Learn to Cross-lingual Transfer with Meta Graph Learning Across Heterogeneous Languages*

Zheng Li[1], **Mukul Kumar**[2], **William Headden**[2], **Bing Yin**[2], **Ying Wei**[1], **Yu Zhang**[3], **Qiang Yang**[1]
[1]Hong Kong University of Science and Technology
[2]Amazon.com Inc
[3]Department of Computer Science and Engineering,
Southern University of Science and Technology
{zlict, qyang}@cse.ust.hk, {mrajk, headdenw, alexbyin}@amazon.com
{yweiad, yu.zhang.ust}@gmail.com

Abstract

The recent emergence of multilingual pre-training language model (mPLM) has enabled breakthroughs on various downstream cross-lingual transfer (CLT) tasks. However, mPLM-based methods usually involve two problems: (1) simply fine-tuning may not adapt general-purpose multilingual representations to be task-aware on low-resource languages; (2) ignore how cross-lingual adaptation happens for downstream tasks. To address the issues, we propose a meta graph learning (MGL) method. Unlike prior works that transfer from scratch, MGL can learn to cross-lingual transfer by extracting meta-knowledge from historical CLT experiences (tasks), making mPLM insensitive to low-resource languages. Besides, for each CLT task, MGL formulates its transfer process as information propagation over a dynamic graph, where the geometric structure can automatically capture intrinsic language relationships to guide cross-lingual transfer explicitly. Empirically, extensive experiments on both public and real-world datasets demonstrate the effectiveness of the MGL method.

1 Introduction

The diversity of human languages is a critical challenge for natural language processing. To alleviate the cost in annotating data for each task in each language, cross-lingual transfer (CLT) (Yarowsky et al., 2001), aiming to leverage knowledge from source languages that are sufficiently labeled to improve the learning in a target language with little supervision, has become a promising direction.

To bridge the gaps between languages, numerous CLT algorithms have emerged, ranging from early translation-based methods (Prettenhofer and Stein, 2010), cross-lingual word representation learning (Conneau et al., 2018a), to powerful mPLM (Devlin et al., 2019; Lample and Conneau, 2019), from which the versatile multilingual representations derived suffice it to become a mainstream approach for various downstream CLT tasks. However, existing mPLM-based methods focus on designing costly model pre-training while ignoring equally crucial downstream adaptation. With simply fine-tuning on the downstream labeled data for CLT tasks, mPLM often underperforms on low-resource target languages, especially for the languages distant from the source ones since the generalization ability of mPLM highly relies on lexical overlap across languages (Huang et al., 2019). On the other hand, existing adaptation approaches for mPLM behave as a black box without explicitly identifying intrinsic language relations.

To address the issues, we propose meta graph learning (MGL), a meta learning framework to learn how to cross-lingual transfer for mPLM. Specifically, MGL models each CLT process as heterogeneous information propagation over a dynamic graph, which captures latent language correlations and makes the downstream CLT adaptation more interpretable. However, solely learning the

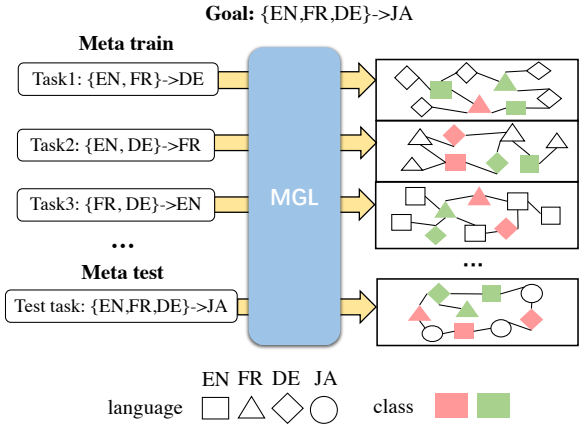


Figure 1: The meta learning process of the MGL.

The work was done when Zheng Li was an intern at Amazon.com Inc. We thank the support of Hong Kong CERG grants (16209715 & 16244616) and NSFC 61673202.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2290–2301,
November 16–20, 2020.

dynamic graph structures may be insufficient since the graph-based metric space usually favors high-resource languages over low-resource ones.

Meta learning, a.k.a learning to learn, (Finn et al., 2017; Snell et al., 2017), addresses the few-shot problems, by extracting common meta-knowledge from previous tasks (**Meta-train**) that can be rapidly adapted to new tasks with a few examples (**Meta-test**). Inspired by it, MGL takes advantage of historical CLT experiences to quickly adapt the dynamic graphs to our target CLT task. This enables MGL to meta-learn a graph-based cross-lingual metric space that is invariant across languages. For example, suppose we transfer from *English* (**EN**), *French* (**FR**) and *German* (**DE**) to *Japanese* (**JA**), i.e., {**EN**, **FR**, **DE**}→**JA**. We construct previous CLT experiences by *leave-one-out* among source languages: for each source CLT task, we leave one out of source languages as a *pseudo-target language* in turn and use the remaining ones as the *pseudo-source languages*. As such, we expect the MGL can borrow knowledge from source CLT pairs: {**FR**, **DE**}→**EN**, {**EN**, **DE**}→**FR**, {**EN**, **FR**}→**DE** to improve the transfer effectiveness in the target CLT pair {**EN**, **FR**, **DE**}→**JA**, which is illustrated in Figure 1.

Recently, some efforts have been initiated on meta learning for low-resource NLP tasks that straightforwardly views each dataset with its objective as a task (Dou et al., 2019). However, this strategy can only make meta-leaner learn knowledge from each language separately. And meanwhile, most existing meta-learners lack the ability to handle tasks lying in different distributions, especially tasks for heterogeneous languages. On the contrary, MGL resorts to learning how to adapt across languages from each CLT task. Empirically, extensive experiments on both the public multilingual Amazon review dataset (Prettenhofer and Stein, 2010) and the real-world industrial multilingual search relevance dataset (Ahuja et al., 2020) demonstrate the effectiveness of the MGL method.

