Towards a Fine-Grained Multi-Domain Neural Machine Translation Using Inter-Domain Relationships

Anonymous ACL submission

Abstract

While research on the domain adaptation task in neural machine translation has become popular recently, there exists no agreement on what constituents a domain, and most previous studies only focus on coarse-grained domain adaptation and their methods cannot be generalized if the domain size is large. In this work, we argue the necessity to study a fine-grained domain adaptation problem. We build a new multilingual dataset from web sources that focus on fine-grained domains and inter-domain attributes and relationships. We also propose a simple but effective adaptation method to incorporate domain knowledge leveraging models in information networks.

1 Introduction

The success of most machine learning algorithms heavily relies on the assumption that feature spaces and underlying distributions of training data and testing data are similar. In reality, such assumption is often violated. This motivates the study of domain adaptation in many areas including computer vision (Wang and Deng, 2018) (Csurka, 2017), natural language processing (Ramponi and Plank, 2020) and recommendation systems (Pan, 2016).

In the neural machine translation (NMT) task, domain adaptation is also in high demand since previous researches have shown that a general translation system trained on open-domain corpus often performs poorly in specific domains (Koehn and Knowles, 2017). Since in-domain parallel corpora are often insufficient, leveraging both in-domain and out-of-domain resources becomes important. Previous researches exhibit several sub-tasks: a semi-supervised task that uses a small-sized parallel corpus of the target domain (Jia and Zhang, 2020), and a supervised meta-learning task that requires small-sized in-domain support data-set to fast adapt to the target domain (Li et al., 2020; Sharaf et al., 2020). We argue that there are two problems regarding current domain adaptation studies.

What is a domain? There exists no common agreement on what constitutes a domain in NMT. Many researches attribute domain difference to different vocabulary (Blitzer et al., 2006) or system of a document (McClosky, 2010). In practice, people use different predetermined dataset to represent different domains (Plank, 2016; Ramponi and Plank, 2020). Such practices often neglect the heterogeneity inside the corpus, which Plank (2016) put forward a theoretical notion called the *variety space*. A corpus can be seen as a set of sub-domains drawn from the underlying variety space. It is necessary to consider the smaller granularity of domains.

How to utilize the prior knowledge of inter-domain relationships and attributes? Domain adaptation in NMT usually tries to utilize linguistic similarities shared between source and target domains (Britz et al., 2017; Gu et al., 2019). However, there are other attributes outside the corpus to utilize. Although parallel or monolingual training corpus in NMT for a newly-published document is hard to get, we can easily find attributes that are shared with other domains (e.g., authors, categories, tags) and their relation with previous domains (e.g.recommendations). They can be regarded as prior knowledge for domain adaptation methods. In computer vision, Gebru et al. (2017) utilizes domain-level attributes and adopts a multi-task method for fine-grained domain adaptation. However, in NMT, we did not find similar researches.

The goal of our work is to solve the above three problems. The contribution comes in two folds: We build a multi-lingual dataset and provide inter-domain attributes as well as relationships. We believe this dataset is the first public dataset that focuses on fine-grained domains and is a good resource to explore the open-set domain adaptation and domain-level knowledge integration.

Field Name	Description
id	unique index of a game
lang	language of this sample
name	name of the game
developer	developers of the game
publisher	publishers of the game
category*	game category
type	game type
tag*	most voted tags by users
text	game introduction
recommended	ids of similar games

Table 1: Field description of STEAM. Fields marked with * are multilingual. The text field across two languages for the same game is aligned in sentence-level

We also propose a new approach that utilizes prior knowledge of inter-domain relationships by training an information network (Zhu et al., 2020b), whose output is treated as additional feature embeddings. To the best of our knowledge, this is the first work that tries to incorporate prior inter-domain relationships in an open-set domain adaptation task. Please find more information about the related work and baseline implementation in the appendix.

