Exploring Domain-shared and Domain-specific Knowledge in Multi-Domain Neural Machine Translation

Zhibo Man zhiboman@bjtu.edu.cn
Yujie Zhang* yjzhang@bjtu.edu.cn
Yuanmeng Chen yuanmengchen@bjtu.edu.cn
Yufeng Chen yfchen@bjtu.edu.cn
Jinan Xu jaxu@bjtu.edu.cn
School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

Abstract

Currently, multi-domain neural machine translation (NMT) has become a significant research topic in domain adaptation machine translation, which trains a single model by mixing data from multiple domains. Multi-domain NMT aims to improve the performance of the low-resources domain through data augmentation. However, mixed domain data brings more translation ambiguity. Previous work focused on domain-general or domain-context knowledge learning, respectively. Therefore, there is a challenge for acquiring domain-general or domain-context knowledge simultaneously. To this end, we propose a unified framework for learning simultaneously domain-general and domain-specific knowledge, we are the first to apply parameter differentiation in multi-domain NMT. Specifically, we design the differentiation criterion and differentiation granularity to obtain domain-specific parameters. Experimental results on multi-domain UM-corpus English-to-Chinese and OPUS German-to-English datasets show that the average BLEU scores of the proposed method exceed the strong baseline by 1.22 and 1.87, respectively. In addition, we investigate the case study to illustrate the effectiveness of the proposed method in acquiring domain knowledge.

1 Introduction

In recent years, Neural Machine Translation (NMT) has shown excellent performance in various translation tasks, as evidenced by state-of-the-art (SOTA) results reported in studies such as (Bahdanau et al., 2015; Wu et al., 2016; Vaswani et al., 2017; Liu et al., 2021; Fernandes et al., 2022), among which Multi-domain NMT aims to construct a single NMT model with the ability to translate sentences across different domains (Wang et al., 2020). Mixed-domain data can improve cross-domain knowledge on low-resource domains by data augmentation. However, word ambiguity increases when we mix data from multiple domains. Therefore, a challenge remains in how to learn the domain-shared and domain-specific knowledge for multi-domain NMT.

To address the above problem, researchers design domain-shared (Zeng et al., 2018, 2019; Pham et al., 2019; Wang et al., 2020) and domain-specific knowledge learning mechanisms

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 99–110
September 4–8, 2023, Macau SAR, China.

Example 1 : Translation based on domain-shared knowledge learning	
Input (*Law*)	promotion centers and technology enterprise incubation base
Reference	促进中心和科技企业孵化✓基地
Mixed	促进中心和技术企业孵化✓基地
Single	促进中心和技术企业潜伏×基地
MDNMT (Jiang et al., 2020)	促进中心和技术企业培养×基地
Mixed data (*Education*)	incubation (孵化✓) lasts anywhere from 24-28 days.
Example 2 : Translation based on domain-specific knowledge learning	
Input (*Science*)	output power can never equal the input power for there always losses.
Reference	输出功率决不可能等于输入功率✓因为总有损耗。
Mixed	输出电力×从来不等于输入电力×由于常有损耗。
Single	输出功率从来不等于输入功率✓因为总有损耗。
MDNMT (Jiang et al., 2020)	输出功率从不等于输入功率✓由于总有损耗。
Mixed data (*News*)	which is also the source of most of Beijing's power (电力×) supply.

Table 1: Two English-Chinese translations of "incubation" and "power" with different models.

(Kobus et al., 2016; Britz et al., 2017; Jiang et al., 2020; Lee et al., 2022). Nevertheless, these strategies have inherent limitations, such as Example 1 in Table 1 shows that the word "incubation" is incorrectly translated to "潜伏" and "培养" by the Single and Jiang et al. (2020), respectively. On the one hand, this suggests that domain-specific data only has the translation "潜伏" in the Law domain. On the other hand, MDNMT (Jiang et al. (2020)) learns domain-specific features using a domain discriminator, resulting in translations relying on the results of the domain discriminator. In addition, the word is translated to "孵化" by Mixed, illustrating that mixing multiple domains' data can improve domain-shared knowledge. In contrast, Example 2 in Table 1 shows that "power" is incorrectly translated to "电力" by Mixed, showing that this model introduces the ambiguity of "电力" from the News domain, demonstrating the importance of domain-specific knowledge learning. Therefore, effectively representing domain-shared and domain-specific knowledge has become a key issue in multi-domain NMT.

