Not to Overfit or Underfit the Source Domains?
An Empirical Study of Domain Generalization in Question Answering

Md Arafat Sultan   Avirup Sil   Radu Florian
IBM Research AI
arafat.sultan@ibm.com, {avi, raduf}@us.ibm.com

Abstract

Machine learning models are prone to overfitting their training (source) domains, which is commonly believed to be the reason why they falter in novel target domains. Here we examine the contrasting view that multi-source domain generalization (DG) is first and foremost a problem of mitigating source domain *underfitting:* models not adequately learning the signal already present in their multi-domain training data. Experiments on a reading comprehension DG benchmark show that as a model learns its source domains better—using familiar methods such as knowledge distillation (KD) from a bigger model—its zero-shot out-of-domain utility improves at an even faster pace. Improved source domain learning also demonstrates superior out-of-domain generalization over three popular existing DG approaches that aim to limit overfitting. Our implementation of KD-based domain generalization is available via PrimeQA at: `https://ibm.biz/domain-generalization-with-kd`.

1   Introduction

Domain generalization (DG) seeks to train models on a small number of *source* domains in a way that maximizes their zero-shot out-of-domain (OOD) utility (Blanchard et al., 2011; Muandet et al., 2013). Many existing DG methods are rooted in the premise that weak generalization under domain shift occurs due to "source domain overfitting": the learning of spurious source domain correlations, *i.e.*, noise, that are unrelated to the core learning task, and are therefore unlikely to be present in novel target domains. The popular *domain-invariant learning* (DIL) paradigm, for example, proposes to limit overfitting by imposing some form of cross-domain regularization on the empirical risk minimization (ERM) process over labeled multi-domain training data (Wang et al., 2021; Zhou et al., 2021).

We argue in this paper that this emphasis on *not fitting the noise* minimizes the importance of *actually learning the signal* present in the source domains, *i.e.*, **not underfitting** them, for DG. Underfitting occurs when a trained model, due to inadequacies in its capacity or the training procedure, fails to learn training set patterns that are truly representative of the task. In a recent study of DG in computer vision, Gulrajani and Lopez-Paz (2021) find that large models with sufficiently high capacity can demonstrate strong OOD generalization even with ordinary ERM training—without the need for DIL—when properly configured. Here we ask a very similar question, but (*a*) in the NLP context of question answering (QA), and (*b*) for relatively small models that are more prone to underfitting, *e.g.*, a BERT-*base* QA model (Devlin et al., 2019). Concretely, we investigate if enabling smaller QA models to learn their source domains better also improves their generalization to new domains and how the results compare to DIL.

This formulation essentially views OOD generalization in low-capacity models as a natural extension of the more familiar and arguably simpler problem of in-domain generalization in a multi-domain setting. A key advantage of such an approach is that familiar supervised learning methods such as knowledge distillation (KD) (Hinton et al., 2014)—which was specifically designed for the very purpose of training small yet high-accuracy models—can now be leveraged for DG. KD generally provides stronger source domain supervision than ERM by minimizing a surrogate risk, which utilizes the soft predictions of a larger *teacher* model (*e.g.*, a BERT-*large* QA model) as the learning targets for the smaller model being trained, now called the *student*. Additionally, synthesized input in large quantities has been found to further enhance the performance of KD (Chen et al., 2020; Liang et al., 2021). Here we extend the application of these methods to DG for QA.

3752

We evaluate our methods on a multi-dataset reading comprehension benchmark (Fisch et al., 2019) and compare their accuracy with three popular DIL approaches: domain adversarial training (Ganin et al., 2016; Lee et al., 2019), episodic training (Li et al., 2019)—for which we propose a novel variant suitable for deep transformer models—and meta learning (Finn et al., 2017; Li et al., 2018). We also design experiments to answer more targeted questions such as: (1) Are the improvements more prominent on in-domain validation data than on OOD test instances, which could be indicative of weak generalization? (2) Do the proposed methods falter on input cases where domain-invariant approaches thrive, potentially indicating weakness on extremely distant test cases? In all these different evaluations, our methods exhibit far superior DG than the three existing methods, whereas the latter only marginally outperform ERM.

