and perform reverse-brainwashing, whereupon he remembers his past and joins their efforts. Miles becomes jealous when he catches Luna kissing the rebel leader, Erno Windt (John Beck), and she tells him that she believes in free love."

Question: Who catches onto the scientists' project?
Correct: authorities
DistilBERT: Miles becomes jealous when he catches Luna
Adversarial Model: The authorities

We see that our adversarial model is able to correctly identify the subjects with which the writer struggles while the baseline model simply returns some subjects it identifies. This is a recurring trend in that the baseline tends to return some answer that matches the type of the answer but fails to identify the correct answer. The baseline also tends to fail to remove unnecessary punctuation. The adversarial model's ability to generalize seems to lead to greater accuracy in identifying specific person, place, or object.

7 Conclusion

In summary, we propose an optimal implementation of domain adversarial training for robust question answering. This implementation uses a DANN [4] with partial domain independence for the features and Wasserstein distance and focal loss used for the discriminator with the following hyperparameters $\lambda = 0.01$, $\alpha = 0.25$, $\beta = 0.95$, and $\gamma = 2.0$. Compared to our baseline model trained without an adversarial component, adding the discriminator improved performance in terms of F1-Score and Exact Match (EM). Developing features with partial domain independence also improved the model's performance on unseen data.

While our dataset was heavily imbalanced, it remains unclear whether focal loss improved overall performance. Due to time limitations, we were unable to fully explore the utility of the alpha parameter, for example, which could have resulted in significant performance changes.

While several combinations of hyperparameters were tested, a more extensive and organized hyperparameter search needs to be conducted to make conclusions on the utility of Wasserstein distance and focal loss. Future directions of this work can include experimenting with various training times, iterations, and learning rates on this same model and further finetune our hyperparameters. In addition, we could explore other current methods of improving out-of-domain performance, including mixure-of-experts, few shot learning, meta learning.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 2019. Association for Computational Linguistics.

[2] Cs 224n default final project: Building a qa system (robust qa track). 2022.

[3] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019*, 2019.

[4] Seanie Lee, Donggyu Kim, and Jangwon Park. Domain-agnostic question-answering with adversarial training. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, 2019.

[5] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. Facebook AI Research (FAIR), 2018.

[6] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.

[7] Bryan Zhu. Robust question answering using domain adversarial training. 2021.

[8] clcarwin. Focal loss for dense object detection in pytorch. 2017.

[9] Daniel Daza. Approximating wasserstein distances with pytorch. 2019.

[10] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250, 2016.

[11] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

[12] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[13] Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. Duorc: Towards complex language understanding with paraphrased reading comprehension. 2018.

[14] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. 2017.

[15] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. *CoRR*, abs/1706.04115, 2017.

[16] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. Association for the Advancement of Artificial Intelligence, 2018.