To tackle the above issues, we found that some research work has proven that parameters play a key role in Multilingual NMT (Wang and Zhang, 2022; Sachan and Neubig, 2018) and Multilingual Speech Translation (Wang et al., 2022). In our work, we calculate the gradient from different domains based on cosine similarity as domain-specific parameters, and then obtain the domain-shared and domain-specific parameters of the model to represent the corresponding knowledge.

To summarize, our contributions are three-fold:

- To the best of our knowledge, our model is the first to explore domain-shared and domain-specific parameters of multi-domain NMT.
- We design different mechanisms to dynamically acquire domain-shared and domain-specific knowledge, respectively.
- Experimental results and analyses on multiple language pairs show that the proposed model improves over several baselines, then we further analyze the approach insights into its actual contributions in multi-domain NMT.

2 Related Work

According to the domain representation learning strategy, we divide it into domain-shared and domain-specific knowledge methods: **Domain-shared knowledge learning**: Mixed domain data is a simple and convenient method to obtain domain-shared knowledge. Additionally, Zeng et al. (2018) designed the domain-shared discriminator to learn cross-domain features. Pham et al. (2019) proposed isolating domain-agnostic from domain-specific lexical representations while sharing most of the network across domains. Furthermore, Wang et al. (2020) proposed two complementary supervision signals by leveraging the power of knowledge distillation and adversarial learning. **Domain-specific knowledge learning**: From a sentence-level perspective, training a discriminator to detect and embed the domain tag for a sentence has become the mainstream approach (Kobus et al., 2016; Britz et al., 2017; Tars and Fishel, 2018; Aharoni and Goldberg, 2020; Lee et al., 2022). Both Zeng et al. (2018) and Su et al. (2021) propose a maximum weighted likelihood estimation method, where the weight is obtained by masking the domain-aware word level to encourage the model to pay more attention to the domain-specific representation of words. Recent work proposes Domain Proportion to improve the adaptability of each word (Jiang et al., 2020; Lai et al., 2021; Zhang et al., 2021). Some works propose the domain proportion of words for MDNMT, where each word in the sentence has a corresponding proportion in each domain (Jiang et al., 2020; Zhang et al., 2021; Lai et al., 2021). However, this approach may also affect the performance of the domain discriminator in the target language to some extent, potentially leading to translation ambiguity.

Compared with the previous approaches, there are two salient features in our methods: (1) Our method can capture domain-shared and domain-specific knowledge simultaneously within the framework of multi-domain NMT instead of separately. (2) Our method learns domain-shared and domain-specific knowledge from the perspective of parameter learning, rather than utilizing domain discriminators.

3 Our model

Multi-domain NMT task: The objective of this task is to create a unified model using mixed-domain data, aiming to maximize performance across all domains (Wang et al., 2020). Specifically, there are J subsets, denoted as $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_J$. Each subset $\mathcal{D}_j$ consists of pairs of input-output sequences, represented as $\mathcal{D}_j = \{[\mathbf{x}_j^m, \mathbf{y}_j^m]\}_{m=1}^{M_j}$, where j indicates the domain and m denotes the index within the domain, M_j represents the number of all sentences in the j-th domain. The training objective can be formulated as follows:

$$\mathcal{L}_{MDNMT}(\theta) = \arg\max_{\theta} \frac{1}{J} \sum_{j=1}^{J} \mathcal{L}_j(\theta) \tag{1}$$

where θ represents the learnable parameters in the model, and $\mathcal{L}_j$ denotes the training objective for each specific domain.

3.1 Parameter Differentiation

Figure 1 gives the process of parameter differentiation (Wang and Zhang, 2022). This method enables the model to identify language-specific parameters during the training of multi-lingual NMT task. Shared parameters in this approach have the ability to dynamically specialize into different types, akin to cellular differentiation. Moreover, Wang and Zhang (2022) define the differentiation criterion as inter-task gradient cosine similarity (Yu et al., 2020; Wang et al., 2021). Consequently, parameters exhibiting conflicting inter-task gradients are more likely to be language-specific. As the key problem of multi-domain NMT is how to learn domain-shared and domain-specific knowledge. Inspired by the parameter differentiation of multi-lingual NMT. In

our work, we consider domain-specific knowledge learning as the process of parameter differentiation, the model determines which parameters should be domain-specific during training, and other parameters are domain-shared knowledge.

