Multi-source Meta Transfer for Low Resource Multiple-Choice Question Answering

Ming Yan[1], **Hao Zhang**[1,2], **Di Jin**[3], **Joey Tianyi Zhou**[1,*]
[1]Institute of High Performance Computing, A*STAR, Singapore
[2]SCSE, Nanyang Technological University, Singapore
[3]CSAIL, Massachusetts Institute of Technology, USA
mingy@ihpc.a-star.edu.sg, zhang_hao@ihpc.a-star.edu.sg
jindi15@mit.edu, joey_zhou@ihpc.a-star.edu.sg

Abstract

Multiple-choice question answering (MCQA) is one of the most challenging tasks in machine reading comprehension since it requires more advanced reading comprehension skills such as logical reasoning, summarization, and arithmetic operations. Unfortunately, most existing MCQA datasets are small in size, which increases the difficulty of model learning and generalization. To address this challenge, we propose a multi-source meta transfer (MMT) for low-resource MCQA. In this framework, we first extend meta learning by incorporating multiple training sources to learn a generalized feature representation across domains. To bridge the distribution gap between training sources and the target, we further introduce the meta transfer that can be integrated into the multi-source meta training. More importantly, the proposed MMT is independent of backbone language models. Extensive experiments demonstrate the superiority of MMT over state-of-the-arts, and continuous improvements can be achieved on different backbone networks on both supervised and unsupervised domain adaptation settings.

1 Introduction

Recently, there has been a growing interest in making machines to understand human languages, and a great progress has been made in machine reading comprehension (MRC). There are two main types of MRC task: 1) extractive/abstractive question answering (QA) such as SQuAD (Rajpurkar et al., 2018) and DROP (Dua et al., 2019); 2) multiple-choice QA (MCQA) such as MultiRC (Khashabi et al., 2018) and DREAM (Sun et al., 2019a). Different from extractive/abstractive QA whose answers are usually limited to the text spans exist in the passage, the answers of MCQA may not appear in the text passage and may involve complex language inference. Thus, MCQA usually requires more advanced reading comprehension abilities, including arithmetic operation, summarization, logic reasoning and commonsense reasoning (Richardson et al., 2013; Sun et al., 2019a), and etc. In addition, the size of most existing MCQA datasets is much smaller than that of the extractive/abstractive QA datasets. For instance, all the span-based QA datasets, except CQ (Bao et al., 2016), contain more than $100k$ samples. In contrast, the data size of most existing MCQA datasets are far less than $100k$ (see Table 1), and the smallest one only contains 660 samples.

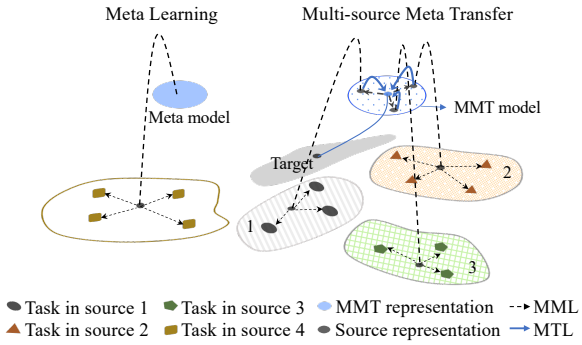

Figure 1: Comparison of meta learning and multi-source meta transfer learning (MMT). "MTL" denotes meta transfer learning, and "MML" denotes multi-source meta learning.

The above two major challenges make MCQA much more difficult to optimize and generalize, especially for the low resource issue. In order to achieve better performance on downstream NLP tasks, it is inevitable to fine-tune the pre-trained deep language models (Devlin et al., 2019; Raffel et al., 2019; Dai et al., 2019; Liu et al., 2019; Yang et al., 2019) with a large number of supervised target data for reducing the discrepancy between the training source and target data. Due to the low resource nature, the performance of most existing

* Corresponding author.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7331–7341
July 5 - 10, 2020.

MCQA methods is far from satisfactory. To alleviate such issue in MCQA, one straightforward solution is to merge all available data resources for training (Palmero Aprosio et al., 2019). However, the data heterogeneity of datasets (*e.g.,* resource domains, answer types and varies diversity of choice size across different MCQA datasets.) hinders the practical use of this strategy.

