INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Domain-Aware Self-Attention for Multi-Domain Neural Machine Translation

Shiqi Zhang[1], *Yan Liu*[2], *Deyi Xiong*[2*], *Pei Zhang*[1], *Boxing Chen*[1]

[1]Alibaba Group Inc, China
[2]College of Intelligence and Computing, Tianjin University, China
{zsq169774, xiaoyi.zp, boxing.cbx}@alibaba-inc.com, {yan_liu, dyxiong}@tju.edu.cn

Abstract

In this paper, we investigate multi-domain neural machine translation (NMT) that translates sentences of different domains in a single model. To this end, we propose a domain-aware self-attention mechanism that jointly learns domain representations with the single NMT model. The learned domain representations are integrated into both the encoder and decoder. We further propose two different domain representation learning approaches: 1) word-level unsupervised learning via a domain attention network and 2) guided learning with an auxiliary loss. The two learning approaches allow our multi-domain NMT to work in different settings as to whether the domain information is available or not. Experiments on both Chinese-English and English-French demonstrate that our multi-domain model outperforms a strong baseline built on the Transformer and other previous multi-domain NMT approaches. Further analyses show that our model is able to learn domain clusters even without prior knowledge about the domain structure.

Index Terms: Transformer, machine translation, domain adaptation, unsupervised learning

1. Introduction

Neural machine translation (NMT) has achieved remarkable progress in translation quality recently [1, 2, 3]. This success benefits from both advanced neural architectures and the availability of a huge amount of in-domain training data. However, the availability of such sufficient in-domain data is always an issue for many languages, and words may have different senses across domains. Therefore, it is necessary for NMT to take domain information into account.

In this paper, we consider a domain-related open problem for NMT: how can a single domain-aware NMT model be built on training data from multiple domains?

Conventional approaches to the multi-domain setting are to train a generic model first and then fine-tune the model on a specific domain to maximize its performance on the domain [4] or to train multiple models on different domains for model combination. Both require to build multiple domain-specific models to be invoked by testing sentences, which may increase the maintenance cost for the deployment of MT on industrial scenarios [5].

We attempt to approach multi-domain NMT by jointly learning the embeddings of domains with the NMT model. In this way, we force the NMT model to encode and decode with both semantic and domain information. Therefore, a single NMT model can deal with multiple domains in this way.

In order to jointly learn domain representations with the NMT model, we propose a domain-aware self-attention (DSA) mechanism for multi-domain NMT. In DSA, each word has a semantic word embedding and a domain embedding.[1]

In order to learn domain embeddings, we further propose two approaches: 1) unsupervised learning via a domain attention network, and 2) guided learning with an auxiliary loss. With these two approaches, our DSA model is capable of dealing with a variety of multi-domain configurations. The unsupervised domain learning method does not require any prior knowledge about the domain structure of both the training and test data. It can even learn such a domain structure at the word level simultaneously with the training of the NMT model.

2. Preliminaries: Self-Attention of Transformer

Given an input sequence $x = (x_1, ..., x_n)$ of n elements where $x_i \in R^{d_x}$ with a dimension d_x, the output of a single head of the self-attention layer in Transformer is a weighted sum of a linearly transformed input (i.e., value), which can be computed as:

$$o_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V) \tag{1}$$

The attention weight α_{ij} measures the degree of compatibility between a query x_i and key x_j, which can be estimated as the output of a softmax function:

$$\alpha_{ij} = \frac{exp((e_{ij}))}{\sum_{k=1}^{n} exp((e_{ik}))} \tag{2}$$

$$e_{ij} = \frac{x_i W^Q (x_j W^K)^T}{\sqrt{d_x}} \tag{3}$$

where $W^Q, W^K, W^V \in R^{d_x * d_x}$ are projection matrices for the query, key and value, which can be learned during the training phase.

3. Multi-Domain NMT

In the following we first describe the overall architecture of Transformer-based multi-domain NMT and then elaborate the DSA model .