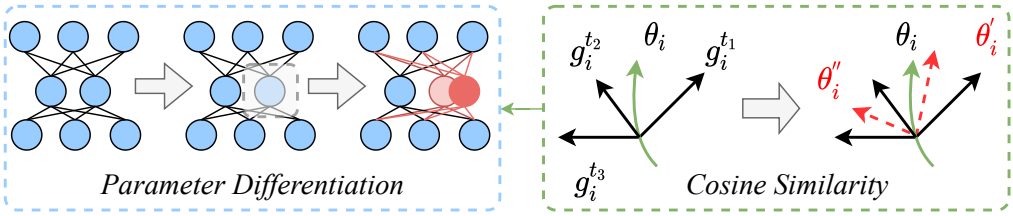


Figure 1: The process of Parameter Differentiation

3.2 The Framework of Our Model

As shown in Figure 2, we design a model consisting of an encoder-decoder based on Transformer (Vaswani et al., 2017). The blue box in Figure 2 represents domain-shared and domain-specific parameters in the encoder and decoder, respectively. Specifically, we obtain domain-shared and domain-specific parameters in the following ways: **(1) Domain-shared Knowledge Learning:** We design a unified multi-domain NMT framework that allows parameter sharing for each domain translation. Shared parameters of different layers in the encoder and decoder of Transformer are updated to exploit commonalities and differences across tasks. **(2) Domain-specific Knowledge Learning:** We calculate the gradient conflict of parameters in different layers of Transformer between different domains as the basis for domain-specific knowledge.

3.3 Domain-shared Knowledge Learning

Parameter sharing strategies are mainly used in multilingual *one-to-many* or *many-to-many* scenarios (Sachan and Neubig, 2018; Wang and Zhang, 2022; Wang et al., 2022). To be precise, all subtasks are passed through individual encoders and decoders simultaneously. As shown in Figure 2, we migrate parameter strategies from multilingual translation to multi-domain NMT. The parameters are described below:

Encoder Parameter Setting Individual encoder and decoder are set for source domain and target domain, respectively. The source domain parameters θ_{enc}=$\{W_{\text{K}}^{\text{enc}}, W_{\text{Q}}^{\text{enc}}, W_{\text{V}}^{\text{enc}}, W_{\text{F}}^{\text{enc}}, W_{\text{L}_1}^{\text{enc}}, W_{\text{L}_2}^{\text{enc}}\}$ are shared among different source domains, where $W_{\text{K}}^{\text{enc}}, W_{\text{Q}}^{\text{enc}}, W_{\text{V}}^{\text{enc}}, W_{\text{F}}^{\text{enc}}$ are the self-attention weights, $W_{\text{L}_1}^{\text{enc}}, W_{\text{L}_2}^{\text{enc}}$ are the FFN sublayer parameters.

Decoder Parameter Setting Regarding decoding stage, θ_{enc}, θ_{dec}=$\{W_{\text{K}_1}^{\text{dec}}, W_{\text{Q}_1}^{\text{dec}}, W_{\text{V}_1}^{\text{dec}}, W_{\text{F}_1}^{\text{dec}}, W_{\text{K}_2}^{\text{dec}}, W_{\text{Q}_2}^{\text{dec}}, W_{\text{V}_2}^{\text{dec}}, W_{\text{F}_2}^{\text{dec}}, W_{\text{L}_1}^{\text{dec}}, W_{\text{L}_2}^{\text{dec}}\}$ are shared for different target domains, where $W_{\text{K}_1}^{\text{dec}}, W_{\text{Q}_1}^{\text{dec}}, W_{\text{V}_1}^{dec}, W_{\text{F}_1}^{\text{dec}}$ are the self-attention weights of the decoder, θ_{enc}, $W_{\text{K}_2}^{\text{dec}}, W_{\text{Q}_2}^{\text{dec}}, W_{\text{V}_2}^{\text{dec}}, W_{\text{F}_2}^{\text{dec}}$ are parameters in the encoder-decoder attention sublayer, and $W_{\text{L}_1}^{\text{dec}}, W_{\text{L}_2}^{\text{dec}}$ are the feed-forward parameters shared in each decoder block.