To better discover the hidden knowledge across multiple data sources, we propose a novel framework termed **Multi-source Meta Transfer (MMT)**. In this framework, we first propose a module named *multi-source meta learning* (MML) that extends traditional meta learning to multiple sources where a series of meta-tasks on different data resources is constructed to simulate low-resource target task. In this way, a more generalized representation could be obtained by considering multiple source datasets. On the top of it, the *meta transfer learning* (MTL) is integrated into multi-source meta training to further reduce the distribution gap between training sources and the target one. Different from traditional meta learning that assumes tasks generated from the similar distribution/same dataset, MMT is able to discover the knowledge across different datasets and transfer it into the target task. More importantly, MMT is agnostic to the upstream framework, *i.e.,* it can be seamlessly incorporated into any existing backbone language models to improve performance. Figure 1 briefly illustrates both meta learning and the proposed MMT.

2 Related Work

2.1 Meta Learning

Meta learning, *a.k.a* "learning to learn", intends to design models that can learn general data representation and adapt to new tasks with a few training samples (Finn et al., 2017; Nichol et al., 2018). Early works have demonstrated that meta learning is capable of boosting the performance of natural language processing (NLP) tasks, such as named entity recognition (Munro et al., 2003) and grammatical error correction (Seo et al., 2012).

Recently, meta learning gains more and more attention. Many works explore to adopt meta learning to address low resource issues in various NLP tasks, such as machine translation (Gu et al., 2018; Sennrich and Zhang, 2019), semantic parsing (Guo et al., 2019), query generation (Huang et al., 2018), emotion distribution learning (Zhao and Ma, 2019), relation classification (Wu et al., 2019; Obamuyide and Vlachos, 2019) and etc. These methods have all achieved good performance due to their powerful data representation ability. Meanwhile, the strong learning capability of meta learning also provides deep models with a better initialization, and boosts deep models fast adaptation to new tasks under both supervised (Qian and Yu, 2019; Obamuyide and Vlachos, 2019) and unsupervised (Srivastava et al., 2018) scenarios. Unfortunately, meta learning is seldom studied in multiple-choice question answering in existing methods. To our best knowledge, it is also the first time to extend meta learning into multi-source scenarios.

2.2 Multiple-Choice Question Answering

Multiple-choice question answering (MCQA) is a challenging task, which requires understanding the relationships and handle the interactions between passages, questions and choices to select the correct answer (Chen and Durrett, 2019). As one of the hot track of question answering tasks, MCQA has seen a great surge of challenging datasets and novel architectures recently. These datasets are built through considering different contexts and scenes. For instance, Guo et al. (2017) present an open-domain comprehension dataset; Lai et al. (2017) build a QA dataset from examinations, which requires more complex reasoning on questions; and Zellers et al. (2018) introduce a QA dataset that requires both natural language inference and commonsense reasoning. Meanwhile, various approaches have been proposed to address the MCQA task using different neural network architectures. Some works propose to compute the similarity between question and each of the choices through an attention mechanism (Chaturvedi et al., 2018; Wang et al., 2018). Kumar et al. (2016) construct the context embedding for semantic representation. Liu et al. (2018) and Yu et al. (2019) apply the recurrent memory network for question reasoning. Chung et al. (2018) and Jin et al. (2019) further incorporate an attention mechanism into recurrent memory networks for multi-step reasoning. Most existing works only strive to increase the reasoning capability by constructing complex models, but ignore the low resource nature of those available MCQA datasets.

3 Methodology

Many existing MCQA tasks suffer from the low-resource issue, which requires a special training strategy to tackle it. Recent advance of meta learning shows its advantages in solving the few-shot learning problem. Typically, it can rely on only a very small number of training samples to train a model with good generalization ability (Finn et al., 2017; Nichol et al., 2018). Unfortunately, the existing meta learning algorithms are unable to be applied in our problem setting directly, since they are based on the assumption that the meta tasks are generated from the same data distribution (Fallah et al., 2019). For example, one of the most popular benchmarks is the Mini-ImageNet dataset that was proposed by Lake et al. (2011), and it consists of 100 sub-classes from ImageNet dataset. All the meta tasks generated from the same training dataset have similar properties. In contrast, in our studied problem MCQA, data properties such as answer, question type, and commonsense are greatly vary across the MCQA datasets. Specifically, the passages and questions come from different scenarios (such as exams, dialogues, and stories), and the answering choice contains more complex semantic information than the fixed categories in Mini-ImageNet. Therefore, simply combining all the data resources into one and feeding it into existing meta learning algorithms is not an optimal solution (the experimental results in Figure 5 also support this point).

To address the data heterogeneity challenge and cater to the MCQA task, we extend the traditional meta learning method to multiple training sources scenarios, where we fully exploit multiple inter-domain sources to learn more generalized representations. Specifically, multi-source meta learning performs meta learning among multiple sources in sequence, thereby completing one iteration. However, multi-source meta learning alone cannot guarantee the desirable performance due to the data distribution gap between multiple sources and target data. Therefore, transfer learning from multi-sources to target is required. Here we introduce meta transfer learning into each meta learning iteration, which aims at reducing the discrepancy between the learned meta representation from multi-source and target.

