Cross-lingual Text Classification with Heterogeneous Graph Neural Network

Ziyun Wang[1*], **Xuan Liu**[2*†], **Peiji Yang**[1] , **Shixing Liu**[1] , **Zhisheng Wang**[1]
Tencent, Shenzhen, China
[1]{billzywang, peijiyang, shixingliu, plorywang}@tencent.com
[2]lxstephenlaw@gmail.com

Abstract

Cross-lingual text classification aims at training a classifier on the source language and transferring the knowledge to target languages, which is very useful for low-resource languages. Recent multilingual pretrained language models (mPLM) achieve impressive results in cross-lingual classification tasks, but rarely consider factors beyond semantic similarity, causing performance degradation between some language pairs. In this paper we propose a simple yet effective method to incorporate heterogeneous information within and across languages for cross-lingual text classification using graph convolutional networks (GCN). In particular, we construct a heterogeneous graph by treating documents and words as nodes, and linking nodes with different relations, which include part-of-speech roles, semantic similarity, and document translations. Extensive experiments show that our graph-based method significantly outperforms state-of-the-art models on all tasks, and also achieves consistent performance gain over baselines in low-resource settings where external tools like translators are unavailable.

1 Introduction

The success of recent deep learning based models on text classification relies on the availability of massive labeled data (Conneau et al., 2017; Tian et al., 2020; Guo et al., 2020). However, labeled data are usually unavailable for many languages, and hence researchers have developed the setting where a classifier is only trained using a resource-rich language and applied to target languages without annotated data (Xu et al., 2016; Chen and Qian, 2019; Fei and Li, 2020). The biggest challenge is to bridge the semantic and syntactic gap between

languages. Most existing methods explore the semantic similarity among languages, and learn a language-agnostic representation for documents from different languages (Chen et al., 2018; Zhang et al., 2020a). This includes recent state-of-the-art multilingual pretrained language models (mPLM) (Devlin et al., 2019; Conneau and Lample, 2019), which pretrain transformer-based neural networks on large-scale multilingual corpora. The mPLM methods show superior cross-lingual transfer ability in many tasks (Wu and Dredze, 2019). However, they do not explicitly consider syntactic discrepancy between languages, which may lead to degraded generalization performance on target languages (Ahmad et al., 2019; Hu et al., 2020).

On the other hand, there usually exists sufficient unlabeled target-language documents that come naturally with rich information about the language and the task. However, only a handful of previous researches have taken advantage of the unlabeled data (Wan, 2009; Dong and de Melo, 2019).

To integrate both semantic and syntactic information within and across languages, we propose a graph-based framework named Cross-Lingual Heterogeneous GCN (CLHG). Following the work of TextGCN (Yao et al., 2019), we represent all the documents and words as graph nodes, and add different types of information into the graph. We utilize mPLM to calculate the representation of all the nodes, and connect documents nodes with semantically similar ones to extract the knowledge in mPLM. Words are connected with documents based on the co-occurrences as in previous works. However, we choose to separate different word-doc edges by part-of-speech (POS) tags of words to inject some shallow syntactic information into the graph, as POS taggers are one of the most widely accessible NLP tools, especially for low-resource languages. In-domain unlabeled documents are added to the graph if available. To further absorb
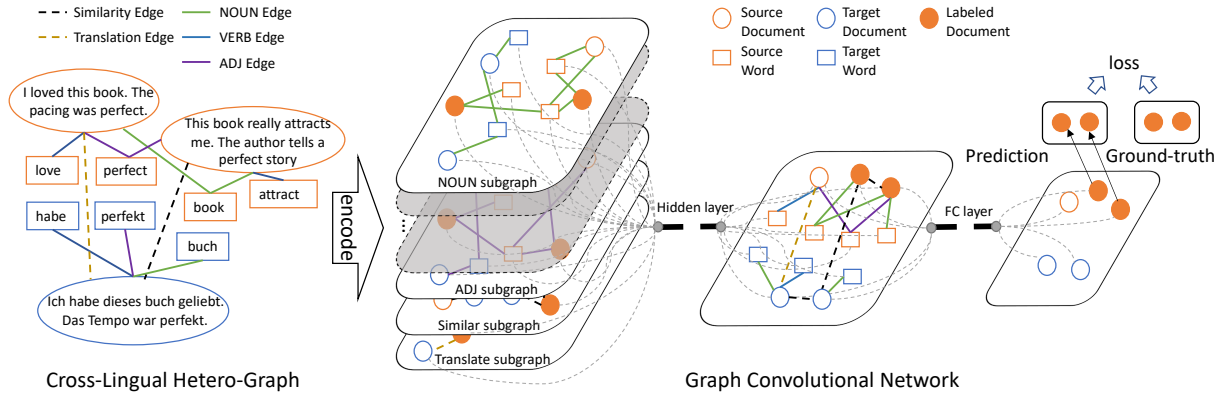

