ICL: Iterative Continual Learning for Multi-domain Neural Machine Translation

Zhibo Man, Kaiyu Huang, Yujie Zhang[†], Yuanmeng Chen, Yufeng Chen, Jinan Xu
Beijing Key Lab of Traffic Data Analysis and Mining,
School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China
`{zhiboman, kyhuang, yjzhang, yuanmengchen, yfchen, jaxu}@bjtu.edu.cn`

Abstract

In a practical scenario, multi-domain neural machine translation (MDNMT) aims to continuously acquire knowledge from new domain data while retaining old knowledge. Previous work separately learns each new domain knowledge based on parameter isolation methods, which effectively capture the new knowledge. However, task-specific parameters lead to isolation between models, which hinders the mutual transfer of knowledge between new domains. Given the scarcity of domain-specific corpora, we consider making full use of the data from multiple new domains. Therefore, our work aims to leverage previously acquired domain knowledge when modeling subsequent domains. To this end, we propose an **I**terative **C**ontinual **L**earning (**ICL**) framework for multi-domain neural machine translation. Specifically, when each new domain arrives, **(1)** we first build a pluggable incremental learning model, **(2)** then we design an iterative updating algorithm to continuously update the original model, which can be used flexibly for constructing subsequent domain models. Furthermore, we design a domain knowledge transfer mechanism to enhance the fine-grained domain-specific representation, thereby solving the word ambiguity caused by mixing domain data. Experimental results on the UM-Corpus and OPUS multi-domain datasets show the superior performance of our proposed model compared to representative baselines.

1 Introduction

Multi-domain neural machine translation (MDNMT) aims to train a single model with mixed-domain data and has shown great performance in recent years (Lu and Zhang, 2023; Zhang et al., 2023; Man et al., 2024a,b; Wu et al., 2024). Traditional work (Chu et al., 2017) fine-tunes a model for each domain based on an original model. However, fine-tuning models individually for each domain increases maintenance costs and limits the ability to utilize diverse knowledge. Therefore, a smarter and more practical approach is to continuously update the original translation model with new domain data, rather than fine-tuning multiple domain-specific models.

In this scenario, the key challenge is how to continuously and sequentially learn new domain knowledge while avoiding catastrophic forgetting, which is a challenge in the field of Continual Learning (CL). Currently, existing studies can be divided into three lines: **(i) Replay-based methods**: These methods retain part or all training data from previous tasks (de Masson d'Autume et al., 2019; Peng et al., 2020; Liu et al., 2021; Kanwatchara et al., 2021; Garcia et al., 2021). **(ii) Regularization-based methods**: These methods aim to approximate the loss incurred on previous tasks and are usually in quadratic form (Luong and Manning, 2015; Castellucci et al., 2021; Gu et al., 2022; Shao and Feng, 2022). These methods effectively learn the knowledge between different domains. However, they cannot completely avoid the problem of catastrophic forgetting. To address this limitation, researchers propose **(iii) Parameter isolation-based methods**: These methods design separate pluggable modules and freeze all original parameters to completely retain the performance on previous tasks (Bapna and Firat, 2019; Madotto et al., 2021; Huang et al., 2022, 2023b; Lu and Zhang, 2023). These methods are also called "plug and play" (Nguyen et al., 2017; Dathathri et al., 2019; Tiong et al., 2022). However, the parameters introduced for each domain in these methods are learned independently, so subsequent domains cannot leverage knowledge from previous domains.

To deal with this challenge, we propose an **I**terative **C**ontinual **L**earning (**ICL**) method for multi-domain neural machine translation based on a parameter isolation framework. In this work, we

---

[†] Yujie Zhang is the corresponding author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 7732–7743
November 12-16, 2024 ©2024 Association for Computational Linguistics

aim to address the challenge of utilizing knowledge across multiple new domains. Specifically, our approach consists of three stages: First, we train original and incremental models for the original and new domains, respectively. Then we attach parameters to the original model to create pluggable modules via domain knowledge transfer, including the word embeddings layer and FFN layers. Finally, as new domains arrive, the pluggable modules from the previous domain are continuously integrated into the original model for the new domain. Compared to previous work, our approach not only prevents catastrophic forgetting but also makes full use of the knowledge across domains. Moreover, comprehensive experimental results and analyses on multiple language pairs demonstrate that our proposed model improves across all baselines. In conclusion, our contribution can be summarized as follows:

- We propose a novel iterative incremental learning framework that acquires knowledge from new domains and updates the original model to enable knowledge transfer between new domains.
- Our framework adapts stably to multiple domains, even when the learning order changes, demonstrating its robustness in iterative incremental learning.
- We further design a domain knowledge transfer strategy to resolve word ambiguities and enhance domain-specific representations during incremental learning.

2 Related Work

Recent work on continual learning of NMT can be divided into three categories:

Replay-based methods. The first category of methods requires retaining part or all of the training data from previous tasks (Lakew et al., 2018; Sun et al., 2019; Feyisetan et al., 2020; Liu et al., 2021; Garcia et al., 2021). However, these methods result in higher training costs, particularly when applied to large-scale pre-trained multi-lingual neural machine translation (MNMT) models. Furthermore, previous training data may be inaccessible due to privacy concerns or storage limitations. In contrast, our approach does not require additional data and offers greater flexibility for continual learning.