3.4 Domain-specific Knowledge Learning

The main challenge in parameter differentiation is to define the criterion for differentiation, which assists in identifying shared parameters that should be specialized into specific types. Our approach defines the differentiation criterion using inter-task gradient cosine similarity, allowing us to identify parameters that encounter conflicting gradients and are likely domain-specific. Therefore, we first build the model as completely shared and initialize the parameters with a pre-trained model. Following prior work (Wang and Zhang, 2022), parameter differentiation consists of differentiation criterion and differentiation granularity.

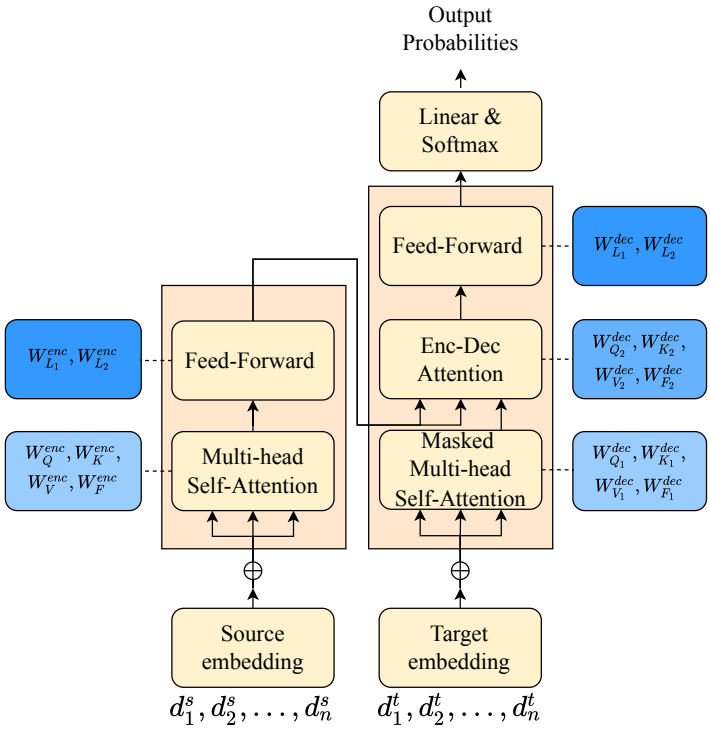


Figure 2: The Framework of Our Model

Differentiation Criterion: To assess the level of specialization for a shared parameter, we quantify its interference degree across three tasks using inter-task gradient cosine similarity. The i-th parameter θ_i in an multi-domain NMT model is shared by a set of tasks T_i, the interference degree $\mathcal{I}$ of the parameter θ_i is defined by:

$$\mathcal{I}(\theta_i, T_i) = \max_{t_j, t_k \in T_i} -\frac{g_i^{t_j} \cdot g_i^{t_k}}{\| g_i^{t_j} \| \| g_i^{t_k} \|} \tag{2}$$

where $g_i^{t_j}$ and $g_i^{t_k}$ are the gradients of task t_j and t_k respectively on the parameter .

Differentiation Granularity contains Layer{encoder layer, decoder layer}, Module{self-attention, FFN, Enc-Dec attention}, and Operation{linear projection, layer normalization}, "Layer granularity" refers to distinct layers within the model, while "Module granularity" refers to individual modules within a layer. On the other hand, "Operation granularity" encompasses the fundamental transformations in the model that possess trainable parameters. Each granularity level groups parameters into separate units for differentiation. For instance, at Layer level granularity, parameters within a layer are combined into a vector and differentiated as a single entity, which is known as a differentiation unit.

3.5 Training Method

In our method, we incorporate dynamic changes to the model architecture, resulting in distinct computational graphs for each task. To achieve this, we construct batches from multi-domain data, ensuring that each batch exclusively contains samples from a single task. This approach differs from training a conventional completely shared multi-domain NMT model, thus we train

the model of each domain from $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_J$. Specifically, to stabilize the training of θ_i on task t_i, we reinitialize the optimizer states by performing a warm-up update for those differentiated parameters(Wang and Zhang, 2022):

$$m_t^{'} = \beta_1 m_t + (1 - \beta_1)(g_i^{t_i}) \tag{3}$$

$$v_t^{'} = \beta_2 m_t + (1 - \beta_2)(g_i^{t_i})^2 \tag{4}$$

where m_t and v_t are the Adam states of θ_i, and $g_i^{t_1}$ is the gradient of task t_1 on the held-out validation data.