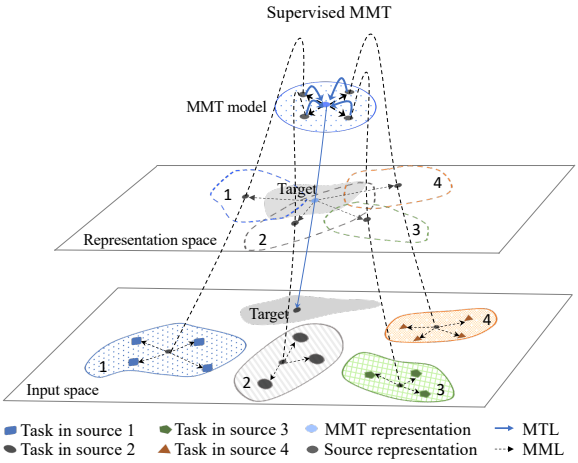

Figure 2: Architecture of multi-source meta transfer (MMT), where dot-line denotes multi-source meta learning (MML) and solid-line represents meta transfer learning (MTL).

3.1 Multi-source Meta Transfer

The proposed multi-source meta transfer (MMT) method consists of two modules: multi-source meta learning (MML) and meta transfer learning (MTL). As shown in Figure 2, the MML contains fast adaptation, meta-model update and target fine-tuning steps; and the MTL performs to transfer the knowledge initialized by MML to the target task. Note that MMT is agnostic to backbone models, *i.e.,* it can be seamlessly incorporated into any stronger backbone to boost performance. In this work, we select pre-trained BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) as the backbone for MMT. Generally, MMT first learns meta features from multiple sources of inputs such that those features could be mapped into a latent representation space. Then, the fine-tuning step performs to reduce the representation gap between different sources and the meta representation. Finally, MTL is applied to transfer the well-initialized meta representations to the target task.

The details of MMT are summarized in Algorithm 1, where the procedures of MML and MTL are presented in lines 2-16 and lines 17-21, respectively. In MML, we sequentially sample data to construct the tasks τ in meta learning from multiple source distributions $\{p^s(\tau); s \in S\}$, where S denotes the sources index set. Note that the support-tasks and query-tasks, in one iteration of MML, should be sampled simultaneously to satisfy the same distribution requirement. The learning rates for each of the learning modules are different,

where α denotes the learning rate for fast adaptation module, β is utilized for both meta-model updating and target fine-tuning, and λ represents the learning rate for MTL. Moreover, the parameter of MMT is initialized from the backbone language model, *i.e.,* BERT, RoBERTa.

In the sequence, we introduce each step in multi-source meta learning (MML) module. The first step is fast adaptation (lines 4-8), which aims to learn the meta information from support-tasks τ_i^s. The task-specific parameter θ' is updated by

$$\theta' = \theta - \alpha \nabla_\theta L_{\tau_i^s}(f(\theta)), \tag{1}$$

where the gradient $\nabla_\theta L_{\tau_i^s}(f(\theta))$ is computed by the cost function $L_{\tau_i^s}(f(\theta))$ with respect to model parameter θ.

The second step is meta-model update (line 9), where its cost function, $\sum_{\tau_i^s \sim p^s(\tau)} L_{\tau_i^s}(f(\theta'))$, is calculated with respect to θ', and it is adopted to evaluate the performance of fast adaptation on the corresponding newly sampled query-tasks (τ_i^s).

It is worth noting that $f(\theta')$ is an implicit function of θ (see Equation 1), and the second-order Hessian gradient matrix is required for the gradient computation (Nichol et al., 2018). However, the use of second derivatives is computationally expensive, so we employ a first-order approximation (Obamuyide and Vlachos, 2019) to update the meta-model gradient by

$$\theta = \theta - \beta \nabla_\theta \sum_{\tau_i^s \sim p^s(\tau)} L_{\tau_i^s}(f(\theta')). \tag{2}$$

The last step of MML is target fine-tuning (lines 10-14). Although the learnt meta representations carry sufficient semantic knowledge and are well generalized, the data distribution discrepancy between meta representation and target still exists. This fine-tuning step is utilized to reduce the distance between the meta representation and target task on the latent representation space.

Generally, all the steps in MML are sequentially conducted until the meta-model converges. After performing MML, the meta transfer learning (MTL) module will be applied upon the learnt meta representations for the final transfer learning on target data.

Algorithm 1: The procedure of MMT.