Regularization-Based Methods. The second category of works alleviates catastrophic forgetting by adding penalty terms to the learning objective, balancing performance between previous and new tasks (Kirkpatrick et al., 2017; Thompson et al., 2019; Castellucci et al., 2021). In this scenario, Gu et al. (2022) uses a hard constraint to update parameters in regions with a low risk of forgetting. Shao and Feng (2022) introduce an online knowledge distillation approach, where previous models assist in training the current model. In contrast to these methods, our framework naturally prevents catastrophic forgetting.

Parameter-Isolation Based Methods. The third category of works designs separate pluggable modules and freezes original parameters to retain performance on previous tasks (Bapna and Firat, 2019; Madotto et al., 2021; Zhu et al., 2022). In particular, Huang et al. (2023b) propose a knowledge transfer method to efficiently adapt MNMT models to diverse incremental language pairs. Furthermore, Huang et al. (2023a) propose a two-stage approach that encourages original models to acquire language-agnostic multilingual representations from new data while preserving the model architecture without adding new parameters. However, the parameters introduced for each domain in this method are independent of one another. By contrast, our method aims to better utilize knowledge across incremental domains, preventing catastrophic forgetting.

3 Method

In this work, our goal is to address the challenge of how to learn knowledge across multiple new domains. As shown in Figure 1 (a), the original model needs to build a single pluggable module for each new domain, and the knowledge between multiple new domains cannot learn from each other within the framework of Pluggable Incremental Learning (PIL) (Huang et al., 2023b). However, in real-world scenarios, the domain-specific corpus is typically low-resource, and new domains often arrive in stages over time. Consequently, we aim to effectively utilize knowledge from multiple new domains. Specifically, we focus on learning knowledge from different domains by iteratively updating the original model, as shown in Figure 1 (b). Compared with PIL, our approach differs in two key aspects: domain knowledge transfer strategy (§ 3.2) and iterative continual learning framework (§ 3.3).

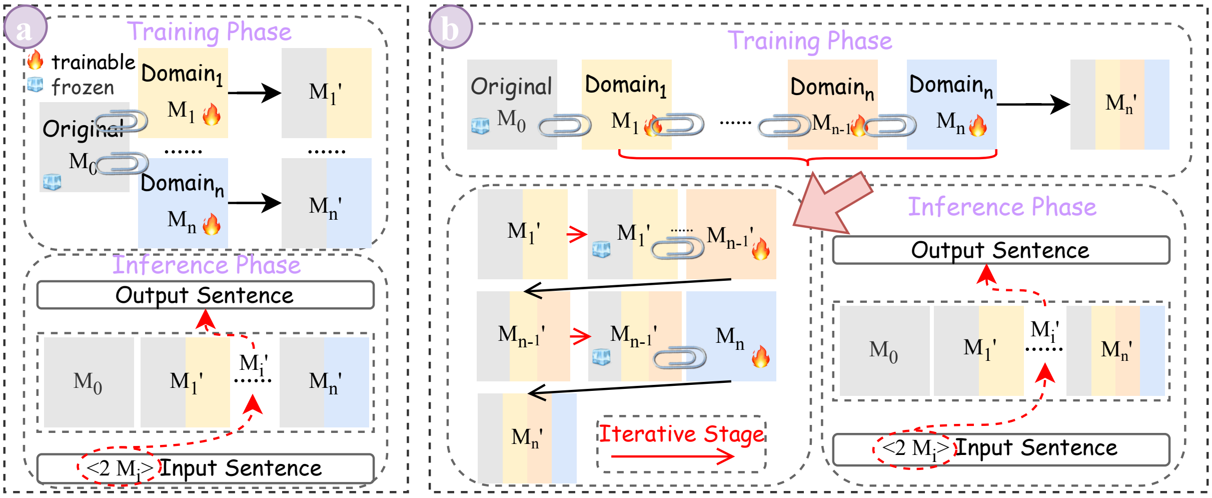

Figure 1: Comparison of (a) Pluggable Incremental Learning (PIL) and (b) Iterative Continual Learning (ICL) frameworks. Our motivation is to leverage knowledge from multiple domains; our base model is updated with new domain data. The parameters of each updated model are frozen.

3.1 Task Definition

In our work, the scenario of continual learning involves adding new domain data one by one based on the original MDNMT model while retaining translation qualities on original language pairs without accessing previous training data. We define our scenario by referencing the approach of MNMT (Huang et al., 2023b). Formally, an MDNMT model is trained on initially selecting a set of available parallel data $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_J\}$ from J different domains. Accordingly, the source sentence is denoted as x, while the target sentence is denoted as y, and $\mathcal{D}_j$ represents the original domain corpus on the i-th domain. The training objective :

$$\mathcal{L}(\theta) = \sum_{\mathcal{D}_j \in \mathcal{D}} \sum_{(x,y) \in \mathcal{D}_j} \mathrm{logp}(\mathrm{y}|\mathrm{x};\theta) \tag{1}$$

where θ represents the trainable parameters of the external models.

Continual learning is updating the original MDNMT model on an updated set of parallel data $\mathcal{D}^{(ALL)} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_J, ..., \mathcal{D}_N\}$, which covers N domains. The training data $\mathcal{D}$ of original model is often unavailable. Therefore, we use each domain data $\mathcal{D}^{(New)} = \{\mathcal{D}_{J+1}\}, ..., \{\mathcal{D}_N\}$ to continually train the original model, and $\mathcal{D}_i$ is the incremental parallel training corpus on the i-th domain, and the optimization objective:

$$\mathcal{L}(\theta) = \sum_{\mathcal{D}_i \in \mathcal{D}^{(New)}} \sum_{(x,y) \in \mathcal{D}_i} \mathrm{logp}(\mathrm{y}|\mathrm{x};\theta) \tag{2}$$

where the number of domains supported by the MDNMT model increases from J to N.