3.1. Overview

The overall architecture is shown in Figure 1. The goal is to allow the Transformer to encode the domain information of an input sequence and decode domain-specific target translations with such encoded domain information. To this end, we make two significant changes. The first change is the domain-aware self-attention where domain representations are added to the

*Corresponding author

[1]In this paper, the term "domain embeddin" is used exchangeably with "domain representation".

http://dx.doi.org/10.21437/Interspeech.2021-1477

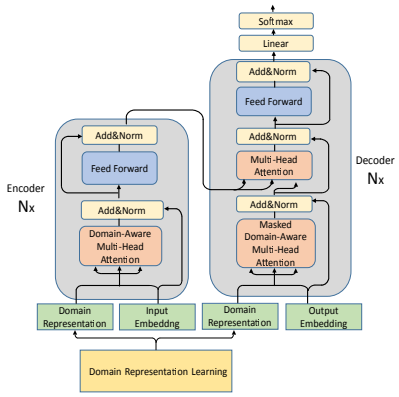


Figure 1: *The architecture of Transformer-based multi-domain NMT.*

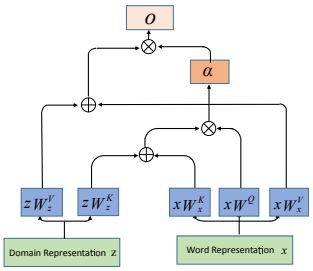


Figure 2: *The illustration of the domain-aware self-attention.*

key and value vectors of the original self-attention. The output of the domain-aware self-attention function is a weighted sum of domain-informed values. The proposed domain-aware self-attention can be used on the self-attention layer of the encoder, decoder or both. It can also be used in the encoder-decoder attention layer. In our preliminary study, we find that using the domain-aware self-attention in the self-attention layer of both the encoder and decoder simultaneously is better than other cases. The second change is to add a domain representation learning module to learn domain embeddings, which will be introduced in Section 4.

3.2. Domain-Aware Self-Attention

We assume that each element in a sequence has a domain representation. We let the sequence $z = (z_1, ..., z_n)$ of the same length as x be the domain representations corresponding to x. z_i ($\in R^{d_z}$) of dimension d_z is the domain representation for element x_i. The dimensions of x and z can be the same (i.e., $d_z = d_x$) if we do not use additional linear transformations. By forcing all elements to have identical domain representations, i.e., $z_1 = z_2 = ... = z_n$, the word-level domain model will be collapsed into a sentence- or text-level model.

The proposed domain-aware self-attention is shown in Figure 2. By adjusting Eq. (1) to include the domain representations for all input elements x_j, the output element o_i of DSA can be computed as follows.

$$o_i = \sum_{j=1}^{n} \alpha_{ij}(x_j {W_x}^V + z_j {W_z}^V) \quad (4)$$

Intuitively, the output will contain both the semantic information from x and domain information from z. We hope that the domain information can help disambiguate source words if the domain-aware self-attention is used in the encoder and choose right target words if it is in the decoder.

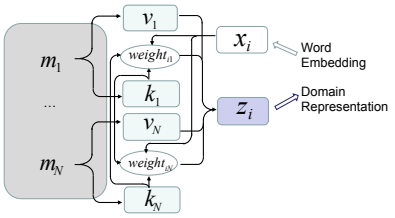


Figure 3: *The architecture of the domain attention network used in the unsupervised learning method. v, k denote value and key vectors respectively.*

Similarly, the function computing the compatibility between the query and key vectors is also changed to take domain representations into account as follows.

$$e_{ij} = \frac{x_i W^Q (x_j W_x^K + z_j {W_z}^K)^T}{\sqrt{d_x}} \quad (5)$$

W^Q, ${W_x}^K$, ${W_x}^V \in R^{d_x * d_x}$, ${W_z}^k$, ${W_z}^V \in R^{d_z * d_x}$ are transformation matrices to be learned.

4. Domain Representation Learning

The key to the domain-aware self-attention is to learn domain representations z. In this section, we present two different approaches to learning such representations.

4.1. Word-Level Unsupervised Learning

In many cases, we do not know which domain a sentence or text comes from. For this, we propose an unsupervised domain representation learning approach. We initialize a set of domain vectors $m = \{m_1, ..., m_N\}$ where $m_i \in R^{d_m}$ of a dimension d_m, N is a hyper-parameter to be predefined for the number of domain clusters. We treat the domain representation z_i of element x_i in the domain-aware self-attention as a weighted mixture of N domain vectors in m, instead of assigning z_i to be one of domain vectors in m.

In order to learn the weights for the domain mixture model, we build a domain attention network (shown in Figure 3) to allow each element x_i to attend to all domains in m. The mixture weights are calculated as attention weights via the compatibility function where the queries are elements from x and the keys are domain vectors. The scaled dot-product function [1] is used for the compatibility as follows.

$$\alpha_{ij} = \frac{exp((e_{ij}))}{\sum_{k=1}^{N} exp((e_{ik}))} \quad (6)$$

$$e_{ij} = \frac{x_i {W_u}^Q (m_j {W_m}^K)^T}{\sqrt{d_m}} \quad (7)$$

With these weights, the domain representation z_i can be computed as follows.

$$z_i = \sum_{j=1}^{N} \alpha_{ij}(m_j {W_m}^V) \quad (8)$$

${W_u}^Q \in R^{d_x * d_m}$, ${W_m}^K \in R^{d_m * d_m}$, ${W_m}^V \in R^{d_m * d_z}$ are parameter matrices.