Input: Task distribution over source $p^s(\tau)$, data distribution over target $p^t(\tau)$, backbone model $f(\theta)$, learning rates in MMT α, β, λ.
Output: Optimized parameters θ.

1 Initialize θ from backbone model;
2 **while** *not done* **do**
3 **for** *all source* S **do**
4 Sample batch of tasks $\tau_i^s \sim p^s(\tau)$;
5 **for** *all* τ_i^s **do**
6 Evaluate $\nabla_\theta L_{\tau_i^s}(f(\theta))$ with respect to K examples;
7 Compute gradient for fast adaption: $\theta' = \theta - \alpha \nabla_\theta L_{\tau_i^s}(f(\theta))$;
8 **end**
9 Meta model update:$\theta = \theta - \beta \nabla_\theta \sum_{\tau_i^s \sim p^s(\tau)} L_{\tau_i^s}(f(\theta'))$;
10 Get batch of data $\tau_i^t \sim p^t(\tau)$;
11 **for** *all* τ_i^t **do**
12 Evaluate $\nabla_\theta L_{\tau_i^t}(f(\theta))$ with respect to K examples;
13 Gradient for target fine-tuning: $\theta = \theta - \beta \nabla_\theta L_{\tau_i^t}(f(\theta))$;
14 **end**
15 **end**
16 **end**
17 Get all batches of data $\tau_i^t \sim p^t(\tau)$;
18 **for** *all* τ_i^t **do**
19 Evaluate $\nabla_\theta L_{\tau_i^t}(f(\theta))$ with respect to batch size;
20 Gradient for meta transfer learning: $\theta = \theta - \lambda \nabla_\theta L_{\tau_i^t}(f(\theta))$;
21 **end**

3.2 Unsupervised Domain Adaptation

In this section, we extend MMT to the unsupervised domain adaptation setting, where no labeled data from the target domain will be given. In this setting, the difficulty of unsupervised domain adaptation arises due to the different number of choices between source and target datasets. This issue hinders the pre-trained model to be applied to the target task whose choices differ from the source task, *i.e.,* only the knowledge of feature encoders are transferable. To address this issue, unsupervised MMT constructs the support/query-tasks by sampling, which makes the choice number of tasks in the source equal to the target task. With this manner, the unsupervised MMT is able to transfer the knowledge of both feature encoders and classifier to the target task. Some prior works (Chung et al.,

2018) also investigated on the unsupervised transfer learning in QA, but they did not well solve the category difference issue exists in multi-sources learning. To the best of our knowledge, we are the first to apply meta learning to address knowledge transfer issue between tasks with different choices in the unsupervised domain adaptation setting. Next, we term our proposed method as unsupervised MMT in short.

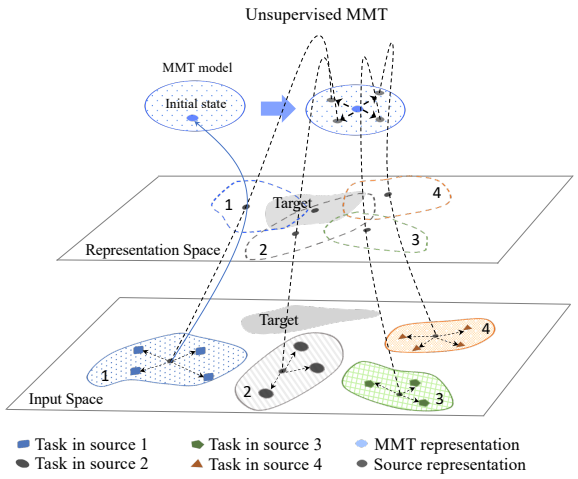

Figure 3: The framework of unsupervised MMT. The initial state of our unsupervised MMT is the pre-trained knowledge transferred from one specific source.

The framework of unsupervised MMT is shown in Figure 3. A specific source is pre-trained, as an initial state of meta model, to reduce the optimization cost of MMT learning without prior information. With this initial state, unsupervised MMT conducts meta learning by the steps of fast adaption and meta-model update iteratively. Correspondingly, the training of unsupervised MMT is implemented by removing the fine-tuning procedures (lines 10-14 and lines 17-21) in Algorithm 1. By this manner, unsupervised MMT shortened the target representation discrepancy from the specific transferred representation to a generalized meta representation. Moreover, unsupervised MMT fast adapts to category variable tasks without supervised fine-tuning, which relaxes the fixed-category constraint in transfer learning.