3.2 Domain Knowledge Transfer via Pluggable Modules

To better learn the knowledge of new domains into original model, inspired by PIL (Huang et al., 2023b), we also introduce two types of pluggable modules to bridge the representation gap of different domains, and we further incorporate domain knowledge into the pluggable module to enhance the features of different domains.

Domain Embedding Transfer. In multilingual neural machine translation, the significant differences between languages due to different characters lead to considerable variation (Dabre et al., 2020). Therefore, the purpose of MNMT is to address the alignment between different languages. However, in MDNMT, the corpora from different domains belong to the same language pair, and words can become ambiguous due to varying contexts. Thus, MDNMT aims to resolve the ambiguity of words across different domains (Chu and Wang, 2018). For example, the word “column” has the translation “柱” and “列” in Spoken and Education domains, respectively. This phenomenon reflects the domain-specific representation with context in different domains. Therefore, we expand the vocab in the embedding layer and concatenate the embedding of different domain tokens between the original and incremental models, enhancing the domain representation features by adding domain tags. This mechanism can enhance the domain features, thereby relieving the word ambiguity caused by the mixed domain data.

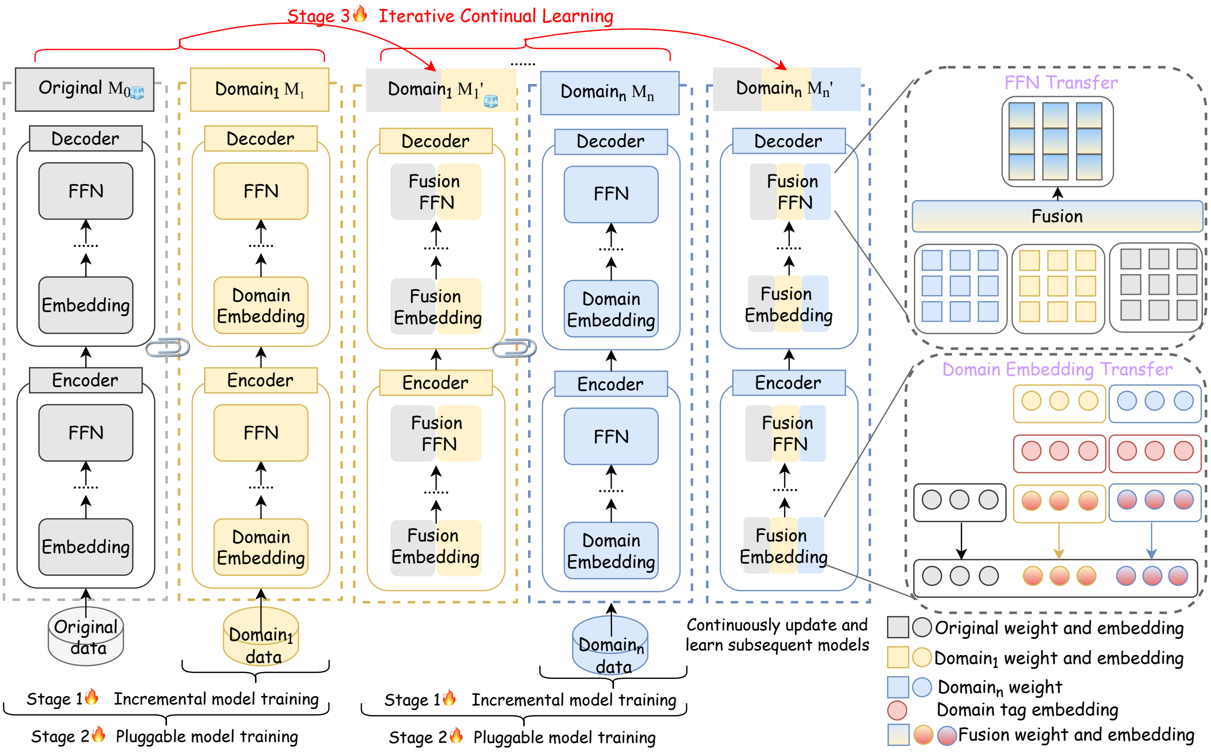
Stage 3🔥 Iterative Continual Learning
Original M_0
Domain$_1$ M_1
Domain$_1$ M_1'
Domain$_n$ M_n
Domain$_n$ M_n'
Decoder
FFN
Embedding
Encoder
Domain Embedding
Fusion FFN
Fusion Embedding
Original data
Domain$_1$ data
Domain$_n$ data
FFN Transfer
Fusion
Domain Embedding Transfer
Continuously update and learn subsequent models
Original weight and embedding
Domain$_1$ weight and embedding
Domain$_n$ weight
Domain tag embedding
Fusion weight and embedding
Stage 1🔥 Incremental model training
Stage 2🔥 Pluggable model training

Figure 2: This framework consists of three stages: Incremental Model Training, Pluaggable Model Training, and the Inference Phase. Moreover, it is important to highlight another aspect of our work: for all updated original models, we freeze their parameters to retain the performance in the training phase.