While more experiments on additional tasks, datasets and baselines are needed before a firm conclusion can be reached on the superiority of the proposed formulation of DG over DIL (or vice versa), our findings do point to a need for a better understanding of optimal source domain learning as an approach to DG. A primary goal of this paper is to motivate future explorations of this important research question.

2 Methodology

This section describes the proposed KD-based DG approaches as we apply them to the problem of reading comprehension.

2.1 The Reading Comprehension Task

Given a question q and a passage p that answers q, reading comprehension (RC) outputs the answer a. In *extractive* RC, which is the form used in our experiments, a is assumed to be a subtext of p; the goal is therefore to locate a in p.

2.2 Multi-Dataset Knowledge Distillation

For improved multi-domain training, we rely on knowledge distillation (KD), which naturally trains small yet highly accurate models by leveraging the predictions of a larger teacher model. We first train a single multi-domain teacher using ERM on labeled data from all source domains. We follow standard RC training procedure for this step, which separately trains an answer start and an end predictor, as described in (Devlin et al., 2019).

The knowledge of this teacher is then distilled into a smaller student model—equal in size with the baselines—by minimizing the following MSE loss on the same set of training examples:

$$\mathcal{L}_{KD} = \|z_s(x) - z_t(x)/\mathcal{T}\|_2^2 \qquad (1)$$

where $z_s(x)$ and $z_t(x)$ are the logits computed for an input question-passage pair x by the student and the teacher, respectively; $\mathcal{T}$ is the temperature. Two separate KD losses are again minimized per training example, one for the start and the other for the end of the answer.

2.3 Augmenting KD with Synthetic Questions

To facilitate the distillation of further knowledge from the teacher, we synthesize additional questions using a sequence-to-sequence model. An encoder-decoder language model (Lewis et al., 2020; Raffel et al., 2020) is first fine-tuned for each source domain, where question-passage pairs from the corresponding dataset constitute the training examples: the passage is the source and the question is the target. Note that we use the teacher's soft answer predictions as targets during KD (Eq. 1), and therefore do not need to provide any answers as part of the synthetic data.

Sultan et al. (2020) demonstrate that sampling-based generation, *e.g.*, with a top-p top-k sampler, produces more useful synthetic training data than deterministic approaches like greedy or beam search decoding. Moreover, Chen et al. (2020) find that (*a*) large amounts of diverse synthetic data can be more effective at supporting KD than typically smaller amounts of human-labeled *gold* examples, and (*b*) leveraging both in a two-stage process yields the best results. We incorporate these suggestions into our work by sampling examples from our generators and performing KD with first synthetic and then gold training data. Unlike those earlier studies, however, we apply the above procedure to the multi-source DG problem.

3 Experiments

Here we describe the baseline DIL approaches our proposed methods are evaluated against, our experimental setup and results.

3.1 Baselines

We select three existing DIL methods as baselines on account of their recency, popularity and general

applicability to different machine learning problems. Each of these methods imposes additional requirements on top of ERM to incorporate domain invariance into the trained models.

For ***domain-adversarial training*** (Ganin et al., 2016), that added requirement is for the model to learn domain-agnostic hidden representations of the training inputs. This is accomplished by training a domain classifier in parallel with the RC model and teaching their shared feature extractor to produce adversarial representations for the domain classifier (Lee et al., 2019).