3.3 Source Selection in MMT

Source selection is a prerequisite step for MMT. Due to the data heterogeneity of different sources, the performance of meta learning may drop if we consider some undesirable data sources in training. In other words, these undesirable or called "dissimilar" data sources will cause negative transfer when their distribution is far away from the target one. To eliminate such drawback, we may consider those "similar" datasets from all the available data sources. In the experiments, we also evaluate the transfer performance of the all source datasets on the target task. The more "similar" of source to target data, the better improvements can be achieved through MMT on the target tasks. Therefore, we use the transfer performance as a guidance for the sequential multi-source meta transfer training, *i.e.,* learns from dissimilar sources to a similar one.

4 Experiments

4.1 Dataset

We conduct experiments to evaluate the performance of MMT on the following MCQA benchmark datasets.

DREAM (Sun et al., 2019a) is a dialogue-based dataset designed by education experts to evaluate the English level of nonnative English speakers. It focuses on multi-tune multi-party dialogue understanding, which contains various types of questions, like summary, logic, arithmetic, commonsense, etc.

MCTEST (Richardson et al., 2013) is a fictional stories dataset which aims to evaluate open-domain machine comprehension. The stories contain open domain topics, and the questions and choices are created by crowd-sourcing with strict grammar, quality guarantee.

RACE (Lai et al., 2017) is a dataset about passage reading comprehension, which collected from middle/high school English examinations. Human experts design the questions, and the passages cover various categories of human articles: news, stories, advertisements, biography, philosophy, etc.

SemEval-2018-Task11 (Ostermann et al., 2018) consists of scenario-related narrative text and various types of questions. The goal is to evaluate the machine comprehension for commonsense knowledge.

SWAG (Zellers et al., 2018) is a dataset about rich grounded situations, which is constructed debiased with adversarial filtering and explores the gap between machine comprehension and human.

The statistics of DREAM, MCTEST, RACE, SemEval-2018-Task11 (SemEval) and SWAG are summarized in Table 1.

Name	DREAM	RACE	MCTEST	SemEval	SWAG
Type	Dialogue	Exam	Story	Narrative Text	Scenario Text
Ages	15+	12-18	7+	-	-
Generator	Expert	Expert	Crowd.	Crowd.	AF./Crowd.
Level	High School/College	High/Middle School	Children	Unlimited	Unlimited
Choices	3	4	4	2	4
Samples	6,444	27,933	660	2,119	92,221
Questions	10,197	97,687	2,640	13,939	113,557

Table 1: Statistics of MCQA datasets, where "Crowd." denotes questions generated by crowd-sourcing, and "AF." denotes question generated by adversarial filtering.

Methods	DREAM		MCTEST		SemEval	
	Dev	Test	Dev	Test	Dev	Test
CoMatching (Wang et al., 2018)	45.6	45.5	-	-	-	-
HFL (Chen et al., 2018)	-	-	-	-	86.46	84.13
QACNN (Chung et al., 2018)	-	-	-	72.66	-	-
IMC (Yu et al., 2019)	-	-	-	76.59	-	-
XLNet (Yang et al., 2019)	-	72.0	-	-	-	-
GPT+Strategies (2×) (Sun et al., 2019b)	-	-	-	81.9	-	89.5
BERT-Base	60.05	61.58	70.0	67.98	86.03	87.53
RoBERTa†	82.16	84.37	88.37	87.26	93.76	94.00
MMT (BERT-Base)	**68.38**	**68.89**	**81.56**	**82.02**	**88.52**	**88.85**
MMT (RoBERTa)†	**83.87**	**85.55**	**88.66**	**88.80**	**94.33**	**94.24**

Table 2: Comparison with state-of-the-art methods in MCQA datasets, where "†" denotes the maximal sequence length of RoBERTa-large is limited to 256.

4.2 Experimental Setting

To demonstrating the versatility of MMT, we adopt both BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) as the backbone. Due to the resource limitation, the maximal sequence input lengths of BERT and RoBERTa can only be set as 512 and 256, respectively. For all datasets, the model optimization is performed by Adam (Kingma and Ba, 2014), the initial learning rate of fast adaptation α is set to $1e-3$, and the rest ones are set to $1e-5$.

4.3 Supervised MCQA

The results of MCQA under supervised setting are summarized in Table 2. Note that we reproduce the results of BERT-Base and RoBERTa-Large on the benchmark datasets in our experiment setting for fair comparison. From the results, we can see that MMT(RoBERTa) achieves the best performances overall benchmark datasets and outperforms current SOTAs with significant margins (i.e., from 5% to 13%). Second, MMT is able to boost up performance over different pre-trained language models. While, the weaker backbone network is, the better improvement MMT can achieve. For example, the MMT(BERT-Base) improves BERT-Base over 14% on MCTEST. In contrast, MMT(RoBERTa) only achieves 1.54% on MCTEST. The performance difference between MMT(RoBERTa) and MMT(BERT-Base) is mainly related to the performance of backbone itself and the scale of backbone parameter in MMT optimization. We also want to point out that one of the advantages for MMT is backbone-free, which indicates that its performance can be improved progressively with the advance of language models.

