Instance-Based Domain Adaptation for Improving Terminology Translation

Prashanth Nayak prashanth.nayak@adaptcentre.ie
School of Computing, Dublin City University, Dublin, Ireland

Rejwanul Haque rejwanul.haque@setu.ie
School of Computing, South East Technological University, Carlow, Ireland

John D. Kelleher john.kelleher@mu.ie
Department of Computer Science, Maynooth University, Dublin, Ireland

Andy Way andy.way@adaptcentre.ie
School of Computing, Dublin City University, Dublin, Ireland

Abstract

Terms are essential indicators of a domain, and domain term translation is dealt with priority in any translation workflow. Translation service providers who use machine translation (MT) expect term translation to be unambiguous and consistent with the context and domain in question. Although current state-of-the-art neural MT (NMT) models are able to produce high-quality translations for many languages, they are still not at the level required when it comes to translating domain-specific terms. This study presents a terminology-aware instance-based adaptation method for improving terminology translation in NMT. We conducted our experiments for French-to-English and found that our proposed approach achieves a statistically significant improvement over the baseline NMT system in translating domain-specific terms. Specifically, the translation of multi-word terms is improved by 6.7% over a strong baseline.

1 Introduction

NMT (Vaswani et al., 2017) has been the state-of-the-art in MT research and development for some time. Fine-tuning NMT models usually requires specialised domain data for translating domain text (Luong and Manning, 2015). In recent times, Large-scale pre-trained models (LPTMs) (Devlin et al., 2019; Brown et al., 2020; Liu et al., 2020) have gained significant attention due to their remarkable performance in various Natural Language Processing (NLP) tasks. These models have proven effective in diverse applications, from information extraction to text generation. As a result, the NLP community is increasingly focused on harnessing their potential. One of the key advantages of LPTMs is that they often require smaller amounts of data for domain adaptation compared to traditional machine learning models (Devlin et al., 2019). By leveraging pre-trained knowledge, LPTMs can be fine-tuned on specific domains with relatively limited data, making them a valuable resource for addressing domain-specific challenges in NLP. However, despite significant improvements in translation quality, NMT systems still struggle with translating terminology. Even domain-adapted models are found to have difficulty with accurately translating domain-specific terminology (Sato et al., 2020).

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 222–234
September 4–8, 2023, Macau SAR, China.

This paper proposes a simple yet effective instance-based fine-tuning approach based on terminology-aware mining. We tested our approach on the French-to-English terminology translation task [1] for COVID-19 domain data. Our findings show that the proposed approach helps improve terminology translation in COVID-19 domain data. Our in-depth analysis showed that adapting a single instance for a larger number of epochs helps improve the translation of domain-specific terms.

The rest of this paper is organised as follows. Section 2 discusses work related to our study. Section 3 gives details about the data we used in our experiments. We describe our methodology in Section 4. Our NMT model is explained in Section 5. Experiments and results are covered in Section 6. Finally, Section 7 summarises our work and discusses possible future research ideas.

2 Related Work

Although NMT models have shown significant improvement in many translation tasks, translating terms of specific domains, such as medical or technical (Ao and Acharya, 2021), still remains challenging for NMT. Numerous methods have been proposed to improve term translation in NMT. These include (i) fine-tuning with domain-specific data: these help NMT models understand and translate domain-specific terms more effectively (Nayak et al., 2020), (ii) data augmentation approaches, including generating synthetic data through back-translation or self-training: these methods expose the NMT model to a variety of examples, ultimately enhancing term recognition and translation (Fernando et al., 2020), (iii) incorporating external resources like glossaries, dictionaries, or terminology databases can assist NMT models in understanding and translating specialised terms more effectively (Scansani and Dugast, 2021), (iv) terminology injection during inference, using techniques like inline tags (Dinu et al., 2019), source-target alignments (Dougal and Lonsdale, 2020), or fixed source positions (Niehues, 2021) for reference terms, helps produce translations with accurate domain-specific terminology, and (v) introducing auxiliary objectives during training such as predicting masked source terms or generating domain-specific inflections (Michon et al., 2020) can handle domain-specific terms better during inference.