Overall, our contributions can be summarized as follows: (1) A novel MGL method is proposed to *learn to cross-lingual transfer* (L2CLT) for task-aware adaptation of mPLM by leveraging previous CLT experiences; (2) The MGL automatically captures intrinsic correlations between languages, which improves the interpretability of the downstream adaptation process; (3) Extensive experiments verify the effectiveness of the MGL.

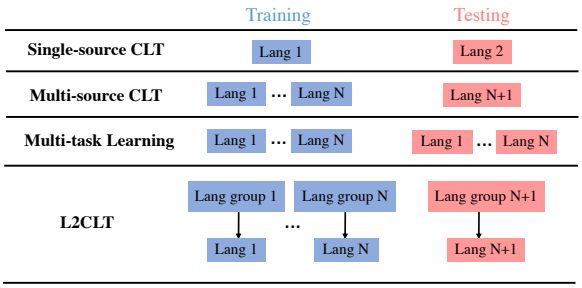


Figure 2: Differences between our work and other existing methods. "Lang" refers to the language.

2 Related Work

2.1 Cross-lingual Transfer

Most CLT studies focus on transferring from a single source language (Wan, 2009; Prettenhofer and Stein, 2010; Zhou et al., 2016b,a; Xu and Yang, 2017; Chen et al., 2018). However, single-source CLT methods would incur the risk of negative transfer when there exists a large language shift. Alternately, multi-source CLT (McDonald et al., 2011; Xu and Wan, 2017; Chen et al., 2019), transferring from multiple source languages, has been proved to increase the stability of the transfer. Another research efforts made on cross-lingual word representation learning (Zou et al., 2013; Mikolov et al., 2013; Conneau et al., 2018a; Chen and Cardie, 2018; Artetxe et al., 2018) and mPLM (Devlin et al., 2019; Lample and Conneau, 2019; Yang et al., 2019; Eisenschlos et al., 2019; Chidambaram et al., 2019), which exploit unsupervised learning on large-scale multilingual corpus to learn versatile multilingual contextualized embeddings.

2.2 Meta Learning

There are mainly three categories of meta learning: (1) Black-box amortized methods (Andrychowicz et al., 2016; Ravi and Larochelle, 2017; Mishra et al., 2017) design neural meta-learners (black-box) to infer the parameters of the base learner; (2) Gradient-based methods (Finn et al., 2017; Nichol et al., 2018; Yao et al., 2019, 2020) learn a good initialization of parameters, which can be adapted to new tasks by a few steps of gradient descent; (3) Metric-based methods (Vinyals et al., 2016; Snell et al., 2017; Garcia and Bruna, 2018; Ying et al., 2018; Sung et al., 2018; Oreshkin et al., 2018; Liu et al., 2019b) learn a task-invariant distance metric. Our work is built upon the third category to learn a cross-lingual metric space rapidly adapted to the low-resource language. Recently, some ef-

forts have been initiated on meta learning for low-resource NLP applications, such as few-shot text classification classification (Sun et al., 2019; Gao et al., 2019; Geng et al., 2019; Bao et al., 2020), natural language understanding (Dou et al., 2019), and medical prediction (Zhang et al., 2019). Very few works have explored meta learning for CLT problems like machine translation (Gu et al., 2018) and cross-lingual named entity recognition (Wu et al., 2019). However, these methods simply combine the MAML (Finn et al., 2017) or its variants for gradient optimization without considering latent relations between languages. Overall, the differences between our MGL paradigm and existing methods are illustrated in Figure 2.

3 Problem Definition

Cross-lingual Transfer Suppose that there are T high-resource (***Source***) languages $\{\ell_i^s\}_{i=1}^T$. Each source language ℓ_i^s has sufficient labeled data $D_{\ell_i^s} = \{\mathbf{x}_{\ell_i^s}^j, y_{\ell_i^s}^j\}_{j=1}^{|\ell_i^s|}$, where $|\ell_i^s|$ is the number of labeled data for the i-th source language ℓ_i^s. Besides, only a few labeled data $D_{\ell^t} = \{\mathbf{x}_{\ell^t}^j, y_{\ell^t}^j\}_{j=1}^{|\ell^t|}$ are available in a low-resource (***Target***) language ℓ^t, i.e., $|\ell^t| \ll |\ell_i^s|, \forall i \in [1, T]$. All languages share the same label space, i.e., the label set $\mathcal{Y}$. Our goal aims to leverage knowledge from high-resource languages $\{\ell_i^s\}_{i=1}^T$ to help the learning in the low-resource language ℓ^t, i.e., $\{\ell_i^s\}_{i=1}^T \rightarrow \ell^t$.

4 Methodology

Our framework is a language-agnostic task-aware model for CLT. On the one hand, we use the mPLM as the base encoder to calculate language-agnostic representations. On the other hand, we propose a meta graph learning (MGL) method to further guide the versatile multilingual representations to be task-aware for downstream CLT tasks.

4.1 Language-Agnostic Backbone

We employ a multilingual BERT (**mBERT**) (Devlin et al., 2019) as the language-agnostic encoder, which harnesses self-supervised learning with shared word piece tokens as the anchor across languages to produce weakly aligned multilingual representations. Our framework is quite general and can be easily compatible with any other mPLMs, e.g., XLM (Lample and Conneau, 2019), mUnicoder (Yang et al., 2019), etc. With the aid of mBERT as the standard encoder, we can demonstrate that the primary efforts come from the design of the task-aware MGL approach.

4.2 Task-aware Adaptation

In this section, we introduce some existing downstream adaptation approaches for CLT tasks.

Common approaches With the power of mPLM, some simple adaptation approaches can yield superior results for downstream CLT tasks, including **Target-Only**: It fine-tunes a mPLM with only the target low-resource language, which is usually regarded as a lower bound for reference; **Fine-tune**: It first trains a mPLM on the source languages and then fine-tunes the model on the target language; **Mix** (Liu et al., 2018): it ignores language characteristics and simply combines the labeled data from all languages to fine-tune a mPLM; **Multi-task** (Liu et al., 2019a): It consists of a shared mPLM encoder with language-specific discriminative layers for multi-task learning.