4.4 Unsupervised Domain Adaptation for MCQA

In this experiment, we further evaluate the performance of MMT under the unsupervised domain adaptation, where no labeled data from the target domain will be available. We use BERT-Base as the backbone, and the model is trained on SWAG and RACE training sources, which is termed as unsupervised MMT(S+R). We also compare it with other SOTAs as well as some transfer learning baselines "TL(∗)". For example, "TL(R-S)" denotes that BERT-Base is first fine-tuned in sequence on RACE and SWAG, and then test on MCTEST.

The results of MCTEST are summarized in Table 3. From the results, we observe that the unsupervised MMT significantly outperforms other unsupervised domain adaptation methods, *e.g.,* MemN2N (Chung et al., 2018) and QACNN (Chung et al., 2018) by a large margin. Moreover, unsupervised MMT can beat some supervised methods, such as BERT-Base, IMC (Yu et al., 2019), even without any labeled data from

Method	Sup.	Test
Bert-Base	Yes	67.98
QACNN (Chung et al., 2018)	Yes	72.66
IMC (Yu et al., 2019)	Yes	76.59
MemN2N (Chung et al., 2018)	No	53.39
QACNN (Chung et al., 2018)	No	63.10
TL(S)	No	50.02
TL(R)	No	77.02
TL(R-S)	No	62.97
TL(S-R)	No	77.38
TL(R+S)	No	79.17
Unsupervised MMT(S+R)	No	**81.55**

Table 3: Unsupervised domain adaptation on MCTEST. "Sup." denotes supervised, "S" denotes SWAG, "R" denotes RACE, and "TL(∗)" denotes transfer learning from different datasets to MCTEST. For example, "TL(R-S)" denotes that Bert-Base is first fine-tuned on RACE, then on SWAG. Unsupervised MMT(S+R) denotes that the meta model is trained on the sources of SWAG and RACE.

the target domain. For a more fair comparison, we also create several transfer learning baselines that can utilize multiple training sources such as TL(R-S) and TL(S-R). From the results, we can conclude that unsupervised MMT is a better solution to make full use of multiple training sources than sequential transfer learning.

Similar observations hold on SWAG dataset. Reported in Table 4, unsupervised MMT outperforms other methods significantly. Note we follow the same setting in KagNet (Lin et al., 2019) that only the development set of SWAG is evaluated.

Method	Sup.	Dev
LSTM+GLV (Zellers et al., 2018)	Yes	43.1
DA+GlV (Zellers et al., 2018)	Yes	47.4
DA+ELMo (Zellers et al., 2018)	Yes	47.7
TL(R)	No	44.83
TL(M)	No	50.03
TL(R-M)	No	46.48
TL(M-R)	No	46.91
TL(M+R)	No	48.65
Unsupervised MMT(R+M)	No	**50.77**

Table 4: Unsupervised domain adaptation on SWAG, where "M" denotes MCTEST, "R" denotes RACE, "DA" denotes decomposable attention, and "GLV" denotes GloVe vectors.

5 Discussion

5.1 Ablation Study

We conduct ablative experiments to analyze the two modules of MMT, *i.e.,* multi-source meta learning (MML) and meta transfer learning (MTL). The MTL is the transfer learning module specifically designed for MML, and TL denotes the traditional transfer learning without MML. The experiments are based on BERT-Base model, and all the results are reported in Table 5.

Dream	Dev	Test
BERT-Base	60.05	61.58
+MML(M)	49.85	52.87
+MML(R)	49.56	51.69
+MML(M∪R)	29.60	29.20
+TL(M)	60.31	60.14
+TL(R)	68.72	67.72
+TL(R-M)	68.97	67.38
+TL(M+R)	68.61	68.15
+MMT(M)	67.99	68.54
+MMT(R)	68.04	68.69
+MMT(M∪R)	61.72	60.12
MMT(M+R)	**68.38**	**68.89**

Table 5: Ablation study of MMT on DREAM. "TL" denotes supervised transfer learning, "M" denotes MCTEST, "R" denotes RACE, and "∪" denotes the task combination of RACE and MCTEST.

