Domain-Aware Adaptation for Unsupervised Machine Translation

Youyuan Lin[1] Rui Wang[2] Chenhui Chu[1]
[1]Kyoto University
[2]Shanghai Jiao Tong University
youyuan@nlp.ist.i.kyoto-u.ac.jp, wangrui12@sjtu.edu.cn, chu@i.kyoto-u.ac.jp

Abstract

Adapting Unsupervised Neural Machine Translation (UNMT) for domain-specific tasks often encounters Domain Mismatch (DM), where one language lacks sufficient in-domain monolingual data. We observe that while in-domain monolingual corpora enhance translation quality for the language they belong to, this improvement does not extend to the paired language. To address DM, we propose Domain-Aware Adaptation (DAA). DAA selects in-domain texts according to assigns higher weights to in-domain texts from open-domain corpora. Experimental results on Japanese-English translation tasks across the IT, Koran, Medical, and TED2020 domains demonstrate that DAA successfully mitigates the quality disparities in translation caused by DM, enhancing overall domain-specific translation performance.

1 Introduction

A widely held belief in Neural Machine Translation (NMT) is that increasing training data enhances model robustness and accuracy. Traditional NMT relies on parallel text pairs, which are often costly and scarce, particularly for many language pairs and specific domains [1]. This scarcity has driven interest in Unsupervised Neural Machine Translation (UNMT), which uses abundant monolingual corpora [2, 3]. With sufficient monolingual data, UNMT can approach the performance of its supervised counterparts.

However, the effectiveness of UNMT depends on the availability of in-domain monolingual data. Domain Mismatch (DM) occurs when in-domain corpora are available in one language of a pair but not the other, a common issue in low-resource settings. Shen et al. [4] highlighted that DM significantly affects methods such as back-translation [5] and self-training [6], reducing their effectiveness in low-resource scenarios.

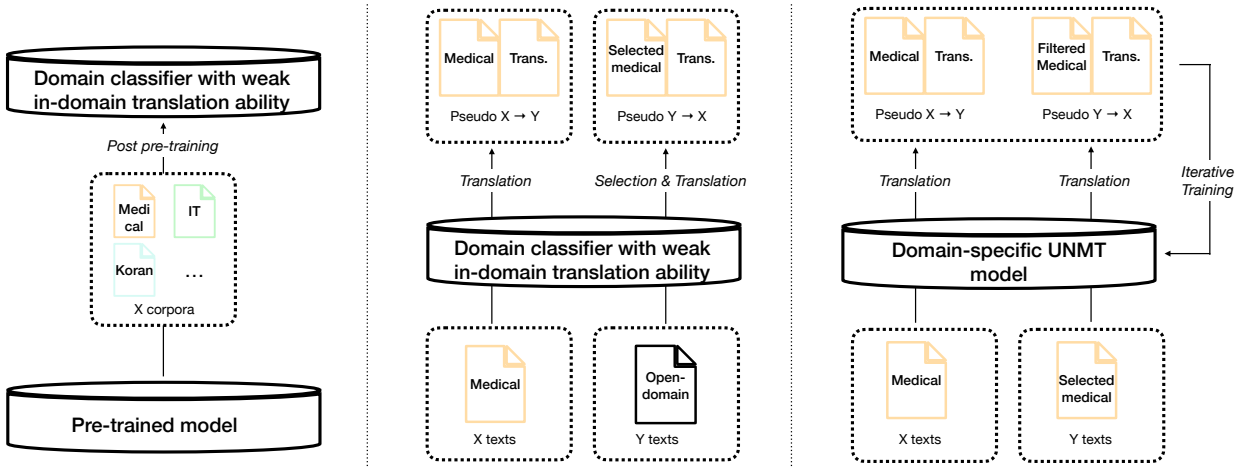
Existing approaches to mitigate DM include data selection techniques that prioritize in-domain data from larger monolingual corpora [7, 8, 9], and tagging methods that incorporate domain information through special tokens [10, 11]. Furthermore, leveraging pre-trained language models for domain-specific translation has shown promise, although often at the cost of increased computational resources and potential over-specialization [12, 13, 14, 15].