Meta approaches There are some efforts on gradient-based meta learning with BERT for low-resource NLU tasks (Dou et al., 2019), including second-order optimization-based **MAML** (Finn et al., 2017) with its first-order variants **FOMAML** and **Reptile** (Nichol et al., 2018). They view each dataset as one task, which may not be able to handle the language heterogeneity. Here, we compare with Reptile that is much faster when deployed to the heavy mPLM and has proved to achieve the best results as observed in (Dou et al., 2019).

4.3 Meta Graph Learning (MGL)

Here, we introduce the MGL that involves: (1) learning to CLT from historical CLT experiences; (2) learning correlations between languages.

4.3.1 Learn to Cross-lingual Transfer

The MGL is optimized over CLT tasks to achieve the L2CLT paradigm. The mechanism aims to learn knowledge from various source CLT pairs (**Meta-train**) to improve the transfer learning effectiveness for a target CLT pair (**Meta-test**).

Meta-train: we simulate the source CLT pairs by *leave-one-out* strategy among source languages $\{\ell_i^s\}_{i=1}^T$. That is, we leave one language out from the T source languages in turn as the *pseudo-target language* ℓ^{tst}, using the remaining languages as the *pseudo-source languages* to constitute a source CLT pair $p^s : \{\ell_i^s\}_{i=1}^T \backslash \ell^{\text{tst}} \rightarrow \ell^{\text{tst}}$. In total, we can obtain T source CLT pairs $\{p_o^s\}_{o=1}^T$. **Meta-test**: we directly use $\{\ell_i^s\}_{i=1}^T \rightarrow \ell^t$ as the target CLT pair

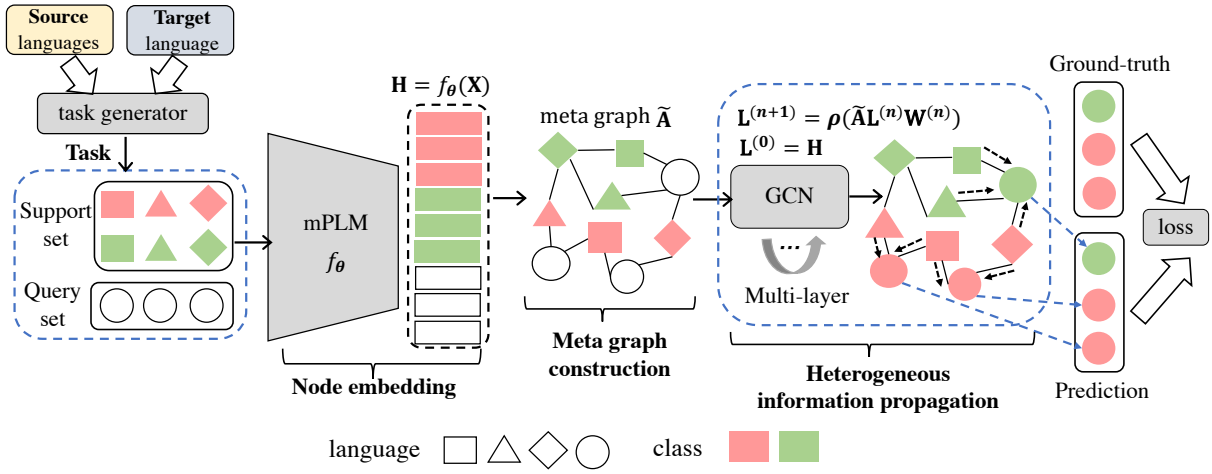


Figure 3: The framework of the proposed Meta Graph Learning (MGL) method.

p^t since the ultimate goal is to improve the learning in the low-resource language ℓ^t.

We follow the effective episodic training strategy (Vinyals et al., 2016) for training a meta-learner. In each episode, we are given a CLT pair p: $\{\ell_i^{\text{trn}}\}_{i=1}^M \rightarrow \ell^{\text{tst}}$, where $\{\ell_i^{\text{trn}}\}_{i=1}^M$ denotes M source languages ($M=T-1$ for Meta-train and $M=T$ for Meta-test) and the ℓ^{tst} is the target language. We then use the label set $\mathcal{Y}$ to randomly sample a *support set* $\mathcal{S}$ and a *query set* $\mathcal{Q}$ from the CLT pair p. The support set $\mathcal{S}$ includes $|\mathcal{Y}|$ different classes and each class contains M source languages, each of which consists of N randomly sampled instances, i.e., $\mathcal{S}=\{\mathbf{x}_j^{\mathcal{S}}, y_j^{\mathcal{S}}\}_{j=1}^{|\mathcal{Y}|\times M\times N}$.While the query set $\mathcal{Q}=\{\mathbf{x}_j^{\mathcal{Q}}, y_j^{\mathcal{Q}}\}_{j=1}^{R}$ includes R different examples of the target language from the same $|\mathcal{Y}|$ classes. $\mathcal{S}$ in each episode serves as the labeled training set on which the model is trained to minimize the loss of its predictions for $\mathcal{Q}$. In the following, we introduce how to organize $\mathcal{S}$ and $\mathcal{Q}$ into a dynamic meta graph and propagate knowledge over it for CLT as illustrated in Figure 3.

4.3.2 Node Embedding

Given a support set $\mathcal{S}$ and a query set $\mathcal{Q}$ of a sampled CLT task, we regard each instance as a node and employ the mPLM, i.e., mBERT, to extract the feature representation of each instance $\mathbf{x}_j \in \mathcal{S}\cup\mathcal{Q}$. Formally, let $\mathbf{h}_j = f_{\boldsymbol{\theta}}(\mathbf{x}_j;\boldsymbol{\theta}) \in \mathbb{R}^{dim_h}$ denote the output representation, where $\boldsymbol{\theta}$ indicates the parameters of the encoder. Then we stack all $\mathbf{h}_j$ to obtain the node embedding matrix $\mathbf{H}=\{\mathbf{h}_j\}_{j=1}^{|\mathcal{S}|+|\mathcal{Q}|}$.