Figure 1: Illustration of our Cross-Lingual Heterogeneous GCN (CLHG) framework. For simplicity, only some POS tags are plotted in this graph. We recommend to view this figure in color as we use different colors to indicate different languages and edge types.

in-domain language alignment knowledge, we utilize machine translation to create translated text nodes. The text classification task is then formalized as node classification in the graph and solved with a heterogeneous version of Graph Convolutional Networks (Kipf and Welling, 2017). Our contributions are summarized as follows:

(1) We propose a graph-based framework to easily comprise heterogeneous information for cross-lingual text classification, and design multiple types of edges to integrate all these information. To the best of our knowledge, this is the first study to use heterogeneous graph neural networks for cross-lingual classification tasks. (2) We conduct extensive experiments on 15 tasks from 3 different datasets involving 6 language pairs. Results show that our model consistently outperforms state-of-the-art methods on all tasks without any external tool, and achieves further improvements with the help of part-of-speech tags and translations.

2 Related Works

Traditional methods for cross-lingual classification usually translate the texts (Wan, 2009) or the classifier model (Xu et al., 2016) with external aligned resources such as bilingual dictionaries (Andrade et al., 2015; Shi et al., 2010) or parallel corpora (Duh et al., 2011; Zhou et al., 2016; Xu and Yang, 2017). Recent works focus on learning a shared representation for documents of different languages, including bilingual word embeddings (Zou et al., 2013; Ziser and Reichart, 2018; Chen et al., 2018), common subword representations (Zhang et al., 2020a), and multilingual pretrained language models (mPLM) (Devlin et al., 2019; Conneau and Lam-

ple, 2019; Clark et al., 2020).

In the past few years, graph neural networks (GNN) have attracted wide attention, and become increasingly popular in text classification (Yao et al., 2019; Hu et al., 2019; Ding et al., 2020; Zhang et al., 2020b). These existing work mainly focus on monolingual text classification, except a recent work (Li et al., 2020) using meta-learning and graph neural network for cross-lingual sentiment classification, which nevertheless only uses GNN as a tool for meta-learning.

3 Method

In this section, we will introduce our CLHG framework, including how to construct the graph and how to solve cross-lingual text classification using heterogeneous GCN. In general, we first construct a cross-lingual heterogeneous graph based on the corpus and selected features, and next we encode all the texts with multilingual pre-trained language models, then we pass the encoded nodes to the heterogeneous GCN, each layer of which performs graph convolution on different subgraphs separated by different edge types, and aggregates the information together. Finally, the graph neural network outputs the predictions of doc nodes, which will be compared with groundtruth labels during training. Figure 1 shows the overall structure of the framework.

3.1 Graph Construction

Inspired by some previous works on GNN-based text classification (Yao et al., 2019; Hu et al., 2019), we construct the graph by representing both documents and words from the corpus in both languages

613

as graph nodes, and augment the corpus by including unlabeled in-domain documents from the target language. To extract more information of language alignments, we further use a publicly available machine translation API[1] to translate the documents in both directions. Then two categories of edges are defined in the graph.