Table 1: *Statistics on the data used for the ZH-EN task.*

ZH-EN	train	dev	test
news-commentary	250K	800	1,200
Laws	220K	800	456
Thesis	300K	800	625
IWSLT	230K	879	1,205
Total	1M	3,279	3,486

Table 2: *Statistics on the data used for the EN-FR task.*

EN-FR	train	dev	test
ECB	194K	500	1,000
KDE4	209K	500	1,000
IWSLT	233K	500	1,000

4.2. Guided Learning with an Auxiliary Loss

In the unsupervised approach, we jointly learn the domain representations with the NMT model. However, we can also use off-the-shelf methods to learn domain representation for each word in the training data independently of the NMT model. For example, we can learn topic models on the training data and treat the distribution of the topic assigned to a word as the domain representation of this word.[2] The basic idea is to adjust the domain representations computed from the domain attention network to be close to the representations learned by the external model.

We therefore introduce an auxiliary loss Δ to measure the disagreement between the domain representation z_i learned by the domain attention network and $\widehat{z_i}$ learned by an external domain model. The final objective for training is to minimize the following loss.

$$-\sum_i logp(y_i|x_i;\theta) + \lambda * \Delta(z_i, \widehat{z_i};\theta) \quad (9)$$

where $\lambda > 0$ is a hyper-parameter that balances the impact of translation likelihood and domain disagreement on the training.

As the auxiliary loss is used only in the training phase, external domain guidance is not required during the testing phase. In this study, we use the distributional vector space model used by [6] as the external model and define Δ as the cross-entropy of two domain representations.

5. Experiments

We conducted experiments on multi-domain Chinese-English and English-French translation.

5.1. Data

For Chinese-to-English (ZH-EN) multi-domain translation task, we used a collection of publicly available corpora from different domains: news-commentary data from WMT2018, Laws and Thesis from UM-Corpus, and IWSLT data from IWSLT2017. For the English-to-French (EN-FR) multi-domain translation task, our data came from OPUS, IWSLT2017 and WMT2015. The training data were composed of three different domains: European Central Bank (ECB), KDE4, IWSLT. Statistics on the data of the two tasks are shown in Table 1 and Table 2.

5.2. Settings

We used the tensor2tensor library to implement our multi-domain NMT systems, both the baseline and the systems enhanced with the proposed domain-aware self-attention and domain representation learning approaches.

We used the Adam optimizer [7]. We used 6 encoder and decoder layers, 8 attention heads, 1024 feed forward inner-layer dimensions, and set dropout = 0.1. We set the dimension of domain representations the same as that of sentence embedding ($d_x = d_m = 512$). The hyper-parameter λ was set to 1 and the number of domain types N in both the word-level unsupervised learning and guided learning was set to 4 in our experiments. We used the same warmup and decay strategy for learning rate as [1] with 8,000 warmup steps. For evaluation, we used beam search with a beam size of 6 and BLEU-4 as the evaluation metric.

[2]We leave the use of guidance from the topic model to our future work.

5.3. Results

The results of the ZH-EN and EN-FR task are shown in Table 3. We compared our models against the following two systems.

- The baseline system is the Transformer that was trained on the combination of all training data from different domains.
- The other system is DomainTag, a Transformer system trained on the combination of all data where each sentence is annotated with a domain ID. Following [8], we treat domain as a tag and append the corresponding domain ID to each training sentence.

In the Chinese-English task, the word-level unsupervised learning and guided learning achieve average improvements of 0.59 and 1.26 BLEU points over the baseline. These results suggest that the proposed domain-aware self-attention is able to incorporate domain information into NMT and indeed enables multi-domain translation with a single NMT model. The improvements are also reasonable. The guided learning obtains higher performance than the unsupervised learning, indicating that external useful domain information can be explored to improve unsupervised domain representation learning. In the EN-FR task, results in Table 3 show a similar trend to what we observe on the Chinese-English translation task.

Unfortunately, the DomainTag method, appending domain tags to the training data [8], does not work for the Transformer. It is even much worse than the baseline. [8] has used the DomainTag method on RNN-based NMT systems while we used it on the self-attention-based Transformer. Appending additional symbols to sequences may be not a good fit for the self-attention architecture.