In this study, we extend the investigation of the impact of DM on UNMT and introduce Domain-Aware Adaptation (DAA) as a novel solution. Our approach utilizes multi-domain corpora in a high-resource language to train a domain classifier, which is then transferred to a low-resource language. This classifier tags open-domain texts, enabling the selection of in-domain data for training a domain-specific UNMT model.

We evaluated DAA in Japanese-English in four domains (Information Technology, Koran, Medical, TED talk) and three open-domain corpora (WMT16, OpenSubtitles, WIKIMatrix). Our results demonstrate that DAA improves translation quality in the low-resource direction and mitigates the DM problem by selectively incorporating relevant in-domain data.

In summary, the contributions of this work are as follows:

- Identification of DM as a critical issue that causes bidirectional disparities in domain-specific translation quality within UNMT models.
- Introduction of DAA, a method that integrates domain classifiers into the UNMT framework to alleviate the effects of DM.



(i) Post pre-train the model by high-resource language X to obtain domain classification ability.

(ii) Select the in-domain texts for language Y from the open-domain texts. And generate translations to combine them into a pseudo corpus.

(iii) Iteratively train the model to domain-specific UNMT model with tagged texts.

Figure 1 An overview of Domain-Aware Adaptation. The low-resource language Y lacks domain-specific texts. We use multi-domain corpora in a high-resource language X to train the model to a domain classifier. We demonstrate that the pretrained model can transfer the domain classification ability to another language. Leveraging this classification ability, we perform domain classification on the open-domain texts of Y and select out the in-domain texts to train the domain-specific UNMT model.

2 Method

The overview of our method is shown in Figure 1.

For training domain-specific UNMT models, a common approach is to utilize Iterative Back-translation (Iterative-BT) with sampled texts from the target domain, as the following equation shows:

$$L_{\mathrm{BT}}(\theta_{i+1}) = -\mathbb{E}_{x\sim P(x)}[\mathbb{E}_{y\sim P(y|x;\theta_i)} \log P(x|y;\theta_{i+1})]. \tag{1}$$

However, this method faces limitations when dealing with low-resource domain adaptation tasks due to the scarcity of in-domain texts.

To address this issue, we propose the DAA method. We aim to select in-domain texts from an open-domain corpus. Hence, we considering the domain probability in Equation (1) as follows:

$$\hat{L}_{\mathrm{BT}}(\theta_{i+1};\mathfrak{X}) = -\mathbb{E}_{x\sim P(x)} \sum_{d\in D} P(d|x)[\mathbb{E}_{y\sim P(y|x,d;\theta_i)} \log P(x,d|y;\theta_{i+1})]. \tag{2}$$

Based on Equation (2), the DAA method can be divided into three steps:

1. **Domain Classification**: Classify x from a open-domain corpus $\mathfrak{X}$ to determine its domain membership. Specifically, x is associated with each domain d in the set of domains D with probability $P(d|x)$. Following the idea of Britz et al. [11], we use the first token output from the decoder to denote the domain of x.
2. **Pseudo Pair Generation**: Generate a pseudo translation pair using the conditional probability $P(y|x,d;\theta_i)$. In this process, we assume that x belongs to the domain d and convey the domain information by entering a domain tag into the decoder. This results in the generation of a domain-specific translation y, which forms the pair (x,y).
3. **Model Update**: Perform a gradient descent on the pseudo-pair, weighted by the domain probability $P(d|x)$.

By following these steps, we perform data selection on arbitrary corpora by assigning higher weights to texts that are more likely to belong to the specific domain based on their probabilities. Some texts may exhibit a low probability $P(d|x)$ of being classified as belonging to the desired domain, i.e., being out of the domain. In practice, to improve computational efficiency, we exclude these out-of-domain texts by applying a probability threshold p.

Establishing a domain classifier during the classification step typically requires the availability of an in-domain

corpus. However, we lack in-domain corpus due to DM. To overcome this challenge, we assume that most domains exhibit cross-linguistic similarities; that is, texts within the same domain share similar patterns regardless of the language. Using this assumption, we train the domain classifier exclusively using a high-resource language such as English to discriminate domains across different languages. This enables us to perform domain-aware adaptation even in the absence of in-domain texts for low-resource languages.

