Simultaneous Domain Adaptation of Tokenization and Machine Translation

Taisei Enomoto[1], Tosho Hirasawa[1], Hwichan Kim[1], Teruaki Oka[2], Mamoru Komachi[1,2]
[1]Tokyo Metropolitan University
[2]Hitotsubashi University
{enomoto-taisei, hirasawa-tosho, kim-hwichan}@ed.tmu.ac.jp
{teruaki.oka, mamoru.komachi}@r.hit-u.ac.jp

Abstract

Domain adaptation through fine-tuning is a well-established strategy to tailor a neural network model trained on a general-domain for a specific target-domain. During the fine-tuning process, the parameters of the model are updated while keeping the general-domain tokenizer unchanged. However, this tokenizer is trained on general-domain data and hence, not entirely optimal for the target-domain. Previous research has shown that simultaneously updating a tokenizer during training a model can enhance the performance of tasks such as classification and machine translation. Building on this concept, our objective is to enhance translation performance in the target-domain by jointly adapting the tokenizer during both pre-training and fine-tuning. Our results demonstrate that domain adaptation of the tokenizer enables the acquisition of a suitable tokenizer for target-domain translation, resulting in improved translation performance for domain-specific inputs.

1 Introduction

The neural machine translation (NMT) model achieves state-of-the-art translation performance in scenarios where abundant resources are available (Bojar et al., 2017; Nakazawa et al., 2017). However, the NMT model has limitations when it comes to accurately translating sentences from domains that differ significantly from those of the training data (Koehn and Knowles, 2017). Furthermore, high-quality training of an NMT model requires a large amount of parallel data, which is only available for a few specific domains. To overcome this problem, domain adaptation—the process of adapting a model to a target-domain—is employed. Luong and Manning (2015) fine-tuned a NMT model with a small amount of target-domain data and demonstrated that this approach improves translation performance for target-domain inputs.

src	**The electrophotographic process** is widely applied ...
ULM	**_The/_electro/pho/t/ographic/_process/**_is/_widely/applied/...
Tgt	**_The/_electro/pho/t/ographic/_process/**_is/_widely/applied/...
Gen→Tgt	**_The/_electro/photo/graph/ic/_process/**_is/_widely/applied/...

(a) Segmentations by tokenizers trained using each method.

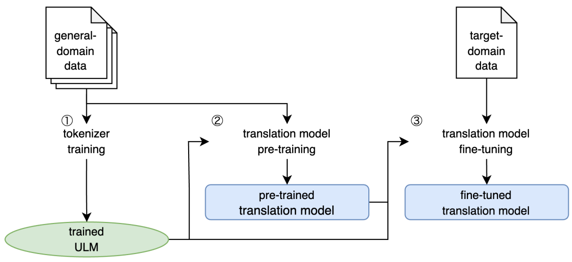
	ref	電子写真プロセスは， ... 広く応用されている。
✗	ULM	.../広く/応用/さ/れ/て/いる/。
✗	Tgt	電気/泳動/法/は/，/.../広く/応用/さ/れ/て/いる/。
✓	Gen→Tgt	電子/写真/プロセス/は/，/.../広く/応用/さ/れ/て/いる/。

(b) Translations from the model trained using each method. The input for each model is the output from (a).

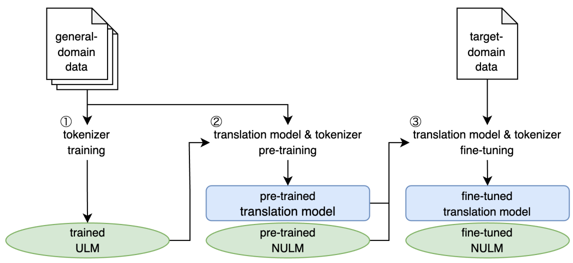
Table 1: Results of tokenization and En-Ja translation. ULM: general-domain tokenization. Tgt: trained only on target-domain data. Gen→Tgt (proposed): simultaneous domain adaptation of tokenization and translation. " ..." is an omission symbol, and "_" is a space symbol.

In domain adaptation, the tokenizer is typically left unchanged during fine-tuning (Figure 1a). However, previous studies have demonstrated that appropriate tokenization varies depending on several factors, and adapting a tokenizer could enhance task performance (Xu et al., 2008; Chang et al., 2008; Nguyen et al., 2010; Hiraoka et al., 2020). We hypothesize that using a suitable tokenizer for target-domain translation has the potential to enhance translation performance in that domain.