Feed-Forward Network Transfer. FFN has previously been investigated by (Sukhbaatar et al., 2019; Huang et al., 2023b), inspired by these work, we utilize the continuous representations in the FFN layers of an external model, containing valuable language knowledge, and transfer this knowledge into the FFN layers of the original model, forming a pluggable module. By combining the outputs of the original FFN layers $\mathrm{FFN}_1(\mathbf{H})$ with these injected modules $\mathrm{FFN}_2(\mathbf{H})$, we share domain knowledge, addressing the representation gap through adaptation in the Feed-Forward layers. The resultant fused FFN output:

$$\mathbf{H}^{(\mathrm{f})} = \mathrm{FFN}_1(\mathbf{H}) + \mathrm{FFN}_2(\mathbf{H}) \tag{3}$$

3.3 Training and Inference

We propose a new framework ICL to continuously iterative and update knowledge from new domains. As shown in Figure 2, our model consists of three essential stages:

Stage 1: Incremental Model Training. We utilize each new domain data to train a incremental machine translation model:

$$\mathcal{L}_{\mathcal{D}^{(New)}}(\theta') = \sum_{\mathcal{D}_i\in\mathcal{D}^{(New)}} \sum_{(x,y)\in\mathcal{D}_i} \mathrm{logp}(\mathrm{y}|\mathrm{x};\theta') \tag{4}$$

where we only retain the parameters in the embedding layer θ'_e and FFN layers θ'_f of the external model as the pluggable modules for the next training stage.

Stage 2: Pluggable Model Training. After the first stage, we migrate the trainable parameters θ'_e and θ'_f in the new model to the original model. The specific training function is:

$$\mathcal{L}_{\mathcal{D}^{(New)}}(\theta'_e, \theta'_f) = \sum_{\mathcal{D}_i\in\mathcal{D}^{(New)}} \sum_{(x,y)\in\mathcal{D}_i} \mathrm{logp}(\mathrm{y}|\mathrm{x};\theta'_{\mathrm{e}}, \theta'_{\mathrm{f}}) \tag{5}$$

Stage 3: Iterative Continual Learning. When a new domain arrives, the pluggable model is updated, which then becomes the new base model for future domain integration. As outlined in Algorithm 1, we start with an initial model, M_0. Upon receiving the first domain data, we train an incremental model, M_1, as described in Stage 1. This

Algorithm 1 Iterative Continual Model Training

```
Require: $\mathrm{M}_0$, $\mathrm{M}_i$, $N$;
Ensure: $\mathrm{M}_0$;
  for $i$ from 1 to $N$ do
    Extract parameters $\theta_e$ and $\theta_f$ from $\mathrm{M}_0$
    Transfer parameters $\theta_e$ and $\theta_f$ to $\mathrm{M}_\mathrm{i}$
    Update the fused model $\mathrm{M}_\mathrm{i}$->$\mathrm{M}_\mathrm{i}^{'}$
  end for
  Output $\mathrm{M}_\mathrm{i}^{'}$
```

process yields a pluggable model, $\mathrm{M}_1^{'}$. We then update the original model by replacing it with $\mathrm{M}_1^{'}$. Lastly, Stage 2 is repeated to train a new plug-in model, M_2, by updating the parameters $\theta_e^{'}$ and $\theta_f^{'}$.

Inference. For the inference phase, as shown in Figure 1, this model can select the appropriate inference model based on the input labels. This framework preserves the performance of each original model while preventing catastrophic forgetting.

4 Experiments

We conduct experiments for our framework (ICL) to explore the following questions: **(i)** Can ICL learn more knowledge between new domains under low-resource domain scene? (§ 4.3) **(ii)** How is the performance of ICL on original domains compared with previous work? (§ 4.3) **(iii)** Which factors affect the performance of ICL? (§ 4.5, § 4.6)

4.1 Datasets

We conduct experiments on the English-to-Chinese[1] multi-domain translation tasks (Tian et al., 2014). For the German-to-English translation task, we utilize the OPUS[2] multi-domains dataset (Kobus et al., 2017), which has been resplit by (Aharoni and Goldberg, 2020). These datasets have been widely used in previous research (Zeng et al., 2018; Jiang et al., 2020). The overall statistics of the datasets are listed in Table 1. The data similar to the test set are filtered out for fair comparison.

Domain Choice. To better approximate the real scenario, we split the original and incremental domains according to the size of the dataset. Notably, the data volume of Microblog being extremely low-resource. The detailed division of the original and incremental domains is shown in Table 1. A detailed description and comprehensive information regarding the datasets for all domains can be found in Appendix A.

English-to-Chinese			
Original	Train	Valid	Test
Education	440K	1996	462
News	440K	1997	1500
Thesis	290K	2000	624
Incremental	Train	Valid	Test
Science	260K	1992	503
Subtitles	220K	2000	596
Spoken	210K	1985	455
Microblog	4.6K	200	266
German-to-English			
Original	Train	Valid	Test
Subtitles	470K	1,899	2,000
Law	434K	1,861	2,000
Medical	233K	1,873	2,000
Incremental	Train	Valid	Test
It	211K	1,888	2,000
Koran	16K	1,872	2,000

Table 1: The statistics of our datasets. The number in Valid/Test columns denotes the amount of sentence pairs in each domain.

4.2 Implementation Details

Baselines for Comparison. We compare our method (ICL) with different multi-domain neural machine translation adaptation methods. All methods utilize the preprocessing script of a shared BPE model with 32k tokens based on the Sentencepiece library[3]. These baselines can be listed as follows:

Base systems. Single, which trained on a single incremental domain parallel data based on Transformer (Vaswani et al., 2017). Mixed, which trained on the mix of original domain parallel data based on Transformer.