4.3.3 Meta Graph Construction

To capture intrinsic correlations between languages, we propose a meta graph construction module to build their manifold structure. A *meta graph* $G=\{V,E\}$ is dynamically learned using the sampled $\mathcal{S}\cup\mathcal{Q}$ in each episode, where V ($|V| = |\mathcal{S}| + |\mathcal{Q}|$) and E denote the sets of nodes and edges, respectively. Thus, each meta graph corresponds to a sampled task's geometric formulation from the given CLT pair in meta learning. In the meta graph G, each instance is regarded as a node. The weights $\mathbf{A}\in\mathbb{R}^{|V|\times|V|}$ of the edges are based on the similarity between their node embeddings.

It is critical to build an appropriate neighborhood graph, where the manifold structure affects the transferability among different languages. Inspired by manifold learning (Chung and Graham, 1997; Zhou et al., 2004), we choose the commonly used Radial Basis Function (RBF) $A_{jj'} = \exp(-\frac{d(\mathbf{x}_j,\mathbf{x}_{j'})}{2\sigma^2})$ to compute the similarity, where d is a distance metric function, i.e., the squared Euclidean distance, and σ is a length-scale parameter. The graph structure behaves differently with respect to various σ. To avoid carefully tuning σ, we propose to instance-wisely learn the scale parameter such that it can be tailored to different language compositions. Specifically, we feed the embedding of each instance $\mathbf{x}_j \in \mathcal{S}\cup\mathcal{Q}$ into a fully-connected layer as

$$\sigma_j = \text{sigmoid}(\mathbf{W}_\sigma f_{\boldsymbol{\theta}}(\mathbf{x}_j) + \mathbf{b}_\sigma), \quad (1)$$

where σ_j is an instance-wise length-scale parameter, $\mathbf{W}_\sigma$ and $\mathbf{b}_\sigma$ are the weight matrix and bias. Then, the adjacency weight matrix $\mathbf{A}$ based on the learnable metric function is calculated as

$$A_{jj'} = \exp(-\frac{1}{2}d(\frac{f_{\boldsymbol{\theta}}(\mathbf{x}_j)}{\sigma_j}, \frac{f_{\boldsymbol{\theta}}(\mathbf{x}_{j'})}{\sigma_{j'}})). \quad (2)$$

We only keep the top K values in each row of $\mathbf{A}$ to retain K nearest neighbors, which makes the episodic training more efficient.

4.3.4 Heterogeneous Information Propagation

After that, we apply a graph convolutional network (GCN) (Kipf and Welling, 2016) on the meta graphs to produce more abstract node embeddings based on properties of their neighborhoods. Here, we only use the simple GCN to propagate heterogeneous information from the support set $\mathcal{S}$ (source languages) to the query set $\mathcal{Q}$ (target language). The generality of MGL makes it be easily adapt to other graph neural networks, e.g., GAT (Veličković et al., 2017). For GCN, it can capture first-order information about immediate neighbors with one-layer of graph convolution. When multiple GCN layers are stacked, higher-order neighbor information can be aggregated layer-wisely. Based on the last n-th layer, the node embeddings $\mathbf{L}^{(n+1)} \in \mathbb{R}^{|V| \times dim_{n+1}}$ at the $(n+1)$-th layer can be obtained as

$$\mathbf{L}^{(n+1)} = \rho(\tilde{\mathbf{A}}\mathbf{L}^{(n)}\mathbf{W}^{(n)}), \tag{3}$$

where $\mathbf{W}^{(n)}$ is the parameter of the layer and ρ is the LeakyRelu (Maas et al., 2013) activation function. We use the embedding matrix $\mathbf{H}$ as the initial node representations, i.e., $\mathbf{L}^{(0)} = \mathbf{H}$. $\tilde{\mathbf{A}}$ is the normalized symmetric adjacency matrix,

$$\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}, \tag{4}$$

where $\mathbf{D}$ is a diagonal degree matrix with D_{jj} to be the sum of the j-th row of $\mathbf{A}$. Stacking proper GCN layers can exploit latent propagation patterns over heterogeneous languages and meanwhile avoid extra noises. Empirically, we consider a two-layer GCN, which can capture the second-order relationships between nodes such that more underlying knowledge from source languages can be aggregated to help the prediction of the target language (e.g., $\ell_1 \rightarrow \ell_3 \rightarrow \ell_2$, where ℓ_1 and ℓ_2 are distant languages while ℓ_3 similar to both can serve as anchors for transitive transfer.) A two-layer GCN is as:

$$\begin{aligned} \mathbf{L}^{(1)} &= \rho(\tilde{\mathbf{A}}\mathbf{H}\mathbf{W}^{(0)}) \\ \mathbf{z} &= \text{Softmax}(\tilde{\mathbf{A}}\mathbf{L}^{(1)}\mathbf{W}^{(1)}), \end{aligned} \tag{5}$$

where $\mathbf{z} \in \mathbb{R}^{|V| \times |\mathcal{Y}|}$ is the probabilistic scores over the label set $\mathcal{Y}$. We let $\mathbf{z}^{\mathcal{Q}} \in \mathbb{R}^{|\mathcal{Q}| \times |\mathcal{Y}|}$ denote the last $|\mathcal{Q}|$ row of the $\mathbf{z}$ to be the query set score, where $\mathbf{z}_j^{\mathcal{Q}} = \mathbf{p}(\mathbf{y}_j^{\mathcal{Q}} | \mathbf{x}_j^{\mathcal{Q}}, \mathcal{S})$ denotes the predicted scores for j-th query instance $\mathbf{x}_j^{\mathcal{Q}}$.

4.3.5 Episodic Training

We follow the episodic training as described in Section 4.3.1 to optimize the MGL meta learner. In each episode, the training objective is to minimize the classification loss between the ground-truth labels and the predictions of the query set (target language) with the aid of the support set (source languages) for the given CLT pair:

$$\mathcal{J} = \sum_{j=1}^{|\mathcal{Q}|} \mathcal{L}(\mathbf{z}_j^{Q}, \mathbf{y}_j^{\mathcal{Q}}), \tag{6}$$

where $\mathcal{L}$ is the cross-entropy loss and $\mathbf{y}_j^{\mathcal{Q}}$ is the one-hot label of the j-th query instance. All the parameters are jointly updated by the gradient descent method in an end-to-end manner during the episodic training. Through learning to CLT, meta graphs can learn a cross-lingual metric space invariant for downstream languages without suffering from overfitting to the low-resource one.