Given a model with parameters Θ to be optimized on mult-domain training data, ***episodic training*** (Li et al., 2019) updates a random subset Θ' in each iteration; values for the remaining parameters $\Theta \setminus \Theta'$ are copied over from a weaker model pretrained on one of the source domains other than that of the current example. This procedure effectively forces the Θ' subnetwork to become more robust to domain shift as it learns to work with an OOD companion $\Theta \setminus \Theta'$. While Li et al. (2019) use a fixed breakdown of Θ into a feature extractor and a task head, we relax this condition for multilayer transformer models to allow a split after a randomly chosen transformer layer.

Finally, ***meta learning*** (Finn et al., 2017) for DG (MLDG) (Li et al., 2018) uses disjoint subsets of the source domains as *meta-train* and *meta-test* domains at each training step. It uses the meta-train set to update model parameters in a way that improves performance on the meta-test set. This is accomplished using a second-order differentiation through the parameter updates of the model.

3.2 Setup

We run our experiments on the the public subset of the DG benchmark by Fisch et al. (2019)[1]. It consists of (*a*) training and in-domain validation data from six source datasets, and (*b*) six target datasets for evaluation. Table 1 shows some key statistics. We refer the reader to the original paper for a detailed description of each dataset.

Given our stated intent to study DG with relatively small models (§1), we fine-tune BERT-*base* (110M parameters) (Devlin et al., 2019) for RC with the different training methods (Wolf et al., 2020). In KD experiments, we use a BERT-*large* teacher (345M parameters) fine-tuned using ERM on the source datasets. To prevent any confounding ef-

---

[1] https://github.com/mrqa/MRQA-Shared-Task-2019#datasets

Source	Train	Dev
SQuAD (Rajpurkar et al., 2016)	86,588	10,507
NewsQA (Trischler et al., 2017)	74,160	4,212
NQ (Kwiatkowski et al., 2019)	104,071	12,836
HotpotQA (Yang et al., 2018)	72,928	5,904
TriviaQA (Joshi et al., 2017)	61,688	7,785
SearchQA (Dunn et al., 2017)	117,384	16,980

Target	Test
BioASQ (Tsatsaronis et al., 2015)	1,504
DROP (Dua et al., 2019)	1,503
DuoRC (Saha et al., 2018)	1,501
RACE (Lai et al., 2017)	674
RelEx (Levy et al., 2017)	2,948
TextbookQA (Kembhavi et al., 2017)	1,503

Table 1: Dataset statistics (Fisch et al., 2019): # of examples.

fects from negative windows (Fisch et al., 2019), only those sliding windows of the training contexts are retained that actually contain an answer. We upsample from smaller training sets to make the number of examples equal across domains, and include examples from only one training set in each mini-batch (which worked better than multi-dataset mini-batches in our experiments). All hyperparameters are tuned by optimizing the macro-averaged F1 score on the in-domain validation sets. See Appendix A.3 for more details.

With every training method, we train six models, each on a unique five-set combination of the six source datasets. Each of these six models is selected using the validation sets of the same five datasets. The performance of the method on an individual test set is the mean F1 score of these six models on the set. Finally, the F1 scores on all six test sets are macro-averaged to measure the method's overall OOD performance.

For synthetic data generation, we fine-tune separate BART-*large* models (Lewis et al., 2020) on the individual source datasets. We generate 500k questions per dataset from Wikipedia contexts using top-p top-k sampling (p=.95, k=10).

3.3 Results

In Table 3 (column 1), we show in-domain validation results for ERM and the two variants of KD: ERM clearly exhibits some underfitting relative to the other two methods, as it has the lowest score of the three. KD using gold instances helps mitigate the underfitting effect to some extent; augmented KD with additional synthetic questions further improves results. These results confirm the utility of each of our methods in improving in-domain generalization of RC systems.