6. Analysis

We further conducted analyses to study the impact of the proposed domain-aware self-attention model by looking into translations and data. Particularly, we are interested in 1) what we can learn for domains and 2) what happens to the NMT model with learned domain representations.

6.1. Analysis and Visualization of Domain Distributions over Words

We analyzed domain-word alignments learned by the guided learning model in the Chinese-English translation task. For each source word, the domain attention network calculates a weight for each connection of the word with one of the N learned

Table 3: *Results of different models on the ZH-EN task and EN-FR task.*

MODELS	ZH-EN					EN-FR		
	DEV	THESIS	NEWS	IWSLT	LAWS	IWSLT	KDE4	ECB
Baseline	29.72	16.16	25.73	22.90	77.07	43.21	45.93	47.82
DomainTag	28.18	16.09	23.55	20.73	72.46	43.65	43.88	47.86
Word-Level Unsupervised learning	29.84	16.64	25.38	**23.05**	79.16	43.64	**46.73**	**48.51**
Guided Learning	**30.16**	**18.33**	25.28	22.88	**80.41**	**43.87**	45.92	48.45

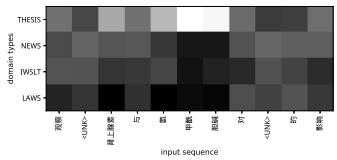
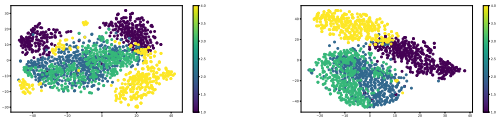


Figure 4: *Visualization of the attention between words and domains in an example source sentence. The lighter the color, the higher the attention.*



(a) (b)

Figure 5: *Visualization of source sentence representations learned by the baseline (a) and the guided learning model (b). Purple: Thesis. Dark blue: News. Green: IWSLT. Yellow: Laws.*

domains. This weight measures the degree that the word is compatible to the corresponding domain. We calculated the average of such attention weights over the six self-attention layers for each word and further averaged them for all occurrences of each word. We then normalized these average weights over all words for each domain. In this way, we can obtain the per-domain word distribution.

We visualize the attention between words and domains in an example sentence shown in Figure 4. We can see that the domain-specific words like "肾上腺素" (isoprenaline), "氨甲酰胆碱" (carbachol) have very high weights of attention to the Thesis domain while domain-insensitive function words such as "与" (and), "的" (of) distribute evenly over all domains.

6.2. Visualization of Sentence Representations

We further conducted an analysis on sentence representations learned by the encoder. We used t-SNE [9] to project representations of source sentences in the Chinese-English test set into 2D space. Figure 5 visualizes these representations learned by the baseline and the guided learning model. We can observe that it is difficult for the baseline to distinguish different domains for sentences. However, our model is able to gather sentences in domain clusters with clear boundaries. We believe that the domain information learned by our model can benefit the decoder in sense disambiguation and translation.

7. Related Work

Domain adaptation for NMT. In the context of NMT, various approaches including fine-tuning, data selection, data weighting have been explored for domain adaptation. [10] uses mixed fine-tuning method that fine-tunes on the mix of in-domain and out-of-domain corpora. [11] employs the data selection method for domain adaptation, which uses sentence embedding to measure the similarity of a sentence pair to the in-domain. Sentence weighting methods are also proposed with the objective function updated by sentence weights when computing the cost of each mini-batch during NMT training. [12] proposes two weighting methods: sentence weighting and domain weighting with a dynamic weight learning strategy. [13] uses classifier probabilities to weight sentences according to their domain similarity when updating the parameters of the neural translation model. [14] introduces a domain similarity metric to evaluate the relevance between a sentence and an available entire domain dataset, which is integrated into the training objective to weight sentences.

Multi-domain NMT. In neural machine translation, [15] has explored model stacking and multi-model ensemble for multi-domain translation. [16] proposes to control domain using word-level domain features in the word embedding layer. [5] dynamically sets the hyper-parameters of the learning algorithm and updates the generic model by exploiting the similarity between each test sentence and the training instances. [17] proposes three models to jointly learn to discriminate domains and translate, including discriminative mixing, adversarial discriminative mixing and target-side token mixing. [18] distinguishes and exploits word-level domain contexts for multi-domain NMT by constructing domain-specific and domain-shared annotations and adjusting the weights of target words in the training objective. [19] explores strategies to share the same encoder-decoder among all domains and to use a private encoder-decoder for each domain separately. [20] combines their methods with mixed fine-tuning using multilingual and multi-domain data. [21] proposes an approach that adapts the model by domain-aware feature embeddings learned from language modeling. [22] presents a multi-domain NMT with word-level layer-wise domain mixing. [23] explores the feature expansion technique of [24] in neural machine translation.