In the first experiments, we present the results of the MML module. When the input source for MML is a single source, MML downgrades to the traditional meta learning. From the results, we observe that MML fine-tuned on MCTEST (MML(M)) is better than that on RACE (MML(R)), which is caused by the large difference between the RACE and DREAM datasets. We also compare the baseline that simply combines RACE and MCTEST datasets to be one large training source, denoted by MML(M∪R), dramatically drops the performance and only achieves 29.20% on DREAM dataset, which is 23.67% lower than that of MML(M). This suggests that a simple combination of the two different training datasets for meta training is not a good choice.

For the transfer learning (TL) module, we can observe that the performance improvement is more significant by transferring knowledge from RACE to DREAM, compared to that from MCTEST. In addition, TL(R-M) also benefits from fine-tuning on RACE and MCTEST sequentially, and achieves better results.

With the help of MTL, MMT further boosts the performance on DREAM and outperforms both MML and TL baselines. For instance, MMT(M) outperforms MML(M) and TL(M) with 15.67% and 8.40%, respectively. Moreover, MMT is also helpful in alleviating the overfitting issue that exists in TL baselines. The results of development set for TL(∗) are higher than the test set, which indicates the poor generalization ability of TL(∗). Fortunately, MMT(∗) is able to address this issue. The MMT(R+M) that is trained on both RACE and MCTEST in meta learning manner, achieves the best results in all evaluated methods.

5.2 Source Selection for MMT

Source selection is a prerequisite step for MMT. In previous experiments, we assume that training resources are given without selection. Due to the data heterogeneity of different sources, the performance of meta learning may drop if we incorporate some undesirable data sources in training. In this experiment, we evaluate the transferability between different datasets and further give the suggestion on the source selection for MMT. The results are summarized in Figure 4. In Figure, the X-axis denotes the source, and Y-axis denotes the target. The values in the boxes indicate transferability from source to the target data in terms of accuracy. For example, 14 denotes transferring RACE to the target MCTEST will obtain 14% accuracy improvement over that only trained on the MCTEST. The negative value in the transferability matrix suggests the negative transfer. There is no source that can be used to improve the performance of SWAG effectively.

Figure 4: Transferability matrix. X-axis denotes the source, and Y-axis denotes the target. The values indicate the transferability from source to the target data in terms of accuracy. The higher the value is, the stronger the transferability is. Taking MCTEST dataset for example, transfer learning pre-trained on RACE leads 14% performance improvement than fine-tuning on MCTEST only.

In MMT, we employ this transferability matrix to guide the source selection for MML training. Specifically, in supervised MMT, we only choose those training sources with the significant positive transfer. In unsupervised MMT, the source with the highest score is selected to be the initial state.

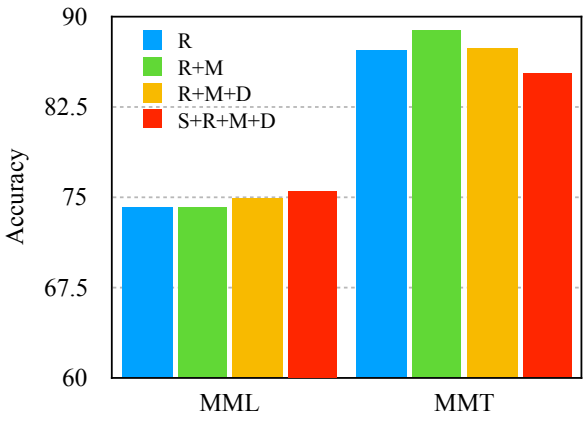

Figure 5: Training with different sources and test on SemEval. Where "S" denotes SWAG, "R" denotes RACE, "M" denotes MCTEST, and "D" denotes DREAM.

To verify the impact of different dataset to MMT, we further study the improvement on target SemEval by training with different sources. The results is shown in Figure 5. The performance of SemEval drops when we incorporate DREAM and SWAG into training. Recall the transferability matrix in Figure 4, the DREAM and SWAG datasets show little help in improving the performance on SemEval compared to RACE and MCTEST. In summary, more source data do not guarantee better performance. Only the "similar" source data will be beneficial for multi-source meta learning.

6 Conclusion

In this work, we propose a novel method named multi-source meta transfer for multiple-choice question answering on low resource setting. Our method considers multiple sources meta learning and target fine-tuning into a unified framework, which is able to learn a general representation from multiple sources and alleviate the discrepancy between source and target. We demonstrate the superiority of our methods on both supervised setting and unsupervised domain adaptation settings over the state-of-the-arts. In future work, we explore to extend this approach for other low resource tasks in NLP.

Acknowledgments

The paper is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project No. A18A1b0045).