3 Experimental Settings

	IT	Koran	Medical	TED2020
En	669K	451K	144K	1,000K
Ja	478K	6K	0K	361K

Table 1 Data statistics for domain-specific monolingual texts.

Datasets We conducted experiments on Japanese-English translation tasks. We selected four domains: (i) IT (GNOME and Ubuntu), (ii) Koran (Tanzil), (iii) Medical (EJMMT) and (iv) TED2020. Note that there are no monolingual texts. We randomly selected 2K parallel texts from each domain for the test set and the validation set, respectively. We used WIKIMatrix as the open-domain monolingual Japanese corpus. We randomly selected 2.8M texts from the WIKIMatrix. All corpora were collected from OPUS [16] except the Medical corpus [17].

Models We evaluated UNMT models fine-tuned on models pretrained with MASS [18]. MASS models are trained for language comprehension and language generation capabilities, which facilitates our use of the model for domain classification and translation. We adopted the pretrained model released by Mao et al. [19].

Hyperparameters We trained and tuned the model for 40 epochs and selected the best model by perplexity on the validation set. The domain threshold was set to $p = 0.5$.

Baselines To simulate the DM, for each of the four domains, the models were trained without a Japanese in-domain corpus, following the paradigm of iterative back-translation, serving as the baseline. For example, in the adaptation task for the IT domain, the Japanese corpora were combined in the Koran, Medical, and TED2020 domains, excluding the IT domain, while all English corpora were used as training data.

For comparison, we used an alternative data selection method based on K-Nearest Neighbors (kNN). We computed sentence embeddings by summing each token's representation in the encoder's last layer. We then determined each domain's representation center by averaging its sentence embeddings. Subsequently, for each missing domain, we selected the k texts closest to the center of the representation of the corresponding domain. Here, k represents the number of missing texts in that domain's corpus. For example, in the En-Ja IT task, we calculated the center of the sentence embeddings using the English IT corpus. Then, we selected k texts from the WikiMatrix corpus closest to the calculated center, forming the pseudo-Japanese IT domain corpus.

In addition, we experimented with domain-specific translations using LLaMA2-7B [20], TowerInstruct-7B [21], and ALMA-7B [22] as baselines to represent the translation capabilities of Large Language Models (LLMs). We used a zero-shot setting and prompt model by "Translating the following text from X to Y.\n$X$:{source text}\n$Y$:", in which X and Y are two languages.

We report the BLEU score [23] as the evaluation metric. The sentence is segmented by Mecab.[1)]

4 Results

4.1 Classification

	IT	Koran	Medical	TED2020	Avg.
En	99.00%	98.95%	97.64%	97.88%	98.37%
Ja	88.41%	81.56%	87.91%	75.62%	83.38%

Table 2 Recall of the domain classifier.

Table 2 shows the results of the recall in each domain. Note that only the English corpus was used to train the domain classifier. In the case of the English corpus, the results demonstrate exceptional recall in accurately classifying texts into their respective domains. For Japanese, despite being slightly dropped compared to English, the domain classifier still outperforms the random level. The results indicate that pretrained models can effectively learn domain patterns through the utilization of a single high-resource language.

Figure 2 shows the visualization results according to t-SNE [24]. For each text, we added all the word representations output by the encoder as its sentence embedding. Texts with the same domain tags are clustered close to each other. This implies that texts in the same domain