4 Experiments

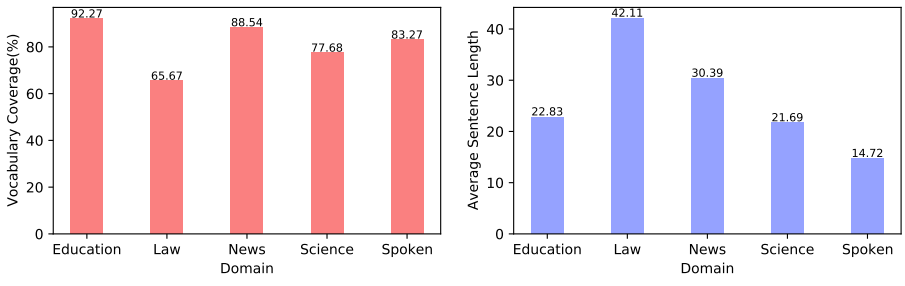
In our experiments, we aim to investigate the following research problems: (1) What is the improved performance of our method against previous work? (3) Can our model learn the more effective domain-shared and domain-specific knowledge?

4.1 Datasets

In our experiments, we use the following datasets for two machine translation tasks: **(1) English-to-Chinese**: We select UM-Corpus as multi-domain dataset [1] containing five domains: News, Spoken, Science, Education, and Laws. **(2) German-to-English**: We also choose OPUS [2] as multi-domain dataset containing five domains: Law, It, Koran, Medical, and Subtitles.

English-to-Chinese				German-to-English			
Domain	**Train**	**Dev**	**Test**	**Domain**	**Train**	**Dev**	**Test**
Education	444,608	1,996	462	It	222,297	1,888	2,000
Law	207,195	1,979	456	Koran	17,982	1,872	2,000
News	443,778	1,997	1,500	Law	467,309	1,861	2,000
Science	263,031	1,992	503	Medcial	248,099	1,861	2,000
Spoken	216,521	1,985	455	Subtitles	14,458,058	1,899	2,000

Table 2: The numbers of sentences in UM-Corpus and OPUS datasets.



(a) Vocabulary Coverage (b) Average Sentence Length

Figure 3: Statistics of English-to-Chinese dataset

[1] http://nlp2ct.cis.umac.mo/um-corpus/

[2] https://github.com/ZurichNLP/domain-robustness

Table 2 provides an overview of the dataset partition, and Figure 3 (a) and (b) shows the statistics of vocabulary coverage and average sentence length on English-to-Chinese dataset. We adopt the same pre-processing as the baseline model (Jiang et al., 2020). To process the English and German sentences, we employ the MOSES script (Koehn et al., 2007) for tokenization. For Chinese sentences, we utilize the Stanford Segmenter (Tseng et al., 2005) for word segmentation. To encode all sentences, we apply byte-pair encoding (BPE) (Sennrich et al., 2016). Specifically, for the German-to-English task, we train a joint BPE vocabulary with 32k merge operations. On the other hand, for the English-to-Chinese task, we separately train BPE vocabularies with a size of 32k for each language.

4.2 Comparative Models

We select seven models to compare the performance of results. Specifically, (1)-(2) are the strategy of domain data and (3)-(7) are the strategy of multi-domain NMT methods: **(1) Single**: This method only uses single domain data. **(2) Mixed**: This method uses mixed domain data. **(3) Disc**: Kobus et al. (2016) uses a sentence-level domain discriminator for domain representation learning. **(4) AdvL**: Britz et al. (2017) This approach is similar to Disc, except that when back-propagating from the discriminator to the encoder, gradients are reversed by multiplying. **(5) PAdvL**: Britz et al. (2017) is a combination of Disc and AdvL, spliting the embedding into half of Disc part and Adv part. **(6) WDCD**: Zeng et al. (2018) integrate Multi-Task Learning (MTL) and AdvL approaches by incorporating word-level domain contexts. **(7) WALDM**: Jiang et al. (2020) uses the domain proportion to learn the representation of each word. In addition, we reproduce the above comparison model based on the same parameter settings with fairseq[3] framework. Table 3 shows that the detailed hyperparameter settings.