5 Experiment

In this section, we present an extensive set of experiments across two datasets. The first experiment is on a public multilingual Amazon review dataset (Prettenhofer and Stein, 2010). In addition, we conduct experiments on a real-world industrial multilingual search relevance dataset (Ahuja et al., 2020) used for E-commerce product search.

5.1 Cross-lingual Sentiment Classification

Dataset The aim of the task is binary sentiment classification, where each review document is classified into positive or negative sentiment. We use the multilingual Amazon review dataset (Prettenhofer and Stein, 2010), which has four languages: English (**EN**), German (**DE**), French (**FR**) and Japanese (**JA**) on three domains: *Books*, *DVD* and *Music*. For statistics, the sizes of the training, validation, and testing data are 1600, 400 and 2000, respectively, for each language of all the domains.

Setting We treat each domain as separate experiments and consider FR, JA, DE as the target language (Here, we do not consider EN since it is usually high-resource) while the remaining three being source languages, which results in 9 total cross-lingual experiments. In the low-resource setting, we only use ***10%*** labeled training data for the target language, i.e., *160 labeled data*. The evaluation metric is Accuracy. All experiments are

Model	EN+JA+DE→FR				EN+FR+DE→JA				EN+FR+JA→DE				*Avg*	Δ
	books	dvd	music	*avg*	books	dvd	music	*avg*	books	dvd	music	*avg*		
methods with cross-lingual parallel data														
MT-BOW	80.76	78.83	75.78	78.46	70.22	71.30	72.02	71.18	79.68	77.92	77.22	78.27	75.97	+(5.36)
CL-SCL	78.49	78.80	77.92	78.40	73.09	71.07	75.11	73.09	79.50	76.92	77.79	78.07	76.52	+(4.81)
CL-RL	78.25	74.83	78.71	77.26	71.11	73.12	74.38	72.87	79.85	77.14	77.27	78.10	76.08	+(5.25)
methods w/o cross-lingual parallel data														
BWE	77.95	79.25	79.95	79.05	54.78	54.20	51.30	53.43	78.35	77.45	76.70	77.50	69.99	+(11.34)
MAN-MoE	81.10	84.25	80.90	82.08	62.78	69.10	72.60	68.16	78.80	77.15	79.45	79.45	76.56	+(4.77)
Lower bound														
mBERT+Target-Only	74.85	72.90	76.80	74.85	69.48	65.75	73.30	69.51	72.50	66.75	73.05	70.77	71.71	+(9.62)
multilingual transfer														
mBERT+Mix	83.05	83.15	81.20	82.47	75.09	75.00	75.90	75.33	81.05	78.35	78.15	78.99	78.93	+(2.40)
mBERT+Multi-task	83.80	81.50	82.40	82.57	75.99	73.20	75.65	74.95	78.70	73.45	80.65	77.60	78.37	+(2.96)
mBERT+Fine-tune	83.00	83.40	81.45	82.62	75.04	73.80	76.80	75.21	81.50	79.40	78.65	79.85	79.23	+(2.10)
Meta learning														
mBERT+Reptile	**84.55**	83.50	81.10	83.05	74.89	73.15	77.85	75.30	81.65	78.55	80.20	80.13	79.49	+(1.84)
mBERT+**MGL**	83.97	**85.07**†	**83.16**†	**84.07**†	**77.41**†	**77.20**†	**78.59**†	**77.73**†	**82.22**†	**81.30**†	**83.01**†	**82.18**†	**81.33**†	-

Table 1: Experimental results (%) on the multilingual Amazon review dataset. Δ refers to the improvements. † means that the MGL significantly outperforms the best baseline **Reptile** with paired sample t-test p-value < 0.01.

repeated 5 times, and we report the average results.

Baselines In addition to mBERT-based baselines mentioned in Section 4.2, we also compare with the state-of-the-art CLT baselines:

- **MT-BOW** uses machine translation to translate the bag of words of a target language into the source language.
- **CL-SCL** (Prettenhofer and Stein, 2010) learns a shared cross-lingual feature space with cross-lingual structural correspondence learning.
- **CL-RL** (Xiao and Guo, 2013) learns cross-lingual representation learning, where part of the word vector is shared among languages.
- **BWE** (Upadhyay et al., 2018) bridges the language gap with Bilingual Word Embedding and weight sharing. We use the unsupervised MUSE (Conneau et al., 2018a) BWE.
- **MAN-MoE** (Chen et al., 2019) exploits both language-invariant and language-specific features with multinomial adversarial training and mixture-of-experts, respectively.

Results Based on the results in Table 1, we can summarize the following observations:

- The MGL achieves the best results on most transfer pairs, significantly outperforming the best baseline Reptile by 1.84% accuracy on average. Our model exceeds the translation-based methods MT-BOW, CL-SCL, and CL-RL by 5.36%, 4.81% and 5.25% accuracy on average, respectively. Without additional translation resources, the embedding-based method BWE shows significant performance degradation. Though MAN-MoE attempts to fully identify both invariant and specific language features, it can only achieve competitive results with translation-based methods. This proves that language correspondences play a critical role in minimizing language gaps. However, obtaining general-purpose alignment usually relies on off-the-shelf translators, e.g., *Google Translate*, making them inflexible and unscalable to the big data.
- Our method achieves 2.40%, 2.96%, and 2.10% average accuracy gains over mBERT with common adaption approaches, i.e., Mix, Multi-task, and Fine-tune, respectively. These methods ignore task-aware adaptation for low-resource target languages and perform poorly for the adaptation between distant languages. On the contrary, our method can effectively alleviate the language gaps by meta-learning previous CLT knowledge.
- Though combining the strengths of both mBERT and meta learning, Reptile can only achieve marginal improvements over common adaptation methods since language heterogeneity hinders the effectiveness of this gradient-based meta learner to adapt across different languages. Differently, our model can alleviate the issue by learning meta graphs over languages to reduce the gaps between them.