Doc-word Edges. Like TextGCN (Yao et al., 2019), documents and words are connected by their co-occurrences. To inject syntactic information more than just co-occurrences, we add part-of-speech (POS) tags to the edges, since different POS roles have different importance in the classification tasks. Adjectives and adverbs are mostly decisive in sentiment classification, while nouns may play a more significant role in news classification. Therefore, we use POS taggers to tag each sentence and create different types of edges based on the POS roles of the words in the document, which could help GNN to learn different propagation patterns for each POS role.

Doc-doc Edges. To add more direct connections between documents, we include two types of document level edges. Firstly, we link each document with similar ones by finding K documents with the largest cosine similarity. The embeddings of the documents are calculated using mPLM. Secondly, we connect nodes created by machine translation with their original texts.

3.2 Heterogeneous Graph Convolution

After building the heterogeneous cross-lingual graph, we first encode all the nodes using mPLM by directly inputting the text to the mPLM and taking the hidden states of the first token. The encoded node features are fixed during training. Next we apply heterogeneous graph convolutional network (Hetero-GCN) (Hu et al., 2019) on the graph to calculate higher-order representations of each node with aggregated information.

Heterogeneous GCN applies traditional GCN on different sub-graphs separated by different types of edges and aggregates information to an implicit common space.

$$H^{(l+1)} = \sigma(\sum_{\tau \in \mathcal{T}} \tilde{A}_\tau \cdot H_\tau^{(l)} \cdot W_\tau^{(l)}) \qquad (1)$$

where $\tilde{A}_\tau$ is a submatrix of the symmetric normalized adjacency matrix that only contains edges with

type τ, $H_\tau^{(l)}$ is the feature matrix of the neighboring nodes with type τ of each node, and $W_\tau^{(l)}$ is a trainable parameter. $\sigma(\cdot)$ denotes a non-linear activation function, which we use leaky ReLU. Initially, $H_\tau^{(0)}$ is the node feature calculated by mPLM.

Empirically, we use two graph convolution layers to aggregate information within second-order neighbors. Then a linear transformation is applied to the document nodes to get the predictions.

4 Experiments

We evaluate our framework on three different classification tasks, including Amazon Review sentiment classification (Prettenhofer and Stein, 2010), news category classification from XGLUE (Liang et al., 2020), and intent classification on a multilingual spoken language understanding (SLU) dataset (Schuster et al., 2019). More details of each dataset is provided in the appendix. For all the tasks, we use only the English samples for training and evaluate on other 6 languages, which are German (DE), French (FR), Russian (RU), Spanish (ES), Japanese (JA), and Thai (TH).

4.1 Experiment Setting

In all our experiments, we use two-layer GCN with hidden size 512 and output size 768. Each document is connected with 3 most similar documents. The model is trained using the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 2×10^{-5} and batch size 256. We train the GCN for at most 15 epochs and evaluate the model with best performance on validation set. XLM-RoBERTa (Conneau et al., 2020) is used to encode all the documents and words, which is finetuned on the English training set of each task for 2 epochs with batch size 32 and learning rate 4×10^{-5}. We set the max length as 128 for intent classification, and 512 for the other two tasks. Each experiment is repeated 3 times and the average accuracy is reported. All the experiments are conducted on an NVIDIA V100 GPU [2].

For part-of-speech tagging, we adopt different taggers for each language [3] and map all the tags to

---

[1] https://cloud.tencent.com/document/api/551/15619

[2] Our codes are available at https://github.com/TencentGameMate/gnn_cross_lingual.

[3] We choose Stanford POS Tagger (Toutanova et al., 2003) for EN, Spacy (https://spacy.io) for DE, FR and ES, MeCab (https://taku910.github.io/mecab/) for JA, tltk (https://pypi.org/project/tltk/) for TH, and nltk (https://www.nltk.org) for RU.