Method	Test Set						Avg.
	BioASQ	DROP	DuoRC	RACE	RelEx	TextbookQA	
ERM	51.7±1.4	37.8±0.2	55.1±1.5	39.0±1.0	83.3±0.3	51.6±1.8	53.1±1.0
*Domain-invariant learning to **limit overfitting**:*							
Domain-Adv	51.7±1.9	38.4±1.1	56.0±1.0	39.3±0.6	83.5±0.7	52.2±1.7	53.5±1.2
Episodic	52.0±1.6	38.4±1.4	56.4±1.1	40.1±0.5	83.3±0.4	52.3±1.9	53.7±1.2
MLDG	52.7±0.9	38.0±1.1	56.1±1.5	39.5±0.9	83.9±0.3	51.1±2.0	53.6±1.1
*Improved source domain learning to **address underfitting**:*							
KD (gold-only)	53.2±0.7	42.2±1.3	58.4±1.6	42.9±0.9	84.1±0.9	56.1±2.6	56.2±1.3
KD (augmented)	**53.4**±0.9	**45.2**±1.7	**60.3**±1.2	**44.2**±0.8	**84.8**±0.6	**58.0**±1.4	**57.6**±1.1

Table 2: Performance (F1 score) of different training methods on OOD test data. Each score is a mean±SD over six models, each trained on a unique five-set combination of the six source datasets. While the domain-invariant methods provide small gains over plain ERM, improved source domain learning demonstrates by far the best results.

Method	ID-Dev	OOD-Test
ERM	75.0	53.1
KD (gold-only)	76.4 (1.9%)	56.2 (**5.8%**)
KD (augmented)	77.2 (2.9%)	57.6 (**8.5%**)

Table 3: F1 score (and relative gain over ERM) for each proposed KD-based method. Gains on the OOD test sets outpace those on the in-domain dev sets, indicating strong generalization.

Figure 1: The ability of different methods to *cover*, *i.e.*, function as a proxy for, other methods. The leftmost plot shows the coverage of domain-adversarial training by other methods, for example. KD-based source domain learning provides the best coverage across the board. B: ERM, E: Episodic, M: MLDG, A: Domain-Adv, Kg: gold-only KD, Ka: KD combined with synthetic data augmentation.

Table 2 summarizes the performance of different training methods on the OOD test sets. The DIL methods do yield small gains over ERM, but improved source domain learning with KD and data augmentation outperforms by a much larger margin. Crucially, we observe improvements on each individual test set. In a two-tailed Student's t-test, we find all differences between (*a*) DIL and ERM, and (*b*) KD-based methods and DIL, to be statistically significant ($p < .0001$).

To further examine the generalization induced by KD, we compare in Table 3 their performance on in-domain validation data versus OOD test data: both methods provide substantially larger relative gains (5.8–8.5%) over ERM in OOD evaluation than in in-domain evaluation (1.9–2.9%). These results clearly indicate that many of the new and/or improved patterns the proposed methods teach do indeed generalize to other domains.

Method	Avg. F1
KD	**56.2**
KD + Domain-Adv	55.4
KD + Episodic	55.5
KD + MLDG	55.8

Table 4: Domain-invariant learning does not complement KD-based source domain learning in OOD tests.

Even though the KD-based methods exhibit stronger overall OOD generalization, it is still possible that DIL teaches certain DG-inducing patterns
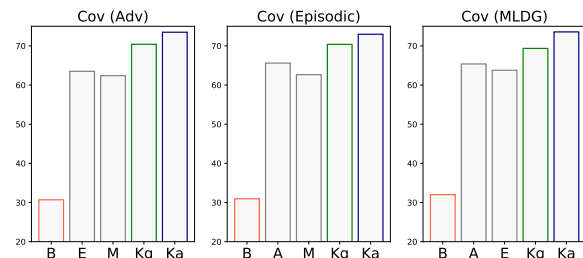
that even powerful source domain supervision fails to reveal, in which case the former should complement the latter well. To test this effect, we train three models using each of the three DIL methods, but replace ERM with KD (gold instances only) as the underlying training mechanism for RC. As Table 4 shows, none of the three combinations does better than KD alone. Along with the results of Table 2, this result indicates that as a learner is exposed to more powerful source domain supervision, DIL starts to lose its ability to complement its already strong OOD generalization.