Hyperparameter	Value
Epoch	50
Optimizer	Adam
(β_1, β_2)	(0.9, 0.98)
Beam Size	5
dropout rate	0.3
Learning Rate	5×10^{-4}
Tokens Per Batch	4096
Minimumm Learning Rate	10^{-9}
Feed-Forward Hidden State	1024
Encoder and Decoder Layers	6
Warmup Initial Learning Rate	5×10^{-4}
Word Embedding Dimensions	512

Table 3: Hyperparameter Settings

4.3 Main Results

The Results of English-to-Chinese Translation Task As shown at the top of Table 4. The BLEU scores of the *Single* on Law and News domains are 74.86 and 35.18, respectively, reaching the highest level compared to other models, reflecting that training on a single domain data avoids introducing noise. However, due to the limited data volume and complexity of content in the Science and Spoken domains, using solely the data from a single domain does not lead

[3]https://github.com/facebookresearch/fairseq

Task	Models	Domain					Avg↑	#Param↓
		Edu	Law	New	Sci	Spo		
English-to-Chinese	Single	30.03	**74.86**	**35.18**	17.93	28.11	37.22	-
	Mixed	35.13	62.76	32.07	27.43	28.14	37.11	145M
	Disc	34.87	62.90	31.92	27.44	28.70	37.17	145M
	AdvL	34.29	63.39	31.73	27.64	28.70	37.15	146M
	PAdvL	34.29	62.82	32.15	27.47	28.32	37.01	145M
	WDCD	33.15	60.87	33.17	27.03	28.40	36.62	211M
	WALDM	**35.87**	67.17	32.50	27.71	28.30	38.31	252M
	Ours	34.72	72.63	33.34	**28.06**	**28.89**	**39.53**	158M

Task	Models	Domain					Avg↑	#Param↓
		IT	Kor	Law	Med	Sub		
German-to-English	Single	66.58	20.07	**76.98**	71.76	**50.98**	57.27	-
	Mixed	64.65	40.03	74.04	69.16	49.77	59.77	70M
	Disc	64.54	40.29	74.62	67.45	49.09	59.20	177M
	AdvL	63.92	41.41	74.42	67.89	49.42	59.41	177M
	PAdvL	63.88	41.32	74.12	67.99	49.84	59.43	177M
	WDCD	63.89	41.11	74.03	67.97	49.68	59.34	204M
	WALDM	64.34	41.19	74.98	67.99	49.94	59.69	220M
	Ours	**67.02**	**42.51**	75.48	**71.92**	50.87	**61.56**	83M

Table 4: BLEU scores on the English-to-Chinese and German-to-English translation task. We bold the best performance results.

to optimal performance. Despite the Law domain having a data volume comparable to both domains, as shown in Figure 3 (a), Law domain data have longer text lengths than other domains, resulting in better performance when training the translation model separately (Chu and Wang, 2018). *Mixed* is a fundamental framework for multi-domain NMT, and it exhibits improvement compared to Single in Education, Science, and Spoken domains, suggesting that employing a mixed data training approach can enhance model performance in these particular domains. *Disc*, *Advl*, and *WADLM* bring+0.06,+0.04, and+1.20 on average BLEU scores compared to *Mixed*, indicating that multi-domain methods have improved with sentence-level or word-level domain discriminators. In addition, our method exceeds *WADLM*+1.22 BLEU scores. Among all the methods, our method is closest to the performance of Law and News domains of *Single*.

The Results of German-to-English Translation Task We further validate the effectiveness of our method on English-to-German datasets. From the bottom section of Table 4, it can be observed that our model achieves the highest average BLEU score of 61.56. These results provide further validation of the robustness and versatility of our model in the task of German-to-English translation. It should be noted that Single obtained the highest BLEU scores of 76.98 and 50.98 in the Law and Subtitles domains, respectively. Our method is closest to the performance of the Law and Subtitles domains of *Single*. In conclusion, the proposed method effectively learns domain-shared and domain-specific knowledge through parameter learning. It is expected to bring improvements when applied to other language translation tasks.

5 Analysis and Discussion

In this section, we first examine the effectiveness of differentiation. Then, we visualize the domain distribution and analyze the case study. It is worth noting that we mainly verify the English-to-Chinese translation task.