Method	EN → DE				EN → FR				EN → JA			
	books	dvd	music	avg	books	dvd	music	avg	books	dvd	music	avg
CLDFA	83.95	83.14	79.02	82.04	83.37	82.56	83.31	83.08	77.36	80.52	76.46	78.11
MVEC	88.41	87.32	89.97	88.61	89.08	88.28	88.50	88.62	79.15	77.15	79.70	78.67
mBERT	84.35	82.85	93.85	83.68	84.55	85.85	83.65	84.68	73.35	74.80	76.10	74.75
XLM	86.85	84.20	85.90	85.65	88.10	86.95	86.20	87.08	80.95	79.20	78.02	79.39
XLM-R	91.65	87.60	90.97	90.07	89.33	90.07	89.15	89.52	85.26	86.77	86.95	86.33
CLHG	**92.70***	**88.60***	**91.62***	**90.97***	**90.67***	**91.38***	**90.45***	**90.83***	**87.21***	**87.33***	**88.08***	**87.54***

Table 1: Sentiment classification accuracy (%) on Amazon Review dataset. * shows the result is significantly better than XLM-R baseline with p-value ≤ 0.05.

Method	DE	FR	ES	RU
mBERT	82.6	78.0	81.6	79.0
XLM-R	84.5	78.2	83.2	79.4
Unicoder	84.2	78.5	83.5	79.7
XLM-R (ours)	83.99	78.66	83.27	80.42
CLHG	**85.00**[+]	**79.58***	**84.80***	**80.91**[+]

Table 2: Classification accuracy (%) on XGLUE News Classification. We re-run the XLM-R model and also report our reproduced results. * shows the result is significantly better than XLM-R baseline with p-value ≤ 0.05, and [+] indicates p-value ≤ 0.1.

Method	EN → ES	EN → TH
CoSDA_ML+mBERT	94.80	76.80
CoSDA_ML+XLM	90.30	86.70
mBERT	74.91	42.97
XLM	62.30	31.60
XLM-R	94.38	85.17
CLHG	**96.81***	**89.71***

Table 3: Intent classification accuracy (%) on multilingual SLU dataset. * shows the result is significantly better than XLM-R baseline with p-value ≤ 0.05.

Universal Dependency (UD) tagset [4].

4.2 Baselines

Our method is compared with different multilingual pretrained models finetuned for each task, which include multilingual BERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019) and XLM-RoBERTa (Conneau et al., 2020), and also with published state-of-the-art results on each dataset.

Amazon Review. CLDFA (Xu and Yang, 2017) utilizes model distillation and parallel corpus to transform a model from source to target language. MVEC (Fei and Li, 2020) refines the shared latent space of mPLM with unsupervised machine translation and a language discriminator.

---

[4] https://universaldependencies.org

XGLUE News Classification. Unicoder (Huang et al., 2019) is another mPLM proposed recently, and is used as the baseline method provided alongside XGLUE benchmark.

Multilingual SLU. CoSDA-ML (Qin et al., 2020) is a data augmentation framework that automatically generates code-switching documents using a bilingual dictionary, which is used when finetuning language models on downstream tasks.

4.3 Results and Analysis

The results are provided in table 1 2 and 3 for each dataset. Our method significantly outperforms state-of-the-art baselines and achieves consistent improvements over XLM-R model. The most performance gain is achieved on the multilingual SLU dataset. Different from the other two, this dataset consists of short texts, and thus the created graph is much cleaner and more suitable for GCN to model.

To verify that the improvement does not come barely from the external data, we conduct another experiment that adds the translated data to the training set and finetunes the baseline XLM-R model on Amazon Review dataset. The results showed very slight improvement (0.09% on average), showing that XLM-R cannot directly benefit from external training data.