As a final test of adequacy for multi-source KD as a DG method, we look at its ability to function as a proxy for the different DIL methods. Let $\mathcal{E}$ be the set of examples for which a training method M has a higher F1 score than ERM, representing the DG capabilities of M. We define the *coverage* of M by another method M′ as the relative F1 score of M′ as a fraction (%) of the F1 score of M on $\mathcal{E}$. This metric essentially quantifies the degree to which the DG capabilities of M is retained by

M′. As the bar charts of Figure 1 demonstrate, the KD-based methods provide the best coverage of all three DIL methods; the latter, while providing considerably better coverage of one another than ERM, lag behind KD in all three cases. These results again suggest that strong source domain learning may potentially be a sufficiently optimal policy for multi-source DG, without the need for an explicit enforcement of domain invariance.

4 Conclusion

This paper puts forward the view with empirical evidence for question answering (QA) that contrary to popular belief, multi-source domain generalization (DG) is better modeled as a problem of addressing model underfitting than overfitting. Our experimental results show that by simply learning the training domains well via knowledge distillation, even when the number of such domains is relatively small, strong out-of-domain (OOD) generalization can be achieved in QA without the need for any cross-domain regularization. These findings point to the need for a re-examination of whether *not fitting the noise* in training data is indeed a more reasonable path forward for DG than actually *learning the signal* in it; we hope that our work will inspire future efforts to answer this important question in greater detail with a wider range of tasks and approaches.

Limitations

We explore the problem of multi-source domain generalization (DG) in QA with new and existing methods. We believe that our findings will generalize to more baselines and datasets, but here we only show proof of concept for a select set of existing baselines and a single multi-dataset DG benchmark.

References

Gilles Blanchard, Gyemin Lee, and Clayton Scott. 2011. Generalizing from Several Related Classification Tasks to a New Unlabeled Sample. In *NeurIPS*.

Yanda Chen, Md Arafat Sultan, and Vittorio Castelli. 2020. Improved Synthetic Training for Reading Comprehension. *arXiv prePrint arXiv:2010.12776*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *NAACL*.

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 *ICML*.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension. In *EMNLP-IJCNLP MRQA Workshop*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*.

Ishaan Gulrajani and David Lopez-Paz. 2021. In Search of Lost Domain Generalization. In *ICLR*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the Knowledge in a Neural Network. In *NeurIPS Deep Learning Worksop*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *ACL*.

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 *CVPR*.

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. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the ACL*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *EMNLP*.

Seanie Lee, Donggyu Kim, and Jangwon Park. 2019. Domain-agnostic Question-Answering with Adversarial Training. In *EMNLP-IJCNLP MRQA Worksop*.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-Shot Relation Extraction via Reading Comprehension. In *CoNLL*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pretraining for Natural Language Generation, Translation, and Comprehension. In *ACL*.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. 2018. Learning to Generalize: Metalearning for Domain Generalization. In *AAAI*.

Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M. Hospedales. 2019. Episodic Training for Domain Generalization. In *ICCV*.

Kevin J Liang, Weituo Hao, Dinghan Shen, Yufan Zhou, Weizhu Chen, Changyou Chen, and Lawrence Carin. 2021. MixKD: Towards Efficient Distillation of Large-scale Language Models. In *ICLR*.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain Generalization via Invariant Feature Representation. In *ICML*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *EMNLP*.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension. In *ACL*.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the Importance of Diversity in Question Generation for QA. In *ACL*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A Machine Comprehension Dataset. In *2nd Workshop on Representation Learning for NLP*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, and Dimitris Polychronopoulos et al. 2015. An Overview of the BioASQ Large-Scale Biomedical Semantic Indexing and Question Answering Competition. *BMC Bioinformatics*.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. 2021. Generalizing to Unseen Domains: A Survey on Domain Generalization. In *IJCAI (Survey Track)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, and Remi Louf et al. 2020. Transformers: State-of-the-Art Natural Language Processing. In *EMNLP System Demonstrations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *EMNLP)*.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2021. Domain Generalization in Vision: A Survey. *arXiv preprint arXiv: 2103.02503*.