5.1 The effectiveness of Differentiation Granularity

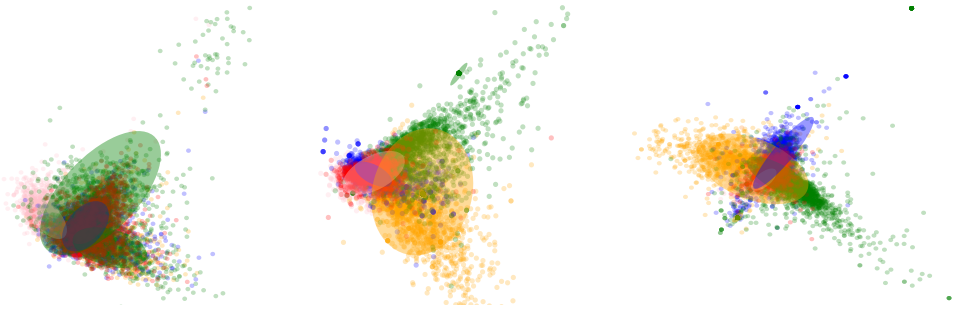
Models	Edu	Law	News	Sci	Spo	Avg	Δ
Mixed	**35.13**	62.76	32.07	27.43	28.14	37.11	-
Domain-specific w *Layer*	34.54	72.01	33.11	28.01	28.54	39.09	+1.98
Domain-specific w *Module*	34.32	72.43	33.02	27.89	28.23	39.18	+2.07
Domain-specific w *Operation*	34.72	72.63	**33.34**	**28.06**	**28.89**	**39.53**	**+2.42**

Table 5: Ablation study on English-to-Chinese, "w" represents with. "Domain-specific" represents domain-specific knowledge learning

We show the effectiveness of differentiation granularity in Table 5. From the average BLEU, "Domain-specific w *Operation*" has the highest improvement on Mixed compared to other granularity, indicating that the finer-grained parameter differentiation can learn more domain-specific knowledge, which is consistent with previous research (Wang and Zhang, 2022; Wang et al., 2022). Moreover, "Domain-specific w *Layer*" exceeds "Domain-specific w *Module*" +0.22, +0.09, and +0.12 on Education, News, and Spoken domains, respectively, indicating that coarse-grained method can obtain more domain knowledge in these domains than fine-grained method.

5.2 Visualization of Domain Distribution

To conduct the effectiveness of our proposed method in domain-specific knowledge learning, we utilize t-SNE (Van der Maaten and Hinton, 2008) to project representations of source sentences. The Visualization of *Mixed*, *WADLM* (Jiang et al., 2020), and Ours are shown in Figure 4 (a), (b), and (c), respectively.



(a) Mixed (b) WADLM (Jiang et al., 2020) (c) Ours

Figure 4: Green, Red, Pink, Orange, and Blue represents Education, Law, News, Science, and Spoken domains, respectively.

As shown in Figure 4, we can observe that *Mixed* does not effectively distinguish the sentences from different domains compared to and *WADLM* and Ours, as the clusters appear more

mixed and overlapping, showing that domain representation learning can improve the source sentences representation. In addition, our approach demonstrates a significant improvement as it successfully organizes the sentences into separate domain clusters with clear and distinct boundaries. This improvement shows that the domain-specific weight parameters of Ours enhances the encoder's ability to disambiguate and translate sentences accurately.

5.3 Case Study

We provide a case study to visually demonstrate the improvements made by our proposal. Example 1 of Table 6 shows the case from the Law domain on *Single*, *Mixed*, Jiang et al. (2020), and our model, respectively. When we mix domain data the word ambiguity introduced at the same time, Jiang et al. (2020) erroneously translates the word "incubation" to "培养", showing that the domain-shared knowledge always be ignored because of the domain discriminator. We can find that in this case, the proposed method corrects the ambiguous translation error, showing that our model can better capture domain-shared knowledge. In addition, Example 2 of Table 6 from the Science domain shows that the word "power" correctly translation into "功率" by *Single*, *Mixed*, Jiang et al. (2020) and our model. It further shows that our method can effectively learn domain-specific knowledge through parameter differentiation to obtain the correct domain when translating words.

Example 1 : Translation based on domain-shared knowledge learning	
Input (*Law*)	promotion centers and technology enterprise incubation base
Reference	促进中心和科技企业孵化✓基地
Mixed	促进中心和技术企业孵化✓基地
Single	促进中心和技术企业潜伏×基地
MDNMT (Jiang et al., 2020)	促进中心和技术企业培养×基地
Ours	促进中心和技术企业孵化✓基地
Example 2 : Translation based on domain-specific knowledge learning	
Input (*Science*)	output power can never equal the input power for there always losses.
Reference	输出功率决不可能等于输入功率✓因为总有损耗。
Mixed	输出电力×从来不等于输入电力×由于常有损耗。
Single	输出功率从来不等于输入功率✓因为总有损耗。
MDNMT (Jiang et al., 2020)	输出功率从不等于输入功率✓由于总有损耗。
Ours	输出功率从不等于输入功率✓因为总有损耗。

Table 6: Case Study

6 Conclusion and Future work

In this paper, we explore domain-shared and domain-specific knowledge in multi-domain NMT. Our method can simultaneously learn domain-shared and domain-specific parameters to resolve word ambiguity. Experimental results on two translation tasks show that our method can bring significant improvements. Further analyses confirm that our method can improve word ambiguity between domains. In future work, we will improve the gradient similarity method to further improve the accuracy of domain-specific parameters.