1) https://github.com/taku910/mecab

En ⇔ Ja	IT		Koran		Medical		TED2020		Avg.	
	⇐	⇒	⇐	⇒	⇐	⇒	⇐	⇒	⇐	⇒
LLaMA2-7B	11.96	4.08	5.11	3.84	13.97	5.29	16.43	10.78	11.87	6.00
TowerInstruct-7B	21.61	2.56	10.63	1.42	28.15	3.52	27.74	5.25	22.04	3.19
ALMA-7B	26.23	9.51	11.14	10.25	32.64	16.92	31.58	20.62	25.40	14.33
MASS-200M [18]	6.63	3.37	3.06	2.27	12.50	4.68	12.30	8.70	8.62	4.75
MASS-200M w/ Iterative-BT	**19.89**	2.90	7.89	2.73	26.97	2.36	22.57	13.29	19.08	5.32
+ WikiMatrix	18.30	5.10	7.85	3.51	25.31	6.53	25.38	16.23	19.21	7.84
+ WikiMatrix, w/ kNN	18.80	5.37	8.20	5.19	25.80	8.24	24.80	18.03	19.40	9.21
+ WikiMatrix, w/ DAA	19.77	**5.74**	**8.90**	**8.12**	**27.22**	**9.93**	**26.17**	**19.18**	**20.49**	**10.74**

Table 3 Translation result.

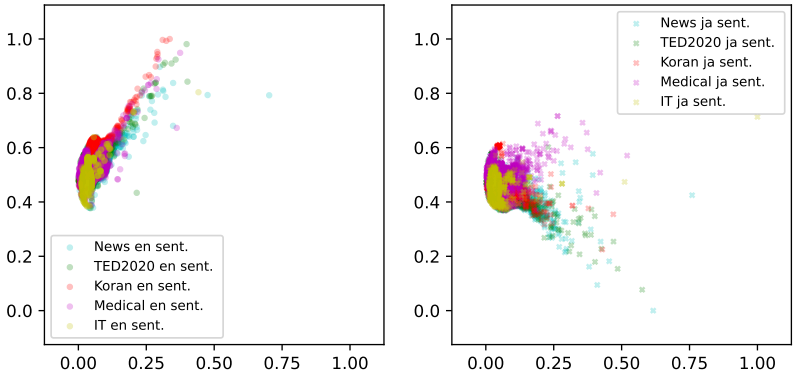


Figure 2 Left: English, Right: Japanese. The t-SNE visualization displays sentence embeddings with distinct domain separation indicated by color. All representations are generated by pretrained models without any fine-tuning. The distribution of patterns in the results is roughly horizontally symmetric, demonstrating the isomorphism of the domain distributions.

have more similar patterns and that there are texts located at "domain boundaries" that exhibit characteristics that fit the patterns of multiple domains. Intuitively, these texts can be used as an in-domain corpus. Previous works have demonstrated the isomorphism of word embeddings across languages [25, 26]. By categorizing texts into domains, we find that sentence embeddings from different domains are also approximately isomorphic, which may be a result of word embedding isomorphism. Hence, it is feasible to use text in the domain of one language to train a domain classifier that classifies the corpus across languages.

4.2 Translation

The results are shown in Table 3. We observed that across all domains, DM contributes to the disparity in translation quality. Taking advantage of the presence of domain-specific training texts, the models demonstrate superior translation performance when translated into English. In contrast, translations into Japanese exhibit subpar quality as a result of the absence of corresponding in-domain training texts. The mere incorporation of WikiMatrix does not consistently improve translation quality within specific domains. DAA outperforms the baseline, suggesting that the domain classifier in DAA more effectively filters out in-domain texts. Compared to the 7B-size LLMs, our model outperformed LLaMA2-7B, while still having a gap of about 4 points compared to ALMA-7B. Note that TowerInstruct-7B sometimes cannot follow the instruction to generate Japanese translations and hence obtained low En-Ja scores.

5 Summary

The DM poses a challenge in UNMT. To address this, we introduced DAA, which enhances domain classification by using multi-domain corpora from high-resource languages. DAA employs domain tagging and weighting to effectively select in-domain texts from open-domain corpora. Our experiments in various domains demonstrate the efficacy of DAA in mitigating the adverse effects of DM.

However, the effectiveness of DAA diminishes when domain-specific monolingual data are limited. It also remains untested in multilingual settings and on models exceeding 200 million parameters. Future work should investigate these scenarios to enhance scalability and robustness.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP23K28144.

References

[1] Emmanouil Stergiadis, Satendra Kumar, Fedor Kovalev, and Pavel Levin. Multi-domain adaptation in neural machine translation through multidimensional tagging. **arXiv preprint arXiv:2102.10160**, 2021.