Multi-domain adaption NMT Baselines. To compare the effectiveness of learning the knowledge from new domains, we compare our approach with Multi-domain adaption NMT Baselines, including: FT (Luong and Manning, 2015), which first trains the NMT model on original domain training corpus, and then fine-tunes it by using incremental domain training corpus. Adapter (Bapna and Firat, 2019), which introduces extra parameters in each FFN layer of the original MNMT model.

---

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

[2] http://opus.nlpl.eu/

[3] https://github.com/google/sentencepiece

Incremental domains (English-to-Chinese)										
Methods	**Subtitles**		**Science**		**Spoken**		**Microblog**		**AVG**	
	`BLEU`	`COMET`	`BLEU`	`COMET`	`BLEU`	`COMET`	`BLEU`	`COMET`	`BLEU`	`COMET`
Single	16.73	0.244	17.65	0.241	21.86	0.301	5.55	0.023	15.45	0.202
FT	17.81	0.261	18.83	0.267	22.13	0.312	13.11	0.126	17.97	0.242
Adapter	17.10	0.252	18.05	0.254	22.33	0.324	12.69	0.120	17.54	0.238
PIL	18.23	0.270	19.65	0.291	23.54	**0.347**	14.20	0.161	18.91	0.267
ICL (Ours)	**18.65**	**0.282**	**20.13**	**0.311**	**23.88**	0.342	**15.14**	**0.172**	**19.45***	**0.277***

Incremental domains (Chinese-to-English)										
Methods	**Subtitles**		**Science**		**Spoken**		**Microblog**		**AVG**	
	`BLEU`	`COMET`	`BLEU`	`COMET`	`BLEU`	`COMET`	`BLEU`	`COMET`	`BLEU`	`COMET`
Single	11.02	0.145	10.05	0.110	14.70	0.130	2.13	0.012	9.48	0.099
FT	12.61	0.150	11.86	0.119	15.48	0.138	10.25	0.117	12.55	0.131
Adapter	12.14	0.149	10.99	0.092	14.98	0.132	10.10	0.113	12.05	0.122
PIL	13.44	0.163	12.01	**0.124**	16.10	**0.144**	11.48	0.129	13.26	0.140
ICL (Ours)	**13.65**	**0.169**	**12.36**	0.121	**16.38**	0.140	**12.16**	**0.136**	**13.64***	**0.142**

Table 2: BLEU and COMET scores on the English-to-Chinese and Chinese-to-English translation directions. We bold the best performance results. The order of domains is that *Subtitles->Science->Spoken->Microblog*. Results with $*$ are statistically (Koehn, 2004) better than the "PIL" with p < 0.01. Others are our re-implementation results using the released code with the same setting for a fair comparison. The highest score is highlighted in **bold**.

Continual Learning Baselines. To compare the effectiveness of mitigating catastrophic forgetting in the original domain, we further compare our method with state-of-the-art methods in continual learning, including: Replay (Sun et al., 2019), which involves creating pseudo data for the original language pairs and jointly training new language pairs using both the pseudo data and incremental training data. EWC (Thompson et al., 2019), which uses the Fisher information matrix to model the importance of parameters, applying more constraints to the crucial ones to ensure they remain close to their original values. PIL (Huang et al., 2023b), which proposes pluggable incremental learning for multilingual machine translation.

Training. Our experiments are conducted under fairseq[4] (Ott et al., 2019) framework, we built on Transformer model (Vaswani et al., 2017) which has 6 encoder and decoder layers with embedding dimension 512, feed-forward dimension 1024, and attention head 4. All experiments are trained with label smoothing cross-entropy loss with a smoothing parameter of 0.1. We use 8 NVIDIA P100 GPU and Adam optimizer with an initial learning rate of 1e-4. In our experiments, we do not use ensembles or n-best reranking, and training is stopped when there is no performance improvement.

Evaluation. We set beam size to 5, and we use the SacreBLEU script for English[5]. For Chinese, we calculate the BLEU at the character granularity, which is consistent with previous work (Jiang et al., 2020; Wang et al., 2020). The `eamt22-cometinho-da` model is used to generate the COMET[6] scores, the scope is 0-1. In particular, we use the paired bootstrap resampling methods (Koehn, 2004) for the statistical significance test.

4.3 Main Results

The performance of English-to-Chinese task Table 2 presents the performances of five systems across four domains in two translation directions. The results show that our approach outperforms several baselines in terms of average BLEU and COMET scores for all incremental translation directions. In comparison with Single, other methods (FT, Adapter, PIL, and ICL) based on the original model achieve significant improvement, demonstrating that the original model trained on large-scale data brings rich translation knowledge. For example, FT exceeds the Single on Microblog with +7.56 BLEU. Because the data in the Microblog domain is extremely low-resource (4.6k sentences, as shown in Table 1), it is evident that fine-tuning

---

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

[5] Signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0

[6] https://github.com/Unbabel/COMET

Incremental domains (German-to-English)				
Methods	**IT**		**Koran**	
	`BLEU`	`COMET`	`BLEU`	`COMET`
Single	41.33	0.689	18.14	0.132
FT	42.56	0.693	19.37	0.145
Adapter	41.23	0.686	19.01	0.141
PIL	42.68	0.697	19.71	0.149
Ours	**42.91**	**0.712**	**19.94**	**0.158**

Table 3: BLEU and COMET scores on the German-to-English translation direction. We bold the best performance results. The order of domains is that *IT->Koran.*

the original model leads to significant performance improvements. Compared with the FT, Adapter, and PIL, ICL shows a better ability to acquire new knowledge between different domains. Additionally, overall performance for English-to-Chinese is better than Chinese-to-English. This trend is consistent with existing studies (Wang et al., 2020), as the BLEU score for Chinese needs to be calculated based on characters due to the unique characteristics of Chinese.