7 Acknowledgements

The present research was supported by the National Nature Science Foundation of China (No. 61876198, 61976015, 61976016). Yujie Zhang is is the corresponding author. We would like thank the anonymous reviewers for their constructive suggestions and insightful comments.

References

Aharoni, R. and Goldberg, Y. (2020). Unsupervised domain clusters in pretrained language models. In *ACL*.

Bahdanau, D., Cho, K. H., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Britz, D., Le, Q., and Pryzant, R. (2017). Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126.

Chu, C. and Wang, R. (2018). A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319.

Fernandes, P., Farinhas, A., Rei, R., De Souza, J., Ogayo, P., Neubig, G., and Martins, A. (2022). Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Jiang, H., Liang, C., Wang, C., and Zhao, T. (2020). Multi-domain neural machine translation with word-level adaptive layer-wise domain mixing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1823–1834.

Kobus, C., Crego, J., and Senellart, J. (2016). Domain control for neural machine translation. *arXiv preprint arXiv:1612.06140*.

Koehn, P., Federico, M., Shen, W., Bertoldi, N., Bojar, O., Callison-Burch, C., Cowan, B., Dyer, C., Hoang, H., Zens, R., et al. (2007). Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding. In *CLSP Summer Workshop Final Report WS-2006, Johns Hopkins University*.

Lai, W., Libovickỳ, J., and Fraser, A. (2021). Improving both domain robustness and domain adaptability in machine translation. *arXiv preprint arXiv:2112.08288*.

Lee, J., Kim, H., Cho, H., Choi, E., and Park, C. (2022). Specializing multi-domain nmt via penalizing low mutual information. *arXiv preprint arXiv:2210.12910*.

Liu, M., Yang, E., Xiong, D., Zhang, Y., Sheng, C., Hu, C., Xu, J., and Chen, Y. (2021). Exploring bilingual parallel corpora for syntactically controllable paraphrase generation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3955–3961.

Pham, M. Q., Crego, J.-M., Yvon, F., and Senellart, J. (2019). Generic and specialized word embeddings for multi-domain machine translation. In *International Workshop on Spoken Language Translation*.

Sachan, D. and Neubig, G. (2018). Parameter sharing methods for multilingual self-attentional translation models. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 261–271.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Su, J., Zeng, J., Xie, J., Wen, H., Yin, Y., and Liu, Y. (2021). Exploring discriminative word-level domain contexts for multi-domain neural machine translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1530–1545.

Tars, S. and Fishel, M. (2018). Multi-domain neural machine translation. *arXiv preprint arXiv:1805.02282*.

Tseng, H., Chang, P.-C., Andrew, G., Jurafsky, D., and Manning, C. D. (2005). A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, Q., Wang, C., and Zhang, J. (2022). Investigating parameter sharing in multilingual speech translation. *Proc. Interspeech 2022*, pages 1731–1735.

Wang, Q. and Zhang, J. (2022). Parameter differentiation based multilingual neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11440–11448.

Wang, Y., Wang, L., Shi, S., Li, V. O., and Tu, Z. (2020). Go from the general to the particular: Multi-domain translation with domain transformation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9233–9241.

Wang, Z., Tsvetkov, Y., Firat, O., and Cao, Y. (2021). Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *International Conference on Learning Representations*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., and Finn, C. (2020). Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836.

Zeng, J., Liu, Y., Su, J., Ge, Y., Lu, Y., Yin, Y., and Luo, J. (2019). Iterative dual domain adaptation for neural machine translation. 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 845–855.

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

Zhang, S., Liu, Y., Xiong, D., Zhang, P., and Chen, B. (2021). Domain-aware self-attention for multi-domain neural machine translation. *Proc. Interspeech 2021*, pages 2047–2051.