In this study, our objective is to improve translation performance for a target-domain through domain adaptation of a tokenizer and a translation model. Initially, we pre-train the tokenizer with a large amount of general-domain data using joint optimization of a tokenizer and a translation model (OpTok4AT) (Hiraoka et al., 2021). Then, we fine-tune the tokenizer and the translation model with a small amount of target-domain data. We evaluate the effectiveness of our approach through experiments on English–Japanese (En-Ja) and English–German (En-De) domain-specific translation tasks.



(a) Conventional domain adaptation of only a translation model in machine translation tasks.



(b) Domain adaptation of a translation model and a tokenizer in machine translation tasks.

Figure 1: Outline of (a) conventional domain adaptation and (b) proposed method.

Our results demonstrate that domain adaptation of a tokenizer achieves suitable tokenization for translation in a target-domain, such as the medical field, resulting in improved translation performance in that domain. Table 1 illustrates an example of how the proposed method enhances tokenization and translation. The proposed method has been shown to prevent mistranslation and untranslation.

We summarize our main contributions as follows:

- We propose simultaneous domain adaptation of a tokenizer and a translation model to enhance translation performance for a target-domain.
- We demonstrate the effectiveness of our method in En-Ja and En-De translation tasks. Our experiments demonstrate that an additional pre-training of a few epochs is sufficient for pre-training the generic tokenizer.
- Our analysis reveals that the adapted tokenizer splits target-domain-specific words into subwords that are semantically appropriate and suitable for translation.

2 Related Work

2.1 Optimization of Subword Tokenization

In many NLP tasks, subwords, which are units smaller than words, have proven to be effective in handling unknown words and rare words (Sennrich et al., 2016; Song et al., 2021). Kudo (2018) proposed a subword tokenization method based on a unigram language model (ULM). In this method, a sentence s is transformed into a series of subwords $s' = w_1, ..., w_I$ such that the likelihood (the product of unigram probabilities from ULM) is maximized. This method trains a tokenizer based on the given training data using EM algorithm, and the tokenizer does not change while the model is trained.

Various methods have been proposed to automatically optimize a tokenizer based on a task (Salesky et al., 2020; He et al., 2020; Hiraoka et al., 2020). Recently, Hiraoka et al. (2021) proposed OpTok4AT. This method comprises a tokenizer and a model and trains them simultaneously in an end-to-end manner. It uses a neural unigram language model (NULM) as a tokenizer; NULM is a ULM comprising a neural network. They reported improved performance in several tasks, including a machine translation task. However, they did not pre-train tokenizers and translation models on general-domain data, but trained them only on target-domain data. It is unclear whether domain adaptation of a tokenizer improves the performance of a target-domain translation model; hence, we verify it.

2.2 Domain Adaptation of NMT

Previous studies have proposed several domain adaptation methods for NMT. For example, Freitag and Al-Onaizan (2016) trained a model on large data and then fine-tuned it on small target-domain data. They reported improvements in the translation performance on the target-domain. Chu et al. (2017) proposed mixed fine-tuning, which combines general-domain data and target-domain data to fine-tune a model on these data. However, in both studies, only the model parameters are updated through domain adaptation, and the tokenizer is fixed. As tokenization affects translation performance, we propose simultaneous domain adaptation of a tokenizer and a translation model to improve the translation performance on the target-domain.

3 Simultaneous Domain Adaptation of a tokenizer and a translation model

This section provides an overview of the NULM used as the tokenizer and the procedure for domain adaptation of the tokenizer using OpTok4AT. Figure 1b illustrates the simultaneous domain adaptation process for the tokenizer and the translation model.

First, the NULM vocabulary V is initialized using ULM (Kudo, 2018). In NULM, the unigram probability $p(w)$ is computed for each subword w in the vocabulary V using scalar values d_w based on the word embedding $\boldsymbol{v}_w$ and the multilayer perceptron $\mathrm{MLP}(\cdot)$ as follows:

$$d_w = \mathrm{MLP}(\boldsymbol{v}_w) \tag{1}$$

$$p(w) = \frac{\exp(d_w)}{\sum_{\hat{w} \in V} \exp(d_{\hat{w}})} \tag{2}$$

NULM updates its parameters based on losses in tasks such as machine translation.