The performance of German-to-English task To further show the advantages of our method, we compare the results of ICL with existing works on the German-to-English translation task. Table 3 presents detailed comparisons, we can see that our model reaches the highest average BLEU score and COMET on IT and Koran, respectively. The results of German-to-English translation tasks further validate the robustness and versatility of our model. Similarly, there is a further performance improvement in the relatively low-resource Koran. This indicates that IT domain knowledge has been utilized and demonstrates the stability of our method.

Degeneration in Continual Learning As shown in Table 4, to demonstrate the reliability and effectiveness of our approach, we examined the degradation in performance on the original domains, comparing it with various methods, our method can maintain the performance of original model. Additionally, the findings reveal that replay-based and regularization-based methods still exists significant degradation in the original translation directions when the original data is not available.

Original domains (English-to-Chinese)			
Methods	**Education**	**News**	**Thesis**
Mixed	35.35	34.79	36.23
Replay	-1.45	-2.16	-2.91
EWC	-0.81	-2.55	-1.13
Ours	**0.00**	**0.00**	**0.00**

Original domains (German-to-English)			
Methods	**Law**	**Medical**	**Subtitles**
Mixed	54.73	50.89	27.74
Replay	-3.32	-3.56	1.24
EWC	-1.12	-2.21	-0.97
Ours	**0.00**	**0.00**	**0.00**

Table 4: Results on the original domains with different continual learning. The darker the color, the closer the performance is to the original domains. The highest score is highlighted in **bold**.

Incremental domains (English-to-Chinese)					
	Sub	**Sci**	**Spo**	**Mic**	**AVG**
Base	20.14	22.45	25.98	18.24	21.70
PIL	22.34	24.67	27.19	19.78	23.50
ICL	**23.06**	**25.76**	**28.80**	**23.33**	**25.24**

Table 5: Results of English-to-Chinese with pre-trained model mBART-nn on the incremental domains.

4.4 Results on Pre-trained Models

As shown in Table 5, we leverage pre-trained multi-lingual machine translation model mBART-nn (Tang et al., 2020) as the external model and investigate the effectiveness of our method, and the average BLEU score of ICL (ours) across four domains exceeds PIL with +1.74 BLEU compared to PIL. It further proves that the large pre-trained model contains more useful knowledge and the effectiveness of our approach.

4.5 Ablation Study

To give a better understanding of our framework ICL, we conduct several ablation study in this section. These studies are taken on the English-to-Chinese translation task.

Effect of Domain knowledge Transfer Strategy To investigate the effectiveness of our approach with different transfer strategy (i.e., Embedding and FFN), we compare our method with four strategies: S1-S4. According to Figure 3, S1 yields relatively bad results, indicating the significance of domain knowledge. In contrast to S1, our approach (S2 and S3) illustrates that each pluggable module can achieve significantly better optimization when treated separately. Furthermore, S4 achieves superior performance across four domains by uti-

Domain Orders	Subtitles	Science	Spoken	Microblog	AVG
Microblog->Science->Subtitles->Spoken	18.44	20.01	**24.33**	13.21	19.00
Subtitles->**Microblog**–>Science->Spoken	18.65	20.24	24.12	14.14	19.29
Subtitles->Science->**Microblog**->Spoken	18.65	20.13	23.02	14.69	19.12
Subtitles->Science->Spoken->**Microblog**	**18.65**	**20.13**	23.88	**15.14**	**19.45**

Table 6: Effect of different domain orders. The performance remains unchanged when the order in a particular domain does not change. The highest score is highlighted in bold.

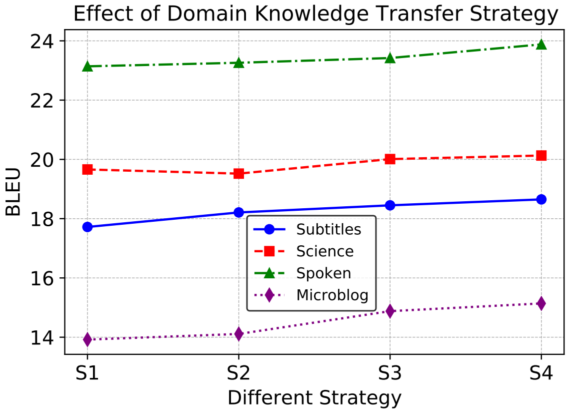

Figure 3: Effect of Domain knowledge Transfer Strategy. S1-S4 represents for the different strategies. Specifically, S1: Embedding (✗), FFN (✗); S2: Embedding (✓), FFN (✗); S3: Embedding (✗), FFN (✓); S4: Embedding (✓), FFN (✓).

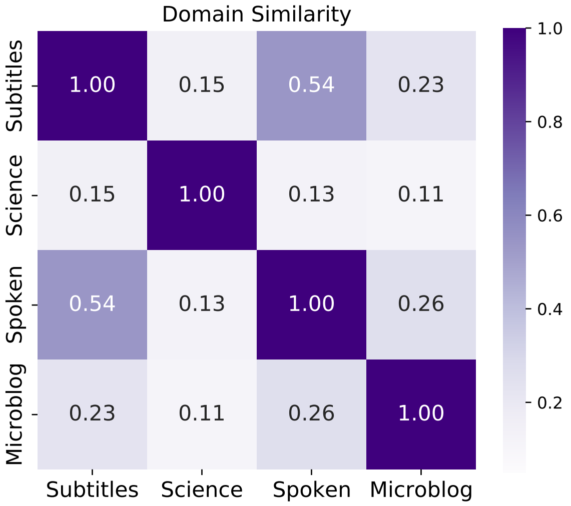

Figure 4: Domain Similarity between incremental domains (Chinese test sentences). We calculate the domain similarity based on cosine. The deeper the color, the better the similarity.

lizing both Embedding and FFN through domain knowledge transfer.