5.2 Cross-lingual Relevance Classification

Dataset The task aims to determine the binary relevance label (relevant or irrelevant) of a pair of user search query and product title. We use a large-scale multilingual search relevance dataset (Ahuja

Model	ES+IT+EN+DE→FR			FR+IT+EN+DE→ES			FR+ES+EN+DE→IT			FR+ES+IT+EN→DE			Avg	Δ
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	F1	F1
methods with task-specific cross-lingual parallel data														
LAPS	44.06	51.79	47.61	52.35	48.74	50.48	48.10	52.38	50.15	37.15	40.68	38.83	46.77	+(4.61)
Lower bound														
mBERT+Target-Only	38.79	40.09	39.43	40.57	49.82	44.72	42.48	57.87	48.99	32.70	35.30	33.95	41.77	+(9.61)
multilingual transfer														
mBERT+Mix	39.57	74.42	51.67	39.44	76.2	52.02	36.06	78.88	49.49	32.56	55.45	41.03	48.55	+(2.83)
mBERT+Multi-task	50.21	50.52	50.37	51.04	49.36	50.19	48.89	57.80	52.97	31.35	53.63	39.57	48.28	+(3.10)
mBERT+Fine-tune	47.63	54.61	50.88	47.17	59.35	52.56	44.89	61.52	51.91	30.88	60.91	40.98	49.08	+(2.30)
Meta learning														
mBERT+Reptile	44.02	65.88	52.77	43.32	70.51	53.67	49.11	58.01	53.19	34.28	50.05	40.69	50.15	+(1.23)
mBERT+**MGL**	50.94	59.24	**54.78**†	45.97	66.06	**54.22**†	48.13	62.33	**54.32**†	38.51	46.64	**42.19**†	**51.38**†	-

Table 2: Experimental results (%) on the multilingual search relevance dataset. Δ refers to the improvements. † means that the MGL significantly outperforms the best baseline **Reptile** with paired sample t-test p-value < 0.01.

Target language	FR	JA	DE	*Avg*	Δ
mBERT+MGL (**Full model**)	84.07	**77.73**	**82.18**	**81.33**	-
mBERT+MGL w/o Meta	83.94	72.86	78.66	78.49	+2.84
mBERT+MGL w/o L2CLT	83.47	75.46	80.69	79.87	+1.46
mBERT+MGL w/o σ	84.14	76.55	80.99	80.56	+0.77
mBERT+MGL (1 GCN layer)	**84.34**	76.45	81.39	80.73	+0.60
mBERT+MGL (3 GCN layer)	83.31	76.51	81.44	80.42	+0.91
mBERT+MGL (4 GCN layer)	83.43	74.09	80.81	79.44	+1.89

Table 3: Ablation results (%): averaged accuracy for each target language on the Amazon review dataset.

et al., 2020), which arises from 5 languages: French (**FR**), Spanish (**ES**), Italian (**IT**), English (**EN**) and German (**DE**). The human-annotated query-product pairs are sampled from the search results from each of the above country-specific services of an E-commerce search engine. The annotators return a binary label that indicates the relevance of the product item to the query.

Setting We use the same setting as described in Section 5.1. Considering the imbalance of the dataset, we use Precision (P), Recall (R), and F1 score as the evaluation metrics.

Baselines Additionally, we compare with the start-of-the-art baseline **LAPS** (Ahuja et al., 2020), which relies on external task-specific cross-lingual parallel data (Ahuja et al., 2020), i.e., product-to-product and query-to-query correspondences among all 5 languages.

Results Based on the results in Table 2, we can summarize the following observations:

- For the imbalanced industrial dataset with more noises, the MGL method consistently achieves the best results for all pairs, significantly exceeding the best baseline Reptile by 1.23% F1 score on average. The efficacy of the LAPS comes from task-specific parallel data, which is usually difficult to obtain in practice. Without any aid of task-specific resources, our MGL method can still achieve a large gain of 4.61% average F1 score by adapting general mPLM with task-aware meta-knowledge for CLT tasks.

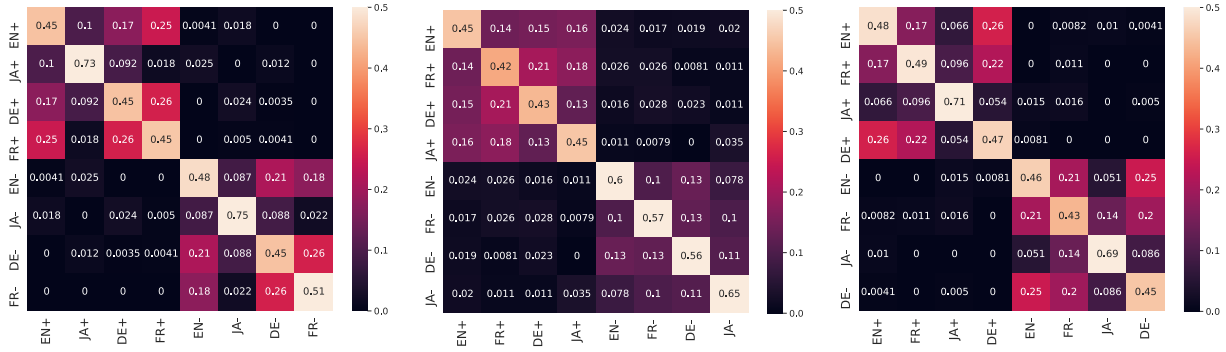
- Compared with mBERT-based methods, we can also obtain consistent observations as on the Amazon dataset. Even with the power of mBERT, common adaptation approaches still cannot handle the low-resource target language. Reptile cannot compete against our MGL method due to its meta-knowledge learned from each separate language.

5.3 Ablation Study

To verify the efficacy of each component, we compare MGL with its ablation variants in Table 3.

No Meta v.s. **Meta** For MGL w/o Meta, we directly learn the dynamic graphs with GCN for cross-lingual transfer without any meta process, i.e., no Meta-train stage. MGL exceeds MGL w/o Meta by 2.84% accuracy on average. This proves that simply adding a GCN cannot work well. Without leveraging historical CLT experiences, the dynamic graphs cannot learn a robust cross-lingual metric space that facilitates knowledge propagation to the target low-resource language.