Second, we employ OpTok4AT to train the tokenizers. In current NLP research, it is common to use publicly available pre-trained models, whose tokenizers are fixed. Following this trend, we investigate two possibilities: (1) training a tokenizer in both pre-training and fine-tuning steps and (2) using a general-domain tokenizer with additional pre-training and fine-tuning. Our proposed methods are as follows:

Gen→Tgt In this setting, we train a tokenizer during pre-training and fine-tuning. The process consists of three steps. First, we train a ULM (Kudo, 2018) on general-domain data. We use the ULM to initialize the vocabulary of NULM. Second, we pre-train a NULM and a translation model on general-domain data. Third, we fine-tune the tokenizer and the translation model using target-domain data.

Gen$^{n\text{ep-Tok}}$→Tgt / Gen$^{n\text{ep}}$→Tgt We perform additional pre-training on general-domain data for n epochs after pre-training a translation model. This setting follows the same processes as Gen→Tgt, except for the second step. In the second step, we pre-train a translation model solely on general-domain data while keeping the tokenizer fixed. Then, we additionally pre-train either the tokenizer alone (Gen$^{n\text{ep-Tok}}$→Tgt) or both the tokenizer and the translation model (Gen$^{n\text{ep}}$→Tgt) using general-domain data.

4 Experiment

4.1 Settings

Datasets As general-domain data, we used JParaCrawl v3.0 (Morishita et al., 2020, 2022) for En-Ja translation and ParaCrawl v9 (Esplà et al., 2019) for En-De translation. We extracted eight million sentence pairs per language pair from the entire data as training data. As target-domain data, we used IWSLT2017 (Cettolo et al., 2017) and ASPEC (Nakazawa et al., 2016) for En-Ja translation and IWSLT2017 and EMEA (Tiedemann, 2012) for En-De translation. IWSLT is created from TED talks, ASPEC from scientific and technical papers, and EMEA from medical documents. We randomly down-sampled the training data of ASPEC and EMEA to match the number of sentences in the IWSLT training data, approximately two hundred thousand. Following the methodology described in the previous study (Hiraoka et al., 2021), we trained the ULM using SentencePiece (Kudo and Richardson, 2018) after applying MeCab (Kudo, 2006) (IPA dictionary) for the Japanese side and Moses tokenizer (Koehn et al., 2007) for the English and German sides. All the tokenizers had a vocabulary size of 32,000.

Training settings We used Transformer (Vaswani et al., 2017) (base) as the translation model [1]. To validate the effectiveness of simultaneous domain adaptation of the tokenizer and the translation model, we compared the proposed method with three baselines: ULM, Gen, and Tgt. These baselines correspond to settings in which the tokenizer is fixed during both pre-training and fine-tuning, fine-tuning only, and pre-training only, respectively. Table 2 summarizes the settings for each method [2]. We trained tokenizers for the source and target languages simultaneously. Subword regularization (Kudo, 2018; Provilkov et al., 2020) was applied in all settings.

Evaluation settings We evaluated the translation performance of each method using automatic and human evaluations. For automatic metrics, we used BLEU (Papineni et al., 2002) with Sacre-

[1]Our implementation is based on the existing code:https://github.com/tatHi/optok4at

[2]We present the settings without pre-training the translation model in Appendix A

Setting	Pre-train		Add Pre-train		Fine-tune		En-Ja				En-De			
							IWSLT		ASPEC		IWSLT		EMEA	
	TM	Tok	TM	Tok	TM	Tok	bleu	comet	bleu	comet	bleu	comet	bleu	comet
ULM	✓				✓		14.83	0.118	27.51	0.634	26.06	0.463	35.17	0.533
Gen	✓	✓			✓		14.94	0.119	27.24	0.634	26.37	0.466	35.19	0.532
Tgt	✓				✓	✓	14.40	0.092	27.12	0.626	26.23	0.462	34.92	0.529
Gen→Tgt	✓	✓			✓	✓	**15.16**	**0.126**	**27.68**	**0.641**	**26.58**	**0.473**	**35.52**	**0.541**
$\text{Gen}^{2\text{ep}}$→Tgt	✓		✓	✓	✓	✓	14.72	0.104	27.27	0.630	26.16	0.463	35.12	0.524
$\text{Gen}^{3\text{ep}}$→Tgt	✓		✓	✓	✓	✓	14.96	0.113	27.53	0.634	26.48	0.468	35.30	0.528
$\text{Gen}^{4\text{ep}}$→Tgt	✓		✓	✓	✓	✓	14.97	0.121	**27.71**	0.638	26.46	0.470	35.26	0.526
$\text{Gen}^{5\text{ep}}$→Tgt	✓		✓	✓	✓	✓	**15.01**	**0.125**	27.64	**0.641**	**26.59**	**0.475**	**35.45**	**0.538**
$\text{Gen}^{5\text{ep-Tok}}$→Tgt	✓			✓	✓	✓	14.97	0.120	27.29	0.636	26.43	0.467	35.43	0.533