2 STEAM: A Multilingual Parallel Dataset for Domain Adaptation

Few datasets are designed specifically for domain adaptation and most datasets do not consider heterogeneity between samples. This corpus is a reaction to the growing importance of NMT in specific domains. While FDMT (Zhu et al., 2020a) make fine-grained domain corpus, it does not provide training data and only has English-Chinese pair. Most researches in multi-domain adaptation only consider a small subset of listed domains and cannot be generalized to the open-set domain adaptation settings. Domains are very distant from each other. No dataset describes inter-domain relationships and attributes.

This dataset is created based on Steam[1], a video game platform that contains thousands of game introductions in many languages. We produced parallel sentences based on HTML positions and LASER-based alignment (Artetxe and Schwenk, 2019). We also find relations between games based on information of game producers and user recommendations.

[1]https://store.steampowered.com/

2.1 Data Collection & Analysis

We use the python framework Scrapy [2] to crawl text information from game introductions. The steps are in the appendix. 8 languages are chosen which are Chinese(ch), German(ge), English(en), Spanish(es), Russian(ru), Japanese(ja), Korean(ko) and French(fr). The field description is shown in 1. The data processing part is divided into two steps, monolingual and bilingual sentence matching. The details can be found in the appendix. The goal is to make the mapping of the same statement in different languages end up in the same neighborhood.

There are 10k games with 90k aligned parallel sentences after the cleaning. Table 3 in the appendix illustrates the detailed information. Codes and examples of bilingual parallel sentences can be found on our Google repository.[3]

3 Problem Formulation and Proposed methods

3.1 Task Definition

D_i is the i^{th} domain, which is a tuple of four elements (S_i, T_i, R_i, A_i). S_i and T_i refer to paralleled source and target data in i^{th} domain. R_i refers to a set of other domains related to the i^{th} domain.A_i refers to domain-level attributes shared to all samples in this domain. The goal is to build a model that can generate best $\hat{T}_i$ based on (S_i, R_i, A_i). Our domain graph is defined as $G = (D, E)$, where E is the set of relationships between domains. We want to generate a feature embedding for each domain by modelling all domains in an undirected graph to avoid the scalability issue occurred.

3.1.1 Inter-Domain Relation Extraction

In this dataset, each domain is regarded as a node. The purpose of the inter-domain is to extract the relationship among nodes. There are several methods to learn nodes representations for inter-domain knowledge. Kipf and Welling (2017) generalizes neural models to work with structured graphs for semi-supervised learning. Tang et al. (2015) proposes LINE to generate node embeddings in a network by preserving the first second order proximities among nodes. We the use graph representation learning method named deep GRAph Contrastive rEpresentation learning (GRACE)(Zhu et al., 2020b). Two graph views are generated at

[2]https://scrapy.org/

[3]https://drive.google.com/drive/folders/1wlj5sNETFL5hatC5sZfrHyAYf3ACK2nd?usp=sharing

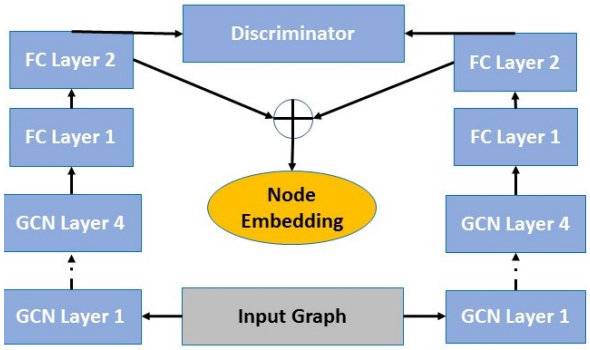

Figure 1: Illustration of the modified GRACE model

each iteration so that each node in the original graph has two embeddings. A discriminator is used to distinguish the two embeddings. We choose the method because it is stated that the framework is suitable for large-scale graphs and which fits our dataset.Also, the connections in our dataset are generated by the recommendation rules of steam and they are not clearly stated. It is rational to assume there are complicated relationships among the connected nodes. We extend the original framework by using a four-layer GCN network instead of two in the generation of node embeddings. Moreover, we increase the number of nodes in the MLP layers to 512 which fits the input of our NMT model. The final embedding of each node is obtained by adding the two hidden layers together and dividing them by two. The modified GRACE framework is shown in 1. The code can be found in Google Drive link[4].