No L2CLT v.s. **L2CLT** For MGL w/o L2CLT, we treat each language as one task like Reptile and change to sample the support set and query set from the same language for each task. As such, this MGL variant solely uses the dynamic graphs to meta-learn knowledge from each language. MGL can outperform MGL w/o L2CLT by 1.46% accuracy on average, especially obtaining more gains for distant language **JA**. The reason is that



(a) **EN+JA+DE→FR** (b) **EN+FR+DE→JA** (c) **EN+FR+JA→DE**

Figure 4: Visualization of meta graphs for different pairs of the *Music* domain on the multilingual Amazon review dataset. + and − denote positive and negative classes, respectively. Brighter colors denote higher correlations.

previous CLT experiences can benefit MGL to transfer across heterogeneous languages in a new target one. For example, for the hard {**EN, FR, DE**}→**JA**, MGL will learn the transfer skill from the comparatively distant source CLT pair: Germanic languages[1] {**EN, DE**} to Romance languages[1] **FR** for Meta-train, and then leverage the skill to rapidly adapt the meta graphs to transfer from {**EN, FR, DE**} to **JA** for Meta-test.

No σ v.s. σ For MGL w/o σ, we remove the scale factor σ in Eq. 2 that controls the learnability of the metric function. MGL outperforms MGL w/o σ by 0.77% average accuracy. This demonstrates that an learnable metric distance function can be more adaptive to measure intrinsic relations among different languages and improve the transferability of the meta graphs.

The number of GCN layers MGL consists of a multi-layer GCN to progressively propagate information across languages over the meta graphs. Multiple GCN layers are necessary to distill more latent propagation patterns. As we can see, increasing the number of GCN layers from 1 to 2 (default) shows significant improvements. However, when further increasing the number of layers to 3 and 4, the performances will be degraded. A possible reason is that more layers may bring in more noises from higher-order neighborhoods in the meta graphs, causing negative transfer.

5.4 Visualization of Meta Graph

To demonstrate that the MGL can automatically capture latent correlations among languages, we perform visualization of the meta graphs on the Amazon review dataset. We visualize the meta graph, i.e., the dynamic normalized symmetric adjacency matrix $\tilde{\mathbf{A}}$ as defined in Eq. 4. For each pair, we only randomly sample one instance for each class of each language, which constitutes a meta graph $\tilde{\mathbf{A}} \in \mathbb{R}^{8\times 8}$ on the Meta-test stage. And we set the number of the neighborhoods for the K-nearest graph to be 2. To be more convincing, we calculate the averaged meta graph obtained by averaging the accumulated meta graphs over a random sampling of 100 times.

First, as shown in Figure 4-(a), we transfer from {**EN, JA, DE**} to **FR**, our model can capture stronger connections from **FR** to {**EN, DE**} than **FR** to **JA**. This is reasonable since {**EN, DE, FR**} all belong to Indo-European languages[1], which are very dissimilar to **JA**. Second, when transferring from {**EN, FR, JA**} to **DE** as shown in Figure 4-(c), **DE** behaves more correlative to **EN** than to **FR**. The possible reason may be that **EN** and **DE** are both Germanic languages, while **FR** belongs to Romance languages[1]. Finally, in Figure 4-(b), when there exists a large gap between the source and target languages, all source languages {**EN, FR, DE**} have weak correlations with **JA**, and thus the results on the target language **JA** (77.73%) are usually worse than the target language **FR** (84.07%) or **DE** (82.18%) on average. This proves that our model can automatically exploit language relations with learnable graph structures to make task-aware adaptation more interpretable.

5.5 Target Labeled Proportion

We vary the labeled proportion of the target language's training set and compare MGL with Reptile, Mix, Multi-task, and Fine-tune based on

[1] https://simple.wikipedia.org/wiki/Language_family

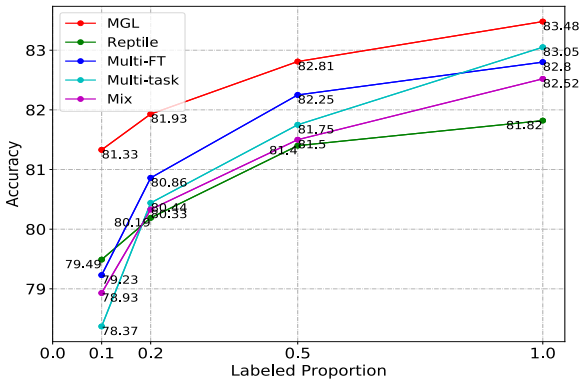


Figure 5: Averaged results w.r.t proportions of the target training data on the Amazon review dataset.

mBERT. We use averaged accuracy on the multilingual Amazon review dataset and change the labeled proportion from 0.1, 0.2, 0.5 to 1.0. As shown in Figure 5, the gap between the MGL and common adaptation methods grows as the training size shrinks, verifying that the MGL is more robust to the drop in the labeled size for the target language. Reptile shows marginal improvements over common adaptation methods and performs worse as the training size grows. This demonstrates that roughly incorporating existing meta learning algorithms into the CLT problem may not work well.

6 Implementation details

Encoder We use the *bert-base-multilingual-cased*[2] model pre-trained on 104 languages as the encoder, which has 12 layers, 768-d hidden size, 12 heads and 110M total parameters. The mBERT is jointly optimized with other parameters during both the Meta-train and Meta-test stages.

Training & Validation & Testing set For each cross-lingual transfer experiment, during the Meta-train stage, its training set is the combination of the training data of the source languages. Meanwhile, it employs the combination of validation data of the source languages as the validation set. As for the Meta-test stage, the training set is the combination of the training data of all languages. The validation data and testing data of the target language will be used for the validation and final evaluation, respectively.

Initialization & Training For all the experiments, the model is optimized by the Adam algorithm (Kingma and Ba, 2014) for training. The weight matrices are initialized with a uniform distribution $U(-0.01, 0.01)$. Gradients with the ℓ_2 norm larger than 40 are normalized to be 40. To alleviate overfitting, we apply the dropout on the node representations of the first GCN layer with the dropout rate 0.5. We also perform early stopping on the validation set during both the Meta-train and Meta-test stages.