Table 2: Automatic metrics scores of baselines and our proposed method for each target-domain data. “TM” and “Tok” represent the translation model and the tokenizer, respectively. The “✓” represents training the relevant component. “Add Pre-train” means additional pre-training, as mentioned in Section 3. Note that we use general-domain data in the “Pre-train” and “Add Pre-train” processes and target-domain data in the “Fine-tune” process.

BLEU [3] (Post, 2018) and COMET [4] (Rei et al., 2020) and reported the average score over three seeds. For human evaluation, we performed a pairwise comparison of the translations of ULM and Gen→Tgt for En-Ja translation based on two attributes: adequacy and fluency. We randomly sampled 100 outputs per method in each target-domain data for human evaluation. Tie was allowed, and system identifiers were shuffled and masked during annotation. We evaluated each output by two annotators [5] and reported the average results.

4.2 Results

Automatic evaluation Table 2 shows BLEU and COMET scores for each target-domain data. The experimental results demonstrate that Gen→Tgt achieving the highest scores. These results indicate that simultaneous domain adaptation of a tokenizer and a translation model improves translation performance for target-domain inputs. Conversely, Tgt performed poorly compared with ULM except in terms of the BLEU score on IWSLT (En-De). This result suggests that it is insufficient to train a tokenizer (NULM) using only a small amount of target-domain data.

In the settings where additional pre-training is performed, the BLEU and COMET scores of $\text{Gen}^{5\text{ep-Tok}}$→Tgt are higher than those of Tgt but lower than those of Gen→Tgt. The scores of $\text{Gen}^{n\text{ep}}$→Tgt improve progressively with each epoch and are similar to those of Gen→Tgt after five epochs. These results indicate that additional pre-training of the general-domain tokenizer of a pre-trained model can improve translation performance in the target-domain. Moreover, during additional pre-training, updating both the tokenizer and the translation model further improves translation performance compared to just updating only a tokenizer. The translation performance improvement with additional pre-training of a small number of epochs can be attributed to the tokenizer being based on MLP and having a simpler structure than the translation model. These results indicate that our approach works well for converged models trained with a fixed tokenizer, such as publicly available pre-trained models.

Human evaluation Figure 2 shows the results of human evaluations for each En-Ja target-domain data. Regarding adequacy, Gen→Tgt outputs are preferred over ULM outputs by more than ten points in both target-domain data. In terms of fluency, Gen→Tgt and ULM outputs are comparable. Moreover, we report the results of a confusion matrix and Cohen’s Kappa (Cohen, 1960) between the two annotations to measure inter-rater reliability. Figure 3 shows the confusion matrix between annotators’ evaluations for each attribute. In terms of adequacy, the evaluations of the two annotators often agree whether on ULM, Tie and Gen→Tgt, and Kappa is 0.746 on IWSLT and 0.707 on ASPEC. According to Landis and Koch (1977), we can determine that the two annotations are substantially consistent and highly reliable. Conversely, in terms of fluency, the evaluations of the two annotators often agree only on Tie and not much on the

[3] https://github.com/mjpost/sacrebleu

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

[5] They are native Japanese speakers and students pursuing a Masters in NLP.