Effect of Different Domain Orders

To verify the stability of our method, we attempt to change the domain order, centering on the Microblog. As shown in Table 6, we primarily draw conclusions by observing the performance changes in Microblog: Changing the order of domains affects the performance of subsequent domains but not the performance of previous domains. For example, the performance of Microblog varies with its position. The later a domain is, the more knowledge it acquires from preceding domains.

4.6 Analysis and Discussion

To further investigate the effectiveness of our approach, we conduct more in-depth studies, divided visualization and case study.

Domain Knowledge Transfer and Similarity

We observed an interesting phenomenon in Table 6 where performance significantly changes when the order of preceding and succeeding domains changes, such as in the Spoken. We speculate this is related to similarities between domains. To investigate this, we calculate the cosine similarity between target languages on different domain training sets. As shown in figure 4, there is a high similarity between the Spoken and the Subtitles. Therefore, when the spoken domain follows the subtitles domain, performance improves.

Case Study

Within the example in Table 7 shows that two translation cases selected from the test datasets in the Spoken and Science domains. For the first case, the English word "columns" is difficult to translate for baseline model as "柱". However, this English word has multiple translations in different domains, for example, it may be translated into "列" in the Science domain. Benefiting from the domain knowledge transfer mechanism, our model can generate the correct translation. Similar to the second case, the English word "column" appears to translation. Our model can successfully translate them, further showing that our method can effectively learn multiple new domain knowledge.

Domain	**Spoken**
Src	Between the columns were light, hollow panel walls of double brick.
Ref	柱间是两砖厚的轻质空心间墙。
PIL	列间是双层砖砌成的轻质空心墙。
ICL (Ours)	柱子之间是双层砖砌成的轻质空心板墙。
Domain	**Science**
Src	4 different information types in the column space left in the margin.
Ref	4种不同的信息类型位于页边空白处的列空间中。
PIL	4种不同的信息类型留在空白处(?)。
ICL (Ours)	4种不同的信息类型位于页边空白处的列空间中。

Table 7: English-to-Chinese translation cases. Blue indicates the correct translation, while red indicates an incorrect translation.

5 Conclusion

In this work, we propose an ICL method for multi-domain neural machine translation, this method can makes full use of knowledge between multiple new domains. Specifically, our framework constantly updated original model to obtain the new knowledge. Moreover, we design the domain knowledge transfer mechanism to enhance the domain-specific features of word. Experimental results demonstrate the effectiveness of our approach.

Limitations

Although our method has achieved outstanding performance compared to current incremental learning methods, it is still has the limitation: e.g., Differences from the real scene, we don't utilize the Large Language Models (LLMs) in the experiments. We have only preliminarily explored the effectiveness of the method, and we will use more models and language directions to verify the proposed method in the future work. To restore the real scenes, we will refer to the original model by the Large Language Models.

Acknowledgements

The research work described in this paper has been supported by the National Nature Science Foundation of China (No. 61876198, 61976015, 61976016). We would thank the anonymous reviewers for their insightful comments.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable 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 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Giuseppe Castellucci, Simone Filice, Danilo Croce, and Roberto Basili. 2021. Learning to solve NLP tasks in an incremental number of languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 837–847, Online. 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.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems*, 32.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th international conference on web search and data mining*, pages 178–186.

Xavier Garcia, Noah Constant, Ankur Parikh, and Orhan Firat. 2021. Towards continual learning for multilingual machine translation via vocabulary substitution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1184–1192, Online. Association for Computational Linguistics.

Shuhao Gu, Bojie Hu, and Yang Feng. 2022. Continual learning of neural machine translation within low forgetting risk regions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1707–1718, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kaiyu Huang, Peng Li, Junpeng Liu, Maosong Sun, and Yang Liu. 2023a. Learn and consolidate: Continual adaptation for zero-shot and multilingual neural machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13938–13951, Singapore. Association for Computational Linguistics.

Kaiyu Huang, Peng Li, Jin Ma, and Yang Liu. 2022. Entropy-based vocabulary substitution for incremental learning in multilingual neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10537–10550, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kaiyu Huang, Peng Li, Jin Ma, Ting Yao, and Yang Liu. 2023b. Knowledge transfer in incremental learning for multilingual neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15286–15304, Toronto, Canada. Association for Computational Linguistics.

Haoming Jiang, Chen Liang, Chong Wang, and Tuo Zhao. 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, Online. Association for Computational Linguistics.

Kasidis Kanwatchara, Thanapapas Horsuwan, Piyawat Lertvittayakumjorn, Boonserm Kijsirikul, and Peerapon Vateekul. 2021. Rational LAMOL: A rationale-based lifelong learning framework. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2942–2953, Online. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Surafel M. Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 54–61, Brussels. International Conference on Spoken Language Translation.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2021. Continual mixed-language pre-training for extremely low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2706–2718, Online. Association for Computational Linguistics.