3.1.2 Handling Isolated Nodes

In the domain adaptation settings, there can be some domains not related to other domains. We call these domains isolated nodes. The proportion of isolated domains in our STEAM dataset is shown in the appendix. We cannot simply feed those isolated nodes into GRACE that takes as input a list of edges. We augment every node with a self-loop to have every node encoded and take an average of all the isolated node embedding vectors to get a final representation for all the isolated nodes.

3.1.3 Domain-Aware Encoder

We use the Transformer (Vaswani et al., 2017) as our backbone NMT model. We are motivated by Dou et al. (2019) to assign a domain-aware feature embedding on the model encoder side as an additional input representation to disentangle encoder representation and learn domain knowledge. Although (Dou et al., 2019) trains different domains embeddings for each encoder layer, we found it not scalable in our setting. We combine three different ways to implement the domain-aware encoder, which is shown in figure2. Given the feature representation of in domain i as d_i, the j^{th} token of the sentence as x_j , the first method is to add the domain features only after the input embeddings. The formula is:

$$h_j^{(0)} = 0.5 * (W_c x_j + p_j + W_d d_i) \quad (1)$$

where $h_j^{(0)}$ is the input embeddings of x_j, W_c is the contextual embeddings of tokens, and p_j is the the positional embeddings. W_d is a trainable linear projection that transforms the graph-based domain representations into domain embeddings for the NMT model. Note that domain is labeled at the sentence level, so tokens within a sentence share the same domain embeddings.

The second method is to add the projected domain features right after the Multi-Head Attention module of each layer. The formula is:

$$\tilde{H}^{(l)} = 0.5 * (\text{Multihead}(Q, K, V) + W_d^{(l)} d_i) \quad (2)$$

where $\tilde{H}^{(l)}$ is the intermediate hidden representation before feed-forward network and residual connection in each layer and $\text{Multihead}(Q, K, V)$ is the projected output of the concatenated attention heads defined in Transformer's original paper.

The third method is to add domain embeddings at the end of each transformer encoder layer, which is similar to the method used in (Dou et al., 2019). The formula is

$$H^{(l)} = 0.5 * (\text{layer}_i(H^{(l-1)}) + W_d^{(l)} d_i) \quad (3)$$

The main difference between our proposed methods and methods introduced in Dou et al. (2019) and Michel and Neubig (2018) (noted as previous methods) is that they use one-hot encoded domain labels as the input of domain embedding/adapter layers, while we use graph-based domain feature representation as the input. We claim two advantages of using graph-based domain feature representation.

Reduced number of parameters The method introduced in Dou et al. (2019) requires at least $r|D|$ more parameters and the method in Michel and Neubig (2018) requires at least $r(|D| + |V|)$

[4]https://drive.google.com/drive/folders/1LCYP3WH489G–KfgnrgsKy8-HluXR26M?usp=sharing

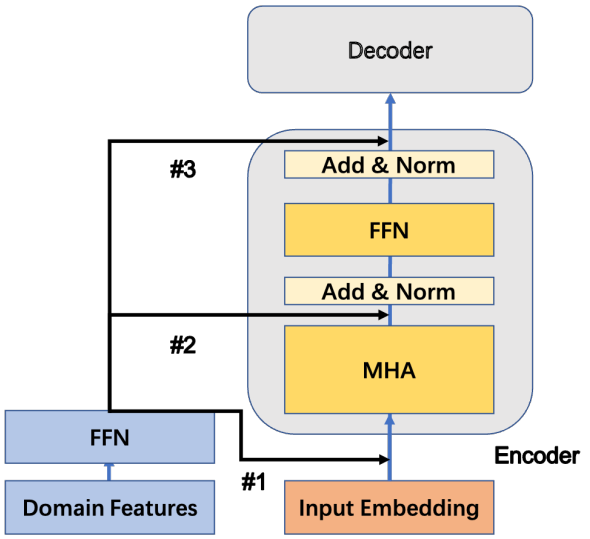

Figure 2: Architecture of the Domain-Aware Encoder. #X denotes the X^{th} method mentioned in 3.1.3