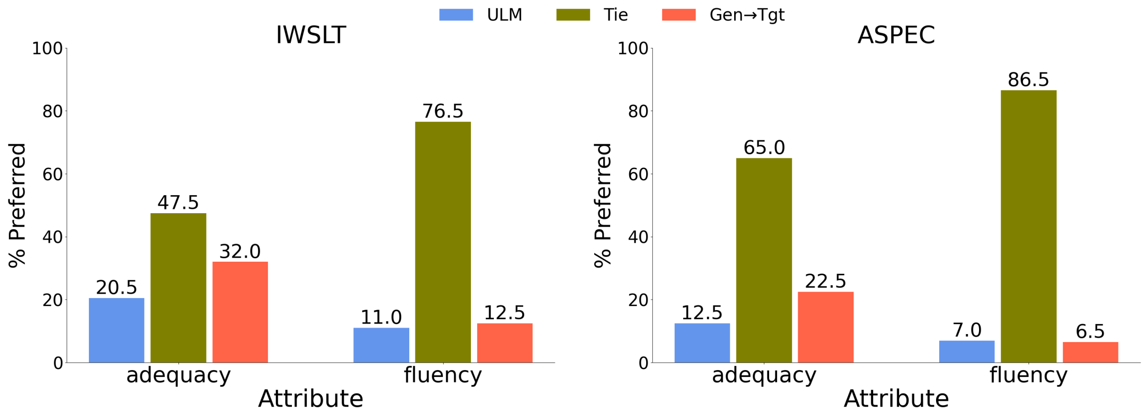


Figure 2: Head-to-head comparison of ULM and Gen→Tgt outputs for En-Ja translation in terms of adequacy and fluency.

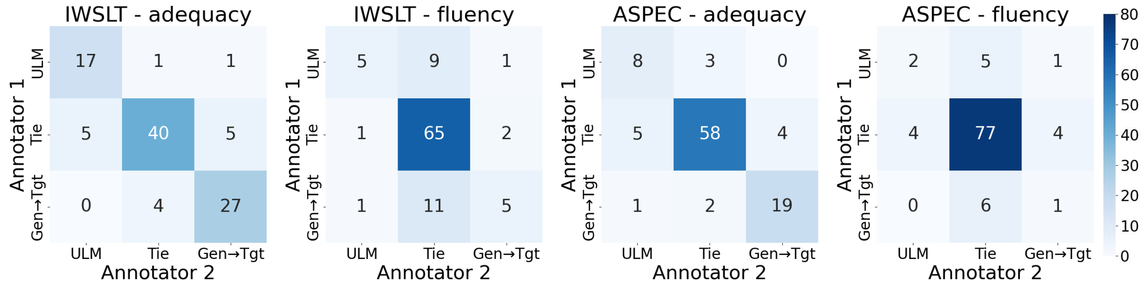


Figure 3: Confusion matrix between the annotators' evaluations for each attribute.

others, and Kappa is 0.372 on IWSLT and 0.177 on ASPEC, which are fair and slight agreement rates, respectively. This finding could be attributed to the tendency of ULM and Gen→Tgt to have lower fluency, making them equivalent, and differences in the annotator's preference might account for the disparities in evaluations. These results indicate that the proposed method improves the translation performance of target-domain data in terms of adequacy but not fluency. We suppose that the improvement in adequacy is driven by enabling a suitable tokenization for the target-domain. We analyze this in Section 5.1.

5 Discussion

5.1 Examples of tokenization and translation

In this section, we analyze how tokenizers alter segmentation through domain adaptation and how these changes subsequently lead to improving translation performance. Table 1 presents examples of tokenization and translation for three settings: ULM, Tgt, and Gen→Tgt. We focus on the string "The electrophotographic process," which is written as "電子写真プロセス" in Japanese. While ULM and Tgt tokenize "electrophotographic" as "_electro / pho / t / ographic," Gen→Tgt tokenizes it as "_electro / photo / graph / ic." Consequently, the translation of the relevant part remains untranslated in ULM [6] and is incorrectly translated as "電気泳動法," meaning "The electrophoresis method," in Tgt, whereas Gen→Tgt produces the correct translation. This result suggests that domain adaptation enables the tokenizer to tokenize in-domain words into appropriate subwords for translating target-domain data. Therefore, acquiring a suitable tokenizer for the target-domain leads to improved translation performance.

5.2 Changes in tokenizers by fine-tuning

We also analyze how the tokenizer, pre-trained on general-domain data, changes when fine-tuned on target-domain data.

Subwords with a large increase in unigram probability Our analysis indicates that fine-tuning a tokenizer on target-domain data increases the unigram probability of subwords that play an important role in the target-domain. Tables 3 and 4 show the subwords with a substantial increase in unigram probability after fine-tuning the tokenizer on

[6]Therefore, the ULM translation in Table 1 does not include a part corresponding to "電子写真プロセス."