Ablation studies are performed on Amazon Review dataset to analyze the effectiveness of different graph structures. From the results provided in table 4, variant 1 containing a homogeneous graph with only word-doc edges (same as TextGCN) performs the worst, while adding more information leads to better performance in general. Comparing variants 4-7, similarity demonstrates to be the most important among all added information. Similarity edges help the model to converge faster and learn better as well, since they provide a "shortcut" between documents that is highly likely to be

	XLM-R	1	2	3	4	5	6	7	full model
word-doc		✓		✓	✓	✓	✓	✓	✓
POS tags						✓	✓	✓	✓
translation edges			✓		✓		✓	✓	✓
similarity edges			✓	✓	✓	✓		✓	✓
unlabeled			✓	✓	✓	✓	✓		✓
EN → DE	90.01	87.60	90.60	90.75	90.77	90.67	89.90	**91.26**	90.97
EN → FR	89.52	90.62	89.95	90.65	90.70	90.37	89.82	90.85	**90.92**
EN → JA	86.61	86.26	87.19	87.31	87.35	87.44	87.18	86.57	**87.54**

Table 4: Ablation study results. The left-most column shows the results of finetuned XLM-R, and others each indicates a variant in graph construction. We conduct experiments on the Amazon Review dataset and report the average accuracy across three domains.

in the same category. Variant 7 shows that unlabeled corpus play an important role in EN→JA setting, but less effective when transferring between similar languages, since unlabeled data inevitably contain some noise and do not provide much help for linguistically-closer languages. Variant 4 also shows that POS tags are more helpful for distant language pairs like EN→JA, and our added experiment on EN→TH shows greater impact of POS tags (89.71→88.06 when removing POS tags). Additionally, we test a variant without any external tool that requires training resources in the target language. Variant 3 does not rely on POS tagger or translation service, and still outperforms the XLM-R baseline with a large margin. This demonstrates that our method can be adopted for real low-resource languages without good tagger or translator.

5 Conclusion

In this study, we propose a novel graph-based method termed CLHG to capture various kinds of information within and across languages for cross-lingual text classification. Extensive experiments illustrate that our framework effectively extracts and integrates heterogeneous information among multilingual corpus, and these heterogeneous relations can enhance existing models and are instrumental in cross-lingual tasks. There may exist some better semantic or syntactic features and combinations of features, which we leave as a future work to explore. We also wish to extend our GNN-based framework to different NLP tasks requiring knowledge transfer and adaptation in the future.

Acknowledgement

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel Andrade, Kunihiko Sadamasa, Akihiro Tamura, and Masaaki Tsuchida. 2015. Cross-lingual text classification using topic-dependent word probabilities. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1466–1471.

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

Zhuang Chen and Tieyun Qian. 2019. Transfer capsule network for aspect level sentiment classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 547–556.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pretraining text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2017. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1107–1116, Valencia, Spain. 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.

Kaize Ding, Jianling Wang, Jundong Li, Dingcheng Li, and Huan Liu. 2020. Be more with less: Hypergraph attention networks for inductive text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4927–4936, Online. Association for Computational Linguistics.

Xin Luna Dong and Gerard de Melo. 2019. A robust self-learning framework for cross-lingual text 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 6307–6311.

Kevin Duh, Akinori Fujino, and Masaaki Nagata. 2011. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 429–433.

Hongliang Fei and Ping Li. 2020. Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5759–5771.

Qipeng Guo, Xipeng Qiu, Pengfei Liu, Xiangyang Xue, and Zheng Zhang. 2020. Multi-scale self-attention for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7847–7854.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Linmei Hu, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short text 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 4821–4830, Hong Kong, China. Association for Computational Linguistics.

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

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Zheng Li, Mukul Kumar, William Headden, Bing Yin, Ying Wei, Yu Zhang, and Qiang Yang. 2020. Learn to cross-lingual transfer with meta graph learning across heterogeneous languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2290–2301, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv*, abs/2004.01401.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1118–1127.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data

augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3853–3860.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805.

Lei Shi, Rada Mihalcea, and Mingjun Tian. 2010. Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1057–1067.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4067–4076, Online. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 235–243.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Ruochen Xu and Yiming Yang. 2017. Cross-lingual distillation for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Vancouver, Canada. Association for Computational Linguistics.

Ruochen Xu, Yiming Yang, Hanxiao Liu, and Andrew Hsi. 2016. Cross-lingual text classification via model translation with limited dictionaries. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 95–104, New York, NY, USA. Association for Computing Machinery.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7370–7377.