more parameters, where r is the hidden dimension of the transformer encoder, $|D|$ and $|V|$ are the size of domains and vocabularies which are linearly correlated with the size of domains. Our method only requires rk more parameters where k is the dimension of domain representation and is more practical when the size of domains is very large. In our dataset, the domain size for the en-zh pair is 9392, and we only use 5% more parameters compared with previous methods.

Inference on unseen domains Previous methods require the model to use either monolingual or bilingual corpus of all domains. Our methods can be used in an open-set domain situation where the corpus of the testing domains is not available in the training step based on domain similarities extracted from the graph-based domain feature representation.

4 Experiments

4.1 Datasets and Mixed fine-tuning

We choose ch-en, fr-en, es-en and ru-en as tested language pairs because they are four distinctively different language families. Domains in the test set are excluded in the training set to guarantee an open-set domain adaptation setting. We use mixed fine-tuning (Chu et al., 2017) as our training strategy. We collect UN Corpus[5] as the out-of-domain data. We first pretrain a vanilla transformer model based on this out-of-domain data for 5 epochs. Then we sample 2 million sentences pairs from UN Corpus and over-sample another 2 million sentence pairs from the STEAM corpus for mixed fine-tuning. For our baseline model, pre-trained NMT models are directly loaded and fine-tuned in the mixed data, while for our proposed models, parameters related to the domain features are randomly initialized in the pretrain step. We also generate noise embeddings for comparison. The mean and variance of the noise embeddings are the same graph embeddings which leads to a similar distribution. The implementation details are in the appendix and our code can be found in our Google link[6].

[5]https://conferences.unite.un.org/UNCorpus/

Method	zh-en	es-en	fr-en	ru-en
Baseline	17.22	28.31	23.18	14.28
Noise Embedding	15.80	27.41	12.86	7.58
DA Encoder	**17.85**	**29.52**	**26.08**	**15.52**

Table 2: BLEU in the test set for the baseline and proposed models. Bold text highligts the best results

4.2 Main Results

Table 2 shows that our proposed models generally improve the overall translation performance in the test set. Regarding that our proposed method only requires very few extra parameters, this improvement is satisfying. Also, with noise embeddings, the results are worse than the baseline.

In Table 4 in the appendix, we can see although the BLEU improvement is relatively small, our proposed methods did improve the quality of translated sentence by correctly translating some key words.

5 Conclusion

This paper presents a new research topic on fine-grained multi-domain adaptation in NMT. We contribute a new dataset that focuses on the fine-grained domains and inter-domain relationships and proposes a novel method to utilize inter-domain relationships.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

[6]https://drive.google.com/drive/folders/1gRSv9B2ZMqzTKdDD9ge9tmDJkEdLTYWz?usp=sharing

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128.

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126, Copenhagen, Denmark. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Gabriela Csurka. 2017. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*.

Zi-Yi Dou, Junjie Hu, Antonios Anastasopoulos, and Graham Neubig. 2019. Unsupervised domain adaptation for neural machine translation with domain-aware feature embeddings. 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 1417–1422.

T. Gebru, J. Hoffman, and L. Fei-Fei. 2017. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1358–1367.

Shuhao Gu, Yang Feng, and Qun Liu. 2019. Improving domain adaptation translation with domain invariant and specific information. 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 3081–3091, Minneapolis, Minnesota. Association for Computational Linguistics.

Chen Jia and Yue Zhang. 2020. Multi-cell compositional LSTM for NER domain adaptation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5906–5917, Online. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Rumeng Li, Xun Wang, and Hong Yu. 2020. Metamt, a meta learning method leveraging multiple domain data for low resource machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8245–8252.