IWSLT		ASPEC	
En	Ja	En	Ja
_verifi	TED	ic	ラーゼ (-lase)
_obsess	プリ (pre, pri)	_augment	ED
_sounds	シティ (-city, -sity)	_defect	_SYN

Table 3: Top three subwords exhibiting a significant increase in unigram probability due to fine-tuning during En-Ja translation.

IWSLT		EMEA	
En	De	En	De
_boost	_Sch	g	kin
_sup	liz	_mugg	tro
ory	rie	ara	ati

Table 4: Top three subwords exhibiting a significant increase in unigram probability due to fine-tuning during En-De translation.

target-domain data for En-Ja and En-De translation, respectively.

On the Japanese side of IWSLT, there is a notable increase in the unigram probability of "TED." [7] As IWSLT is a corpus derived from TED talk subtitles, texts containing the word "TED" frequently appear in the training data, approximately 900 times. In the ASPEC training data, adjectives with the suffix "ic," such as "magnetic," are commonly encountered, leading to an increased unigram probability of "ic" on the English side.

On the German side of EMEA, the most significant increase in unigram probability is observed for "kin." The EMEA training data include the term "pharmakokinetik," a word specific to the medical field, which occurs frequently (approximately 1,500 times). Tokenizing this word into "pharmako / kin / etik" is considered a semantically reasonable segmentation. Moreover, medical words ending in "kin," such as "Interleukin" and "Hodgkin," are frequent, indicating that "kin" is a subword that plays an important role on the German side of EMEA. On the English side, the largest increase in unigram probability is seen for "g." As EMEA pertains to the medical domain, many sentences describe the mass of drugs and other substances. Therefore, mass units such as "g," "mg," and "ng" appear frequently in EMEA.

[7] "TED" is used here instead of "_TED" because "_TED" is not registered in the vocabulary. This is due to the vocabulary being based on JParaCrawl, which has few words that begin with "TED" and many that contain or end with "TED."

	En-Ja		En-De	
	IWSLT	ASPEC	IWSLT	EMEA
source（En）	0.98	4.29	0.52	6.68
target（Ja/De）	0.11	0.18	0.35	6.13

Table 5: Percentage of sentences in which tokenization changed due to fine-tuning of tokenizers.

Percentage of sentences with changed tokenization Table 5 presents the percentage of sentences that exhibit different tokenization when comparing the pre-trained tokenizer trained on general-domain data with the fine-tuned tokenizer trained on target-domain data. Notably, in the En-Ja language pair, the difference in tokenization is more prominent in ASPEC compared to IWSLT. This result can be attributed to the fact that the domain of ASPEC is more dissimilar to the domain of JParaCrawl than the domain of IWSLT (Appendix C), resulting in greater changes in the tokenizer after fine-tuning.

Similarly, for the En-De language pair, the percentage of sentences with altered tokenization is higher in EMEA than in IWSLT. These findings indicate that as the target-domain corpus becomes more distinct in its characteristics (further deviating from the general-domain), the tokenizer undergoes more significant changes during the fine-tuning process. Even in the case of EMEA, which exhibits the highest percentage, the change in tokenization is relatively low at 6.68 %. This result indicates that the tokenizer does not change considerably by fine-tuning and also retains knowledge learned in general-domain. Consequently, these results suggest that fine-tuning the tokenizer on target-domain data requires slight adjustments to enhance translation performance.

6 Conclusion

This study proposed simultaneous domain adaptation of a tokenizer and a translation model. The experiments demonstrated that the proposed method improved the translation performance of the model on target-domain data by training a suitable tokenizer for the target-domain. We also found that the proposed method works well for a pre-trained translation model with additional pre-training of the general-domain tokenizer.

Several studies have demonstrated that pre-trained masked language models (MLMs), such as BART (Lewis et al., 2020) and MASS (Song et al., 2019), enhance translation performance. However, we did not investigate whether our approach works well when the task of pre-training is different from that of fine-tuning. In the future, we will verify whether our approach can improve translation performance when using pre-trained MLMs.

Acknowledgements

This work was supported by TMU research fund for young scientists.

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.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuitho Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 2–14.

Pi-Chuan Chang, Michel Galley, and Christopher D Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the third workshop on statistical machine translation*, pages 224–232.

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. ArXiv:1612.06897.

Xuanli He, Gholamreza Haffari, and Mohammad Norouzi. 2020. Dynamic programming encoding for subword segmentation in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3042–3051, Online. Association for Computational Linguistics.