A Appendix

A.1 Qualtitative Analysis

To better understand what new patterns the proposed method of augmented knowledge distillation (KD-Aug) teaches, we take a closer look at a random sample of the test instances where the baseline model has an F1 score of zero and the student has an F1 score greater than .5. Table 5 shows four such examples from four different test sets. Even in this very small sample, we observe a number of different ways in which the KD-Aug student is better than the plain ERM baseline:

- BioASQ: The student has learned that synonyms are commonly mentioned in parentheses, especially in the beginning of a sentence.

- DuoRC: The word "overdose" can be difficult to model in context as the subject of the verb "overdose" becomes the object in the expression "giving someone an overdose". The student appears to have learned this nuanced semantics, which the baseline model has not.

- RelationExtraction: Unlike the baseline, the student knows that a fictional universe is not a time period, and the word "Universe" at the end of the phrase "DC Universe" is a strong hint to capitalize on in this particular case. Crucially, it has also learned to ignore the spurious surface-level match in this example between the question and the phrase "from the 30th century" in the passage.

- DROP: This one is a somewhat harder case to analyze, where a plausible explanation could be that the student has better knowledge about people's names and only the last name "Gostkowski" was sufficient for it to recognize it as a player's name.

Even though these patterns were taught by KD-Aug using a small number of source domains, their domain-agnostic importance is quite clear and intuitive, which is also supported by the experiments reported in this paper.

A.2 Further Commentary on KD Results

The BERT-*base* model fine-tuned using ERM has an F1 score of 75.0 on the in-domain dev set; the corresponding number for the BERT-*large* teacher is 77.3. As Table 3 shows, the augmented KD student has an F1 score of 77.2, indicating that additional synthetic questions facilitated almost perfect distillation from *large* to *base* on our source domains.

A.3 Model Selection

We use a training batch size of 32 for all models (with gradient accumulation when necessary). To ensure fair comparison, we train and validate all methods on a large set of learning rates: $\{1, 3, 5, 7, 9\} \times 10^{-6}$, $\{1, 3, 5, 7, 9\} \times 10^{-5}$ and $\{1, 3, 5\} \times 10^{-4}$, as the optimal learning rate varied drastically across different methods in our validation experiments. All models are trained for two epochs; we select the best of the epoch 1 and the epoch 2 checkpoint on the validation set for final evaluation on the OOD test set. In Table 6 we provide the optimal values of these two hyperparameters for all models.

Table 7 provides the optimal combinations of method-specific hyperparameters. Below is a brief description of each:

1. λ_{adv} (Domain-Adv): This is the weight of the adversarial loss of the domain classifier. The main ERM loss has a fixed weight of 1.

2. λ_{erm}, $\lambda_{episodic}$ (Episodic): The relative weights of the ERM loss and the episodic training loss in their convex combination.

3. β (MLDG): The weight of the meta-test loss during meta-optimization (second-order differentiation). The meta-train loss has a fixed weight of 1.

4. τ (KD): Temperature (Eq. 1).

A.4 Infrastructure and Computation

We run all experiments on a single V100 GPU with 32GB memory. A vast majority of the runs take less than a day; the longest ones take less than 48h.

Dataset: BIOASQ

Question: Name synonym of Acrokeratosis paraneoplastica.