Standard NMT domain adaptation involves fine-tuning a generic NMT model using domain-specific data. Accordingly, it is essential to consider factors such as similarity or distinct domain features that characterise the specialised field to effectively select the appropriate data. In their study, Farajian et al. (2017) showed that fine-tuning a generic model using a sentence highly similar to the source-test sentence can improve the usage of domain-specific terminology after adaptation. Likewise, Li et al. (2018) conducted an experiment in which they fine-tuned a generic model on a small subset of bilingual training data acquired through a similarity search with the source test sentence. Their findings also indicated an improvement in translation performance. In their experiments, both Farajian et al. (2017) and Li et al. (2018) showed how only a small set of sentences based on similarity to that of the test sentence is sufficient to improve the quality of translation. However, it is crucial that the sentences used for fine-tuning exhibit considerable similarity to the sentence being translated; otherwise, this can lead to a deterioration in translation quality.

Unlike Farajian et al. (2017) and Li et al. (2018), who fine-tuned their models on fewer sentences for each test instance, Chen et al. (2020) took a different approach by employing n-gram matching for the entire test set. Their study focused on matching and selecting n-grams from the training data which are most relevant to the entire test set rather than just individual sentences. By doing so, they were able to create a more comprehensive fine-tuning dataset, which in turn led to improved terminology translation.

[1] https://www.statmt.org/wmt21/terminology-task.html

Numerous studies have investigated ways to better incorporate technical terms into MT systems during inference. For example, Dinu et al. (2019) added special tags to the source text sentence by identifying domain-specific terms. After translating, they found that these tags were correctly replaced with the appropriate terms in the target language. A similar approach was tried by Song et al. (2019), where they replaced specific phrases in the source text with pre-selected, domain-specific translations before translating. This made it easier for the system to use the correct domain-specific terms in the final translation. Michon et al. (2020) carried out a comparative analysis by experimenting with variations of inline terminology tags and discussed the optimal settings in the experiment that helped improve terminology translation. In their work, Dougal and Lonsdale (2020) added domain-specific terminology after the translation process as a post-processing step, replacing incorrect terms with approved ones using source-target alignments. This approach offers the benefit of not requiring the translation model to handle tags, so it could potentially be used to introduce terminology to MT system outputs. However, the effectiveness of this method relies on an effective alignment model. In their work, Chen et al. (2020) developed constraint-aware training data by randomly choosing phrases from the reference translation to serve as constraints and subsequently merging them into the source sentence with the help of a separation symbol. Their method does not require alignments and solely depends on bilingual dictionaries during translation. They inserted the reference terminology at a fixed location in the source text, facilitating the model's learning of proper alignment. Similarly, Niehues (2021) also placed the reference terminology at a fixed point within the source text. However, his primary focus was on using the lemma of the term, which encouraged the model to learn the appropriate inflections for the given terminology. In their experiments, Lee et al. (2021) presented a technique that estimates the range of masked source terms during MT training, facilitating the integration of multi-word domain-specific terms in the translation process. They found that their models produced performance similar to that of Chen et al. (2020) in terms of single-word accuracy but improved performance when it came to translating multi-word terms.

Nayak et al. (2020) conducted an experiment in which they mined sentences from a large general domain corpus based on the presence of domain-specific terms in the test data. They then utilised the extracted data to fine-tune the model and observed improvements in terminology translation. Similar experiments were carried out by Haque et al. (2020), with their approach also demonstrating improvements in terminology translation. In our experiment, we employ an approach similar to that used by Nayak et al. (2020) and Haque et al. (2020). However, we take it further by performing extraction and adaptation for each instance in the test data as in Farajian et al. (2017). This means that, instead of using a predetermined set of sentences containing domain-specific terms, we adapt our model on a per-instance basis, allowing the model to better handle the domain-specific terminology in each test sentence. Our proposed approach aims to provide a more tailored and flexible adaptation process, potentially resulting in more significant improvements in translation performance and domain-specific term management.

3 Dataset

In our experiment, we used French-to-English parallel data from WMT2021,[2] which includes