Mozhi Zhang, Yoshinari Fujinuma, and Jordan Boyd-Graber. 2020a. Exploiting cross-lingual subword similarities in low-resource document classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9547–9554.

Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. 2020b. Every document owns its structure: Inductive text classification via graph neural networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 334–339, Online. Association for Computational Linguistics.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1412.

Yftah Ziser and Roi Reichart. 2018. Deep pivot-based modeling for cross-language cross-domain transfer with minimal guidance. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 238–249.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.

A Datasets

Here we introduce the three datasets we use in our experiments. A summary of the statistics for three datasets are provided in table 5.

Amazon Review [5]   This is a multilingual sentiment classification dataset covering 4 languages (English, German, French, Japanese) on three domains (*Books, DVD*, and *Music*). The original dataset contains ratings of 5-point scale. Following previous works (Xu and Yang, 2017; Fei and Li, 2020), we convert the ratings to binary labels with threshold at 3 points. Since English is not used for testing, we follow the previous works and re-construct the training and validation set by combining the English training and test set. The new training set contains 3,200 randomly sampled documents. We use the training set of target languages for validation. This dataset also provides large amount of unlabeled data for each language, which is used in our framework.

XGLUE News Classification [6]   This is a subtask of a recently released cross-lingual benchmark named XGLUE. This subtask aims at classifying the category of a news article, which covers 10 categories in 5 languages, including English, Spanish, French, German and Russian. This dataset does not contain unlabeled data for target languages, so we do not add unlabeled documents in our graph either.

Multilingual SLU [7]   This dataset contains short task-oriented utterances in English, Spanish and Thai across weather, alarm and reminder domains. We evaluate our framework on the intent classification subtask, which has 12 intent types in total. For each target language, we use the original training set as unlabeled data added in the graph.

B Hyperparameter Search

We perform a grid search to pick the best combination of hyperparameters. The hidden size and output size are chosen among $\{384, 512, 768\}$, and the learning rate within $\{1 \times 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}\}$. For the XLM-R baseline, we also tune the learning rate within $\{2 \times 10^{-5}, 4 \times 10^{-5}, 5 \times 10^{-5}\}$ and number of epochs from 2 to 5. Among all the combination of hyperparameters, we pick the values with the best performance on the German training set from *Books* domain of Amazon Review dataset, and use the same set of hyperparameters for all our experiments. The maximum length of XLM-R model is chosen based on statistics of the datasets.

---

[5] https://webis.de/data/webis-cls-10.html
[6] https://microsoft.github.io/XGLUE/
[7] https://fb.me/multilingual_task_oriented_data

dataset	#category	language	#train	#valid	#test	#unlabeled	avg. length
Amazon Review (Books)	2	English	2,000	/	2,000	50,000	168.31
		German	2,000	/	2,000	165,457	151.27
		French	2,000	/	2,000	32,868	123.84
		Japanese	2,000	/	2,000	169,756	155.05
Amazon Review (Dvd)	2	English	2,000	/	2,000	30,000	167.31
		German	2,000	/	2,000	91,506	158.58
		French	2,000	/	2,000	9,356	138.89
		Japanese	2,000	/	2,000	68,324	150.87
Amazon Review (Music)	2	English	2,000	/	2,000	25,220	146.18
		German	2,000	/	2,000	60,382	143.50
		French	2,000	/	2,000	15,940	142.21
		Japanese	2,000	/	2,000	55,887	131.62
XGLUE NC	10	English	100,000	10,000	10,000	/	553.65
		German	/	10,000	10,000	/	484.69
		French	/	10,000	10,000	/	567.86
		Spanish	/	10,000	10,000	/	533.09
		Russian	/	10,000	10,000	/	426.80
Multilingual SLU	12	English	30,521	4,181	8,621	/	8.05
		Spanish	3,617	1,983	3,043	/	8.74
		Thai	2,156	1,235	1,692	/	5.12

Table 5: Summary statistics of the datasets. The average length shows the average number of words in each document.