[2] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. **arXiv preprint arXiv:1804.07755**, 2018.

[3] Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. Unsupervised neural machine translation with weight sharing. **arXiv preprint arXiv:1804.09057**, 2018.

[4] Jiajun Shen, Peng-Jen Chen, Matthew Le, Junxian He, Jiatao Gu, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. The source-target domain mismatch problem in machine translation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 1519–1533, Online, April 2021. Association for Computational Linguistics.

[5] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. **arXiv preprint arXiv:1511.06709**, 2015.

[6] Zhiwei He, Xing Wang, Rui Wang, Shuming Shi, and Zhaopeng Tu. Bridging the data gap between training and inference for unsupervised neural machine translation. **arXiv preprint arXiv:2203.08394**, 2022.

[7] Roee Aharoni and Yoav Goldberg. Unsupervised domain clusters in pretrained language models. **arXiv preprint arXiv:2004.02105**, 2020.

[8] Javad Pourmostafa Roshan Sharami, Dimitar Shterionov, and Pieter Spronck. Selecting parallel in-domain sentences for neural machine translation using monolingual texts. **arXiv preprint arXiv:2112.06096**, 2021.

[9] Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. Dynamic data selection and weighting for iterative back-translation. **arXiv preprint arXiv:2004.03672**, 2020.

[10] Catherine Kobus, Josep Crego, and Jean Senellart. Domain control for neural machine translation. **arXiv preprint arXiv:1612.06140**, 2016.

[11] Denny Britz, Quoc Le, and Reid Pryzant. Effective domain mixing for neural machine translation. In **Proceedings of the Second Conference on Machine Translation**, pp. 118–126, 2017.

[12] Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Towards making the most of chatgpt for machine translation. **arXiv preprint arXiv:2303.13780**, 2023.

[13] Yongyu Mu, Abudurexiti Reheman, Zhiquan Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. Augmenting large language model translators via translation memories. **arXiv preprint arXiv:2305.17367**, 2023.

[14] Johannes Eschbach-Dymanus, Frank Essenberger, Bianka Buschbeck, and Miriam Exel. Exploring the effectiveness of llm domain adaptation for business it machine translation. In **Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)**, pp. 610–622, 2024.

[15] Jiawei Zheng, Hanghai Hong, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. Fine-tuning large language models for domain-specific machine translation. **arXiv preprint arXiv:2402.15061**, 2024.

[16] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In **Lrec**, Vol. 2012, pp. 2214–2218. Citeseer, 2012.

[17] Takeshi Hayakawa and Yuki Arase. Fine-grained error analysis on english-to-japanese machine translation in the medical domain. In **Proceedings of the 22nd Annual Conference of the European Association for Machine Translation**, pp. 155–164, 2020.

[18] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. **arXiv preprint arXiv:1905.02450**, 2019.

[19] Zhuoyuan Mao, Chenhui Chu, and Sadao Kurohashi. Linguistically driven multi-task pre-training for low-resource neural machine translation. **Transactions on Asian and Low-Resource Language Information Processing**, Vol. 21, No. 4, pp. 1–29, 2022.

[20] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[21] Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. Tower: An open multilingual large language model for translation-related tasks. **arXiv preprint arXiv:2402.17733**, 2024.

[22] Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. **arXiv preprint arXiv:2309.11674**, 2023.

[23] Estelle Bettelli, Yijun Carrier, Wenda Gao, Thomas Korn, Terry B Strom, Mohamed Oukka, Howard L Weiner, and Vijay K Kuchroo. Reciprocal developmental pathways for the generation of pathogenic effector th17 and regulatory t cells. **Nature**, Vol. 441, No. 7090, pp. 235–238, 2006.

[24] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. **Journal of machine learning research**, Vol. 9, No. 11, 2008.

[25] Xilun Chen and Claire Cardie. Unsupervised multilingual word embeddings. **arXiv preprint arXiv:1808.08933**, 2018.

[26] Takashi Wada, Tomoharu Iwata, and Yuji Matsumoto. Unsupervised multilingual word embedding with limited resources using neural language models. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3113–3124, 2019.