[2] https://github.com/huggingface/transformers

Hyper-parameter	Dataset	
	Amazon Review	Search Relevance
dim_0, dim_1	768, 768	768, 768
$\#train\_episodes$	75	100
$\#test\_episodes$	10	20
$\#eval\_times$	5	5
support size N	5	5
query size R	32	64
neighborhoods K	10	100
learning rate	10^{-5}	10^{-5}

Table 4: Settings of hyper-parameters.

Hyperparameter For the multilingual Amazon review dataset, we use the same hyper-parameters, which are manually tuned on 10% randomly held-out training data of the source languages in **EN+FR+DE→JA** on the *Book* domain, for all cross-lingual transfer experiments. As for the multilingual search relevance dataset, the hyper-parameters are manually tuned on 10% randomly held-out training data of the source languages in **ES+IT+EN+DE→FR** and fixed to be used in all cross-lingual experiments. The detailed hyperparameters for the two datasets are listed in Table 4.

7 Conclusion and Future Works

In this paper, we propose a novel MGL method for task-aware CLT adaptation of mPLM by leveraging historical CLT experiences. Extensive evaluations on both the public benchmark and large-scale industrial dataset quantitively and qualitatively demonstrate the effectiveness of the MGL. In the future, the proposed MGL method can potentially applied to more cross-lingual natural language understanding (XLU) tasks (Conneau et al., 2018b; Wang et al., 2019; Lewis et al., 2019; Karthikeyan et al., 2020), and be generalized to learn to learn for domain adaptation (Blitzer et al., 2007), representation learning (Shen et al., 2018), multi-task learning (Shen et al., 2019) problems, etc.

References

Aman Ahuja, Nikhil Rao, Sumeet Katariya, Karthik Subbian, and Chandan K Reddy. 2020. Language-agnostic representation learning for product search on e-commerce platforms. In *WSDM*, pages 7–15.

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. 2016. Learning to learn by gradient descent by gradient descent. In *NIPS*, pages 3981–3989.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *ACL*, pages 789–798.

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2020. Few-shot text classification with distributional signatures. In *ICLR*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, pages 440–447.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. *arXiv preprint arXiv:1808.08933*.

Xilun Chen, Ahmed Hassan, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *ACL*, pages 3098–3112.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *TACL*, 6:557–570.

Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *RepL4NLP workshop, ACL*, pages 250–259.

Fan RK Chung and Fan Chung Graham. 1997. *Spectral graph theory*. 92. American Mathematical Soc.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. In *ICLR*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. Xnli: Evaluating cross-lingual sentence representations. In *EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. *arXiv preprint arXiv:1908.10423*.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. In *EMNLP-IJCNLP*, pages 5702–5707.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *AAAI*, volume 33, pages 6407–6414.

Victor Garcia and Joan Bruna. 2018. Few-shot learning with graph neural networks. In *ICLR*.

Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. 2019. Induction networks for few-shot text classification. In *EMNLP-IJCNLP*, pages 3895–3904.

Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor OK Li. 2018. Meta-learning for low-resource neural machine translation. *arXiv preprint arXiv:1808.08437*.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *EMNLP-IJCNLP*, pages 2485–2494.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *ICLR*.

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

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Qi Liu, Yue Zhang, and Jiangming Liu. 2018. Learning domain representation for multi-domain sentiment classification. In *NAACL-HLT*, pages 541–550.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *ACL*, pages 4487–4496.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sungju Hwang, and Yi Yang. 2019b. Learning to propagate labels: Transductive propagation network for few-shot learning. In *ICLR*.

Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, page 3.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *EMNLP*, pages 62–72.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2017. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*.

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

Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. 2018. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NIPS*, pages 721–731.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *ACL*, pages 1118–1127.

Sachin Ravi and Hugo Larochelle. 2017. Optimization as a model for few-shot learning. In *ICLR*.

Tao Shen, Xiubo Geng, Tao Qin, Daya Guo, Duyu Tang, Nan Duan, Guodong Long, and Daxin Jiang. 2019. Multi-task learning for conversational question answering over a large-scale knowledge base. In *EMNLP-IJCNLP*, pages 2442–2451.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *AAAI*, pages 5446–5455.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *NIPS*, pages 4077–4087.

Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. 2019. Hierarchical attention prototypical networks for few-shot text classification. In *EMNLP-IJCNLP*, pages 476–485.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *CVPR*, pages 1199–1208.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *ICASSP*, pages 6034–6038.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *NIPS*, pages 3630–3638.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *ACL-IJCNLP*, pages 235–243.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual bert transformation for zero-shot dependency parsing. *arXiv preprint arXiv:1909.06775*.

Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F Karlsson, Biqing Huang, and Chin-Yew Lin. 2019. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. *arXiv preprint arXiv:1911.06161*.

Min Xiao and Yuhong Guo. 2013. Semi-supervised representation learning for cross-lingual text classification. In *EMNLP*, pages 1465–1475.

Kui Xu and Xiaojun Wan. 2017. Towards a universal sentiment classifier in multiple languages. In *EMNLP*, pages 511–520.

Ruochen Xu and Yiming Yang. 2017. Cross-lingual distillation for text classification. In *ACL*, pages 1415–1425.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

Huaxiu Yao, Ying Wei, Junzhou Huang, and Zhenhui Li. 2019. Hierarchically structured meta-learning. In *ICML*.

Huaxiu Yao, Xian Wu, Zhiqiang Tao, Yaliang Li, Bolin Ding, Ruirui Li, and Zhenhui Li. 2020. Automated relational meta-learning. In *ICLR*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Wei Ying, Yu Zhang, Junzhou Huang, and Qiang Yang. 2018. Transfer learning via learning to transfer. In *ICML*, pages 5085–5094.

Xi Sheryl Zhang, Fengyi Tang, Hiroko H Dodge, Jiayu Zhou, and Fei Wang. 2019. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In *SIGKDD*, pages 2487–2495.

Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. 2004. Learning with local and global consistency. In *NIPS*, pages 321–328.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016a. Attention-based lstm network for cross-lingual sentiment classification. In *EMNLP*, pages 247–256.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016b. Cross-lingual sentiment classification with bilingual document representation learning. In *ACL*, pages 1403–1412.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, pages 1393–1398.