Tatsuya Hiraoka, Sho Takase, Kei Uchiumi, Atsushi Keyaki, and Naoaki Okazaki. 2020. Optimizing word segmentation for downstream task. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1341–1351, Online. Association for Computational Linguistics.

Tatsuya Hiraoka, Sho Takase, Kei Uchiumi, Atsushi Keyaki, and Naoaki Okazaki. 2021. Joint optimization of tokenization and downstream model. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 244–255, Online. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Taku Kudo. 2006. MeCab: Yet another part-of-speech and morphological analyzer. `http://taku910.github.io/mecab/`.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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.

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. 2022. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6704–6710, Marseille, France. European Language Resources Association.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3603–3609, Marseille, France. European Language Resources Association.

Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Hideya Mino, Isao Goto, Hideto Kazawa, Yusuke Oda, Graham Neubig, and Sadao Kurohashi. 2017. Overview of the 4th workshop on Asian translation. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 1–54, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

ThuyLinh Nguyen, Stephan Vogel, and Noah A. Smith. 2010. Nonparametric word segmentation for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 815–823, Beijing, China. Coling 2010 Organizing Committee.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Elizabeth Salesky, Andrew Runge, Alex Coda, Jan Niehues, and Graham Neubig. 2020. Optimizing Segmentation Granularity for Neural Machine Translation. *Machine Translation*, 34:41–59.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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, Berlin, Germany. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast WordPiece tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jia Xu, Jianfeng Gao, Kristina Toutanova, and Hermann Ney. 2008. Bayesian semi-supervised Chinese word segmentation for statistical machine translation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1017–1024, Manchester, UK. Coling 2008 Organizing Committee.

A Without Pre-training

Table 6 presents the BLEU scores for each target-domain data in scenarios where we do not pre-train the tokenizer and the translation model (i.e., we train them only on target-domain data). The experimental result demonstrate that the performance in the setting in which the tokenizer is trained only on target-domain data tends to be lower compared to the setting in which the tokenizer is not trained. This observation aligns with the trend described earlier.

		En-Ja		En-De	
TM	Tok	IWSLT	ASPEC	IWSLT	EMEA
✓		12.06	25.99	22.78	28.42
✓	✓	11.87	25.65	22.92	28.22

Table 6: BLEU scores for each target-domain data in each setting, without pre-training the tokenizer and the translation model.

B Analysis of Human Evaluation

In this section, we present examples of inter-rater agreement in human evaluation. Table 7 presents examples in which Gen→Tgt is better than ULM in human evaluations. Gen→Tgt is evaluated superior to ULM based on translation of "spinal bifida." While ULM tokenizes "bifida" as "_b/ifi/da," Gen→Tgt tokenizes it as "_bi/fi/da." Although "bifida" is a rare word, Gen→Tgt can translate it correctly by splitting it into "_bi/fi/da," and learning the meaning of "_bi" etc. from other words. On the other hand, table 8 presents examples in which ULM are better than Gen→Tgt in human evaluations. ULM is evaluated superior to Gen→Tgt based on translation of "azoospermic men." This result could be achieved because the subwords that constitute azoospermic do not often appear in the training data of target-domain data, and their meanings cannot be learned correctly.

C Domain Distance of the Corpora

This section discusses the remoteness of the corpus domain utilized in this study. In accordance with Aharoni and Goldberg (2020), we extracted the vectors of the hidden layer of the pre-trained BERT model for each source-side sentence in the corpus. These vectors were then subjected to a 2D visualization using PCA. The results of this visualization for each language pair are presented in Figures 4 and 5. Figure 4 demonstrates that there is a significant overlap between the sentences in IWSLT and JParaCrawl, whereas most of the sentences in ASPEC do not overlap with those in JParaCrawl. This observation indicates that the domain of ASPEC is more distant from JParaCrawl compared to IWSLT. Similarly, Figure 5 suggests that EMEA, as a domain, is further removed from ParaCrawl than IWSLT.