Passage: Acrokeratosis paraneoplastica (Bazex ' syndrome) is a rare but clinically distinctive dermatosis that has been associated in all reported cases , to our knowledge , with either a primary malignant neoplasm of the upper aerodigestive tract or metastatic cancer to the lymph nodes of the neck . Acrokeratosis paraneoplastica was found in a 53-year - old black man with squamous cell carcinoma of the tonsil . A distinctive series of changes was found on histopathologic examination of biopsy specimens taken from his skin lesions , and direct immunofluorescence microscopy of both lesional and nonlesional skin specimens showed immunoglobulin and complement deposition on the epidermal basement membrane . The skin lesions largely resolved following radiation therapy of the neoplasm and of the presumably involved lymph nodes .

GT: ['Bazex syndrome']

ERM Answer: dermatosis

KD-Aug Answer: Bazex ' syndrome

Dataset: DUORC

Question: Who overdoses on insulin?

Passage: The film tells the story of a psychiatrist , Dr. Cross (Vincent Price) , who is treating a young woman , Janet Stewart (Anabel Shaw) , who is in a coma - state , brought on when she heard loud arguing , went to her window and saw a man strike his wife with a candlestick and kill her . It also stars Lynn Bari as Dr. Cross 's nurse / lover , Elaine Jordan . As Stewart comes out of her shock , she recognizes Dr. Cross as the killer . He then takes her to his sanitarium and at Elaine 's urging , gives Janet an overdose of insulin under the pretense of administering insulin shock therapy . He ca n't bring himself to murder her in cold blood , though , and asks Elaine to get the medicine to save her . Elaine refuses , they argue , and he strangles her . A colleague of Dr. Cross , Dr. Harvey , saves Janet 's life and Dr. Cross is taken into custody by a lawyer from the District Attorney 's office .

GT: ['Janet.', 'Janet']

ERM Answer: Dr. Cross

KD-Aug Answer: Janet Stewart

Dataset: RelationExtraction

Question: What is the name of the fictional universe that Polar Boy is from?

Passage: Polar Boy is a fictional character from the 30th century of the DC Universe , initially suggested by reader Buddy Lavigne of Northbrook , Illinois in the letters page of Adventure Comics # 304 , January , 1963 .

GT: ['DC Universe']

ERM Answer: 30th century

KD-Aug Answer: DC Universe

Dataset: DROP

Question: Which player scored the first points of the game?

Passage: The Patriots clinched their fourth straight AFC East title with a close road win . After a scoreless first quarter , the Jaguars responded to a Gostkowski field goal with a Maurice Jones - Drew touchdown run . The Patriots challenged the play , as Jones - Drew appeared to fall down at the line of scrimmage , but the ruling on the field was upheld . New England came back before the halftime to retake the lead at 10 - 7 on a Dillon one - yard touchdown run . The Patriots maintained their lead as the teams traded touchdowns in the second half , including another touchdown by Jones - Drew . A David Garrard fumble with 1:55 left in the fourth quarter , recovered by safety Rodney Harrison , sealed the Patriots ' 11th win of the season .

GT: ['Gostkowski']

ERM Answer: Maurice Jones

KD-Aug Answer: Gostkowski

Table 5: Examples of test cases where KD-based methods improve over plain ERM.