Jinliang Lu and Jiajun Zhang. 2023. Towards unified multi-domain machine translation with mixture of domain experts. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3488–3498.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul Crook, Bing Liu, Zhou Yu, Eunjoon Cho, Pascale Fung, and Zhiguang Wang. 2021. Continual learning in task-oriented dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7452–7467, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhibo Man, Zengcheng Huang, Yujie Zhang, Yu Li, Yuanmeng Chen, Yufeng Chen, and Jinan Xu. 2024a. Wdsrl: Multi-domain neural machine translation with word-level domain-sensitive representation learning.

IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32:577–590.

Zhibo Man, Yujie Zhang, Yu Li, Yuanmeng Chen, Yufeng Chen, and Jinan Xu. 2024b. An ensemble strategy with gradient conflict for multi-domain neural machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(2):1–22.

Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. 2017. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4467–4477.

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.

Wei Peng, Chongxuan Huang, Tianhao Li, Yun Chen, and Qun Liu. 2020. Dictionary-based data augmentation for cross-domain neural machine translation. *arXiv preprint arXiv:2004.02577*.

Chenze Shao and Yang Feng. 2022. Overcoming catastrophic forgetting beyond continual learning: Balanced training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2023–2036, Dublin, Ireland. Association for Computational Linguistics.

Sainbayar Sukhbaatar, Edouard Grave, Guillaume Lample, Herve Jegou, and Armand Joulin. 2019. Augmenting self-attention with persistent memory. *arXiv preprint arXiv:1907.01470*.

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2019. Lamol: Language modeling for lifelong language learning. *arXiv preprint arXiv:1909.03329*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. 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 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, Yi Lu, Shuo Li, Yiming Wang, and Longyue Wang. 2014. UM-corpus: A large English-Chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1837–1842, Reykjavik, Iceland. European Language Resources Association (ELRA).

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C.H. Hoi. 2022. Plug-and-play VQA: Zero-shot VQA by conjoining large pretrained models with zero training. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 951–967, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yong Wang, Longyue Wang, Shuming Shi, Victor OK Li, and Zhaopeng Tu. 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.

Junhong Wu, Yuchen Liu, and Chengqing Zong. 2024. F-MALLOC: Feed-forward memory allocation for continual learning in neural machine translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7180–7192, Mexico City, Mexico. Association for Computational Linguistics.

Jiali Zeng, Jinsong Su, Huating Wen, Yang Liu, Jun Xie, Yongjing Yin, and Jianqiang Zhao. 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, Brussels, Belgium. Association for Computational Linguistics.

Fan Zhang, Mei Tu, Sangha Kim, Song Liu, and Jinyao Yan. 2023. Label-free multi-domain machine translation with stage-wise training. *arXiv preprint arXiv:2305.03949*.

Wenhao Zhu, Shujian Huang, Tong Pu, Pingxuan Huang, Xu Zhang, Jian Yu, Wei Chen, Yanfeng Wang, and Jiajun Chen. 2022. FGraDA: A dataset and benchmark for fine-grained domain adaptation in machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6719–6727, Marseille, France. European Language Resources Association.

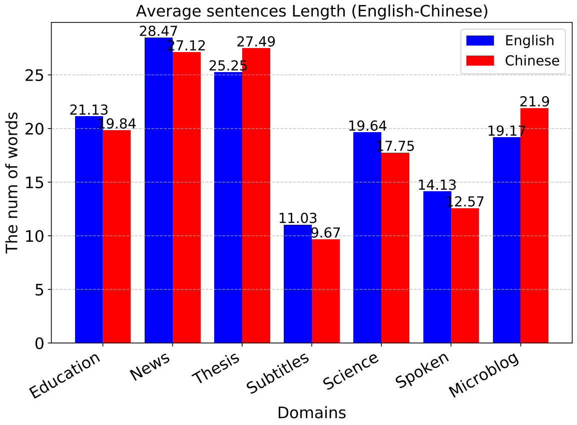

Figure 5: Average sentences Length of English-Chinese translation task

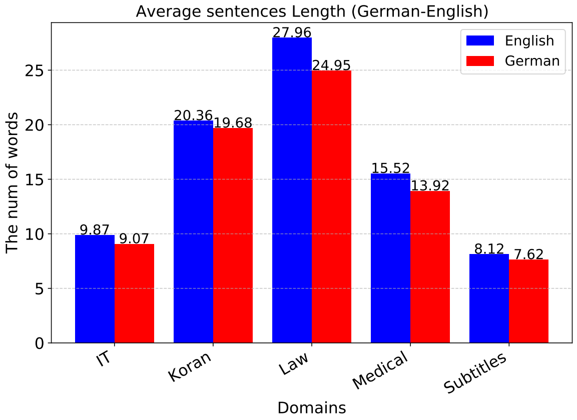

Figure 6: Average sentences Length of German-English translation task

A Detailed Domain Information

Regarding the division between the original domain and new domains, we primarily consider segmenting based on the size of the data in each domain. For the English-to-Chinese translation task, datasets such as Education, News, and Thesis with relatively large data sizes are categorized as the original domain, while datasets like Science, Subtitles, and Spoken with smaller data sizes are treated as incremental domains. Additionally, to explore the performance of our method under extremely low-resource scene, we utilize Microblog data. Similarly, for the English-to-German translation task, we adopt a similar approach for division: original domain including Subtitles, Law, Medical, and incremental domains comprising IT and Koran. Furthermore, to better illustrate the dataset characteristics, we computed the average sentence lengths across different translation tasks, as depicted in Figures 5 and 6.