David McClosky. 2010. *Any domain parsing: automatic domain adaptation for natural language parsing*. Ph.D. thesis, Brown University.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318, Melbourne, Australia. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Weike Pan. 2016. A survey of transfer learning for collaborative recommendation with auxiliary data. *Neurocomputing*, 177:447–453.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Amr Sharaf, Hany Hassan, and Hal Daumé III. 2020. Meta-learning for few-shot NMT adaptation. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 43–53, Online. Association for Computational Linguistics.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.

Wenhao Zhu, Shujian Huang, Tong Pu, Xu Zhang, Jian Yu, Wei Chen, Yanfeng Wang, and Jiajun Chen. 2020a. Fdmt: A benchmark dataset for fine-grained domain adaptation in machine translation.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2020b. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*.

A Example Appendix

A.1 Data Crawling and Cleaning

Setting request cookies such as birth time, age checking, header, initial searching URL, and language code.

Parsing information in the game search page of each language including the URL, game name, game ID, and information language.

Getting detailed information from the game page including review text, game tags, game categories, developer, publisher, and recommendations. The recommendation is to recommend similar games based on the current game.

Monolingual sentence cleaning focuses on filtering abnormal sentences, words, and sentences that are not in the labeled languages, using a language identification tool called langid[7]. We use two steps to build the bilingual parallel sentences and

First we find the passage that can be segmented into small paragraphs based on different positions in the website. Then we use Stanza[8] to segment sentences in the paragraph pair. For each paragraph pair, if the numbers of segments for two paragraphs are the same, each pair of segments is treated as a pair of parallel sentences. Otherwise, we implement LASER[9] to compute the sentence embedding and use cosine similarity to pick out similar sentence pairs. LASER is a toolkit to represent sentences by vectors which are generated with respect to both the input language and the NLP task.

A.2 Implementation details

We use fairseq (Ott et al., 2019) as our code base, and use $1e-4$ as the learning rate with the inverse learning rate decay. All models share the same vocabulary list, which is generated using sentencepiece[10] on the 4m mixed data mentioned above.

[7]https://github.com/saffsd/langid.py
[8]https://stanfordnlp.github.io/stanza/
[9]https://github.com/facebookresearch/LASER
[10]https://github.com/google/sentencepiece

	Sent. Num	Game Num	Avg. Length	Cat/Tag Num	Recommended
zh-en	82819	9392	29(CH)/15(EN)	29/418	12
ru-en	91719	10069	13(RU)/14(EN)	29/419	13
fr-en	93357	10550	17(FR)/15(EN)	30/425	12
es-en	92265	9264	16(ES)/15(EN)	29/419	12
ja-en	58445	10678	40(JA)/15(EN)	28/420	14
ko-en	69539	7982	35(KO)/14(EN)	28/423	12
de-en	93203	9504	17(DE)/15(EN)	29/420	13

Table 3: Statistics of our bilingual parallel dataset.[11]

Reference	**You are afraid. You are forced to hide.** Those who served you till recently have suddenly **turned against you**. Your only option **is to hide, surviving with no hope,** not understanding what's going on around you. You are an outcast, an outlaw because you are human.
Baseline	You feel terrified and forced to crawl on. Now, they have branded your friends and family. Your only option is to survive with despair. You have no idea what you're hiding because you are human, you're the outcasts and outlaws.
Our Model(DA Encoder)	**You're scared and forced to hide.** The robots that served you were suddenly **betrayed**. Your only option is **hiding and survive with despair**. You don't know where they are or why they do this. Because you are a human being, so you are the bearer and the outcasts.

Table 4: Comparison between the system output and the baseline.

Language	Total Games	Total Isolated Games
en	73306	26530
zh	15858	6459
ru	16654	6925
fr	18757	9583
es	17525	8342

Table 5: Statistics of isolated games for each language.