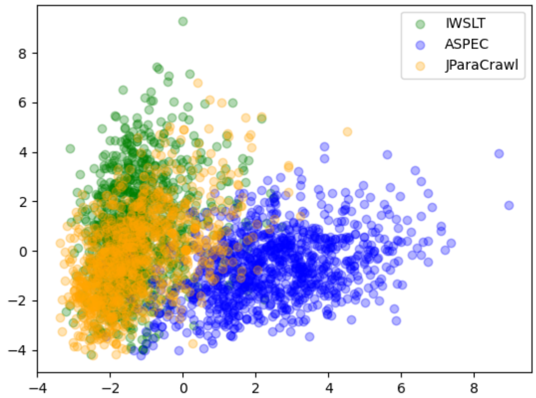


Figure 4: 2D visualization of the BERT hidden layer for the En-Ja dataset using PCA.

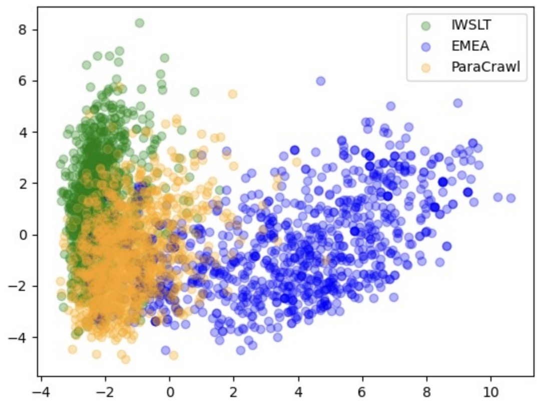


Figure 5: 2D visualization of the BERT hidden layer for the En-De dataset using PCA.

src	From clinical laboratory findings, pathohistological findings and image examination findings, linear scleroderma with **spinal bifida** was diagnosed.
ULM	_From/_clinical/_laboratory/_findings/_,/_path/oh/ist/ological/_findings/_and/_image/_examination/_findings/_,/_linear/_s/cle/rod/er/ma/ _with/**_spin/al/_b/ifi/da/**_was/_diagnos/ed/_.
Gen→Tgt	_From/_clinical/_laboratory/_findings/_,/_path/oh/ist/ological/_findings/_and/_image/_examination/_findings/_,/_linear/_s/cle/rod/er/ma/ _with/**_spin/al/_bi/fi/da/**_was/_diagnos/ed/_.

(a) Segmentations by tokenizers trained using each method.

ref	臨床検査所見，病理組織学的所見及び画像検査所見から，**二分脊椎**を合併した線状強皮症と診断した。
✗ ULM	臨床/検査/所見/,/病理/組織/学/的/所見/,/画像/検査/所見/から/,/脊髄/線/状/強/皮/症/と/診断/し/た/。
✓ Gen→Tgt	臨床/検査/所見/,/病理/組織/学/的/所見/および/画像/検査/所見/から/,/**二/分/脊椎**/を/有する/線形/強/皮/症/と/診断/し/た/。

(b) Translations from the model trained using each method. The input for each model is the output from (a).

Table 7: Examples in which Gen→Tgt translation is better than ULM translation in human evaluations. "_" is a space symbol.

src	In some **azoospermic** men, the region of a Y chromosome including a heat shock transcription factor on a Y chromosome (HSFY) is lost.
ULM	_In/_some/**_a/zo/os/per/mic**/_men/_,/_the/_region/_of/_a/_Y/_chromosome/_including/_a/_heat/_shock/_transcription/_factor/_on/_a/_Y/ _chromosome/_(/_HS/FY/_)/_is/_lost/_.
Gen→Tgt	_In/_some/**_/az/oo/s/per/mic**/_men/_,/_the/_region/_of/_a/_Y/_chromosome/_including/_a/_heat/_shock/_transcription/_factor/_on/_a/_Y/ _chromosome/_(/_HS/FY/_)/_is/_lost/_.

(a) Segmentations by tokenizers trained using each method.

ref	一部の**無精子症**の男性はＹ染色体上熱ショック転写因子（ＨＳＦＹ）を含むＹ染色体の領域を消失している。
✓ ULM	幾つ/か/の/**アゾスペルマミック**/男性/で/は/,/Y/染色/体/(/HSFY/)/上/の/熱/ショック/転写/因子/を/含む/Y/染色/体/の/領域/が/失わ /れ/て/いる/。
✗ Gen→Tgt	Y/染色/体/(/H/SFY/)/上/の/熱/ショック/転写/因子/を/含む/Y/染色/体/の/領域/が/失わ/れる/こと/が/ある/。

(b) Translations from the model trained using each method. The input for each model is the output from (a).

Table 8: Examples in which ULM translation is better than Gen→Tgt translation in human evaluations. "_" is a space symbol.