References

Junwei Bao, Nan Duan, Zhao Yan, Ming Zhou, and Tiejun Zhao. 2016. Constraint-based question answering with knowledge graph. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2503–2514, Osaka, Japan. The COLING 2016 Organizing Committee.

Akshay Chaturvedi, Onkar Pandit, and Utpal Garain. 2018. CNN for text-based multiple choice question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 272–277, Melbourne, Australia. Association for Computational Linguistics.

Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. 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 4026–4032, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhipeng Chen, Yiming Cui, Wentao Ma, Shijin Wang, Ting Liu, and Guoping Hu. 2018. Hfl-rc system at semeval-2018 task 11: hybrid multi-aspects model for commonsense reading comprehension. *arXiv preprint arXiv:1803.05655*.

Yu-An Chung, Hung-Yi Lee, and James Glass. 2018. Supervised and unsupervised transfer learning for question answering. 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 1585–1594, New Orleans, Louisiana. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

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

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 *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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. 2019. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. *arXiv preprint arXiv:1908.10400*.

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*, ICML'17, pages 1126–1135, Sydney, NSW, Australia. JMLR.org.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. 2019. Coupling retrieval and meta-learning for context-dependent semantic parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 855–866, Florence, Italy. Association for Computational Linguistics.

Shangmin Guo, Kang Liu, Shizhu He, Cao Liu, Jun Zhao, and Zhuoyu Wei. 2017. IJCNLP-2017 task 5: Multi-choice question answering in examinations. In *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 34–40, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Po-Sen Huang, Chenglong Wang, Rishabh Singh, Wen-tau Yih, and Xiaodong He. 2018. Natural language to structured query generation via meta-learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 732–738, New Orleans, Louisiana. Association for Computational Linguistics.

Di Jin, Shuyang Gao, Jiun-Yu Kao, Tagyoung Chung, and Dilek Hakkani-tur. 2019. Mmm: Multi-stage multi-task learning for multi-choice reading comprehension. *arXiv preprint arXiv:1910.00458*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. 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 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1378–1387, New York, New York, USA. PMLR.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. 2011. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, pages 2573–2575, Boston, Massachusetts, USA. Curran Associates, Inc.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839, Hong Kong, China. 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, Melbourne, Australia. Association for Computational Linguistics.

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

Robert Munro, Daren Ler, and Jon Patrick. 2003. Meta-learning orthographic and contextual models for language independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 192–195, Edmonton, Canada. Association for Computational Linguistics.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Abiola Obamuyide and Andreas Vlachos. 2019. Model-agnostic meta-learning for relation classification with limited supervision. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5873–5879, Florence, Italy. Association for Computational Linguistics.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 task 11: Machine comprehension using commonsense knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 747–757, New Orleans, Louisiana. Association for Computational Linguistics.

Alessio Palmero Aprosio, Sara Tonelli, Marco Turchi, Matteo Negri, and Mattia A. Di Gangi. 2019. Neural text simplification in low-resource conditions using weak supervision. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 37–44, Minneapolis, Minnesota. Association for Computational Linguistics.

Kun Qian and Zhou Yu. 2019. Domain adaptive dialog generation via meta learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2639–2649, Florence, Italy. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

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

Matthew Richardson, Christopher J.C. 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, Seattle, Washington, USA. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Hongsuck Seo, Jonghoon Lee, Seokhwan Kim, Kyusong Lee, Sechun Kang, and Gary Geunbae Lee. 2012. A meta learning approach to grammatical error correction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 328–332,

Jeju Island, Korea. Association for Computational Linguistics.

Shashank Srivastava, Igor Labutov, and Tom Mitchell. 2018. Zero-shot learning of classifiers from natural language quantification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 306–316, Melbourne, Australia. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019a. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019b. Improving machine reading comprehension with general reading strategies. 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 2633–2643, Minneapolis, Minnesota. Association for Computational Linguistics.

Shuohang Wang, Mo Yu, Jing Jiang, and Shiyu Chang. 2018. A co-matching model for multi-choice reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 746–751, Melbourne, Australia. Association for Computational Linguistics.

Jiawei Wu, Wenhan Xiong, and William Yang Wang. 2019. Learning to learn and predict: A meta-learning approach for multi-label classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4345–4355, Hong Kong, China. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*, pages 5754–5764, NewYork, USA. Curran Associates, Inc.

Jianxing Yu, Zhengjun Zha, and Jian Yin. 2019. Inferential machine comprehension: Answering questions by recursively deducing the evidence chain from text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2241–2251, Florence, Italy. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Zhenjie Zhao and Xiaojuan Ma. 2019. Text emotion distribution learning from small sample: A meta-learning approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3948–3958, Hong Kong, China. Association for Computational Linguistics.