8. Conclusions

In this paper, we have presented a new multi-domain NMT framework. We jointly learn domain representations with the NMT model end-to-end. The learned domain representations are incorporated into the domain-aware self-attention. The unsupervised, and guided learning approaches are proposed to learn domain structures and their representations. Experiments validate the effectiveness of the proposed framework.

9. Acknowledgments

We would like to thank anonymous reviewers for their insightful comments. Yan Liu and Deyi Xiong were partially supported by the National Key Research and Development Program of China (Grant No. 2019QY1802).

10. References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[2] B. Marie, R. Wang, A. Fujita, M. Utiyama, and E. Sumita, "Nict's neural and statistical machine translation systems for the wmt18 news translation task," in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 2018, pp. 449–455.

[3] H. H. Awadalla, A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li *et al.*, "Achieving human parity on automatic chinese to english news translation," 2018.

[4] M.-T. Luong and C. D. Manning, "Stanford neural machine translation systems for spoken language domains," in *Proceedings of the International Workshop on Spoken Language Translation*, 2015, pp. 76–79.

[5] M. A. Farajian, M. Turchi, M. Negri, and M. Federico, "Multi-domain neural machine translation through unsupervised adaptation," in *Proceedings of the Second Conference on Machine Translation*, 2017, pp. 127–137.

[6] B. Chen, R. Kuhn, and G. Foster, "A comparison of mixture and vector space techniques for translation model adaptation," in *AMTA 2014, Proceedings of the 11th conference of the association for machine translation in the Americas*, vol. 1, 2014, pp. 124–138.

[7] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[8] S. Tars and M. Fishel, "Multi-domain neural machine translation," in *Proceedings of the 21st Annual Conference of the European Association for Machine Translation: 28-30 May 2018, Universitat d'Alacant, Alacant, Spain*. European Association for Machine Translation, 2018, pp. 259–268.

[9] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[10] C. Chu, R. Dabre, and S. Kurohashi, "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)*, vol. 2, 2017, pp. 385–391.

[11] R. Wang, A. Finch, M. Utiyama, and E. Sumita, "Sentence embedding for neural machine translation domain adaptation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2017, pp. 560–566.

[12] R. Wang, M. Utiyama, L. Liu, K. Chen, and E. Sumita, "Instance weighting for neural machine translation domain adaptation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1482–1488.

[13] B. Chen, C. Cherry, G. Foster, and S. Larkin, "Cost weighting for neural machine translation domain adaptation," in *Proceedings of the First Workshop on Neural Machine Translation*, 2017, pp. 40–46.

[14] S. Zhang and D. Xiong, "Sentence weighting for neural machine translation domain adaptation," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 3181–3190.

[15] H. Sajjad, N. Durrani, F. Dalvi, Y. Belinkov, and S. Vogel, "Neural machine translation training in a multi-domain scenario," *arXiv preprint arXiv:1708.08712*, 2017.

[16] C. Kobus, J. Crego, and J. Senellart, "Domain control for neural machine translation," *arXiv preprint arXiv:1612.06140*, 2016.

[17] D. Britz, Q. Le, and R. Pryzant, "Effective domain mixing for neural machine translation," in *Proceedings of the Second Conference on Machine Translation*, 2017, pp. 118–126.

[18] J. Zeng, J. Su, H. Wen, Y. Liu, J. Xie, Y. Yin, and J. Zhao, "Multi-domain neural machine translation with word-level domain context discrimination," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 447–457.

[19] S. Gu, Y. Feng, and Q. Liu, "Improving domain adaptation translation with domain invariant and specific information," *arXiv preprint arXiv:1904.03879*, 2019.

[20] C. Chu and R. Dabre, "Multilingual multi-domain adaptation approaches for neural machine translation," *arXiv preprint arXiv:1906.07978*, 2019.

[21] Z.-Y. Dou, J. Hu, A. Anastasopoulos, and G. Neubig, "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)*, 2019, pp. 1417–1422.

[22] H. Jiang, C. Liang, C. Wang, and T. Zhao, "Multi-domain neural machine translation with word-level adaptive layer-wise domain mixing," *arXiv preprint arXiv:1911.02692*, 2019.

[23] M. Q. Pham, J.-M. Crego, F. Yvon, and J. Senellart, "Generic and specialized word embeddings for multi-domain machine translation," in *International Workshop on Spoken Language Translation*, 2019.

[24] H. Daumé III, "Frustratingly easy domain adaptation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007, pp. 256–263.