Method	Source Datasets	Learning Rate	# of Epochs
ERM	C, D, L, P, V	3e-5	2
	D, L, P, V, W	1e-5	2
	L, P, V, W, C	1e-5	2
	P, V, W, C, D	7e-6	2
	V, W, C, D, L	1e-5	2
	W, C, D, L, P	1e-5	2
Domain-Adv	C, D, L, P, V	5e-5	2
	D, L, P, V, W	7e-5	2
	L, P, V, W, C	7e-5	2
	P, V, W, C, D	7e-5	2
	V, W, C, D, L	9e-5	2
	W, C, D, L, P	9e-5	2
Episodic	C, D, L, P, V	9e-6	2
	D, L, P, V, W	9e-6	2
	L, P, V, W, C	1e-5	2
	P, V, W, C, D	7e-6	2
	V, W, C, D, L	7e-6	2
	W, C, D, L, P	9e-6	2
MLDG	C, D, L, P, V	9e-6	2
	D, L, P, V, W	1e-5	2
	L, P, V, W, C	1e-5	2
	P, V, W, C, D	3e-5	1
	V, W, C, D, L	9e-6	2
	W, C, D, L, P	9e-6	2
KD Teacher (ERM)	C, D, L, P, V	9e-6	1
	D, L, P, V, W	7e-6	2
	L, P, V, W, C	7e-6	2
	P, V, W, C, D	9e-6	1
	V, W, C, D, L	1e-5	1
	W, C, D, L, P	7e-6	1
KD (gold-only)	C, D, L, P, V	7e-5	2
	D, L, P, V, W	3e-5	2
	L, P, V, W, C	3e-5	2
	P, V, W, C, D	5e-5	2
	V, W, C, D, L	5e-5	2
	W, C, D, L, P	3e-5	2
KD (augmented)	C, D, L, P, V	5e-5	*synthetic:* 1, *gold:* 2
	D, L, P, V, W	5e-5	*synthetic:* 1, *gold:* 2
	L, P, V, W, C	3e-5	*synthetic:* 1, *gold:* 1
	P, V, W, C, D	5e-5	*synthetic:* 1, *gold:* 2
	V, W, C, D, L	5e-5	*synthetic:* 1, *gold:* 2
	W, C, D, L, P	9e-6	*synthetic:* 1, *gold:* 2
KD (gold) w/ Domain-Adv	C, D, L, P, V	7e-5	2
	D, L, P, V, W	9e-5	2
	L, P, V, W, C	9e-5	2
	P, V, W, C, D	7e-5	2
	V, W, C, D, L	1e-4	2
	W, C, D, L, P	1e-4	2
KD (gold) w/ Episodic	C, D, L, P, V	3e-5	2
	D, L, P, V, W	3e-5	2
	L, P, V, W, C	3e-5	2
	P, V, W, C, D	3e-5	2
	V, W, C, D, L	3e-5	2
	W, C, D, L, P	3e-5	2
KD (gold) w/ MLDG	C, D, L, P, V	5e-5	2
	D, L, P, V, W	5e-5	2
	L, P, V, W, C	5e-5	2
	P, V, W, C, D	5e-5	2
	V, W, C, D, L	7e-5	2
	W, C, D, L, P	5e-5	2

Table 6: Optimal values of shared hyperparameters (learning rate, # of epochs). C: SearchQA, D: SQuAD, L: NaturalQuestions (NQ), P: HotpotQA, V: TriviaQA, W: NewsQA.

Method	Hyperparameters	Grid	Optimal
Domain-Adv	λ_{adv}	$\{0.01, 0.1, 1.0\}$	0.1
Episodic	$(\lambda_{erm}, \lambda_{episodic})$	$\lambda_{erm} \in \{0.25, 0.5, 0.75, 0.9\}$ $\lambda_{episodic} = 1 - \lambda_{erm}$	$(0.75, 0.25)$
MLDG	β	$\{1\}$	1
KD (gold)	τ	$\{1, 2, 4\}$	2
KD (augmented)	τ	$\{4\}$	4
KD (gold) w/ Domain-Adv	(τ, λ_{adv})	$\{1, 2, 4\} \times \{0.01, 0.1, 1.0\}$	$(1, 0.01)$
KD (gold) w/ Episodic	$(\tau, \lambda_{erm}, \lambda_{episodic})$	$\{1, 2, 4\} \times \lambda_{erm} \in \{0.25, 0.5, 0.75, 0.9\}$ $\lambda_{episodic} = 1 - \lambda_{erm}$	$(1, 0.75, 0.25)$
KD (gold) w/ MLDG	(τ, β)	$\{1, 2, 4\} \times \{1\}$	$(4, 1)$

Table 7: Optimal values of hyperparameters specific to different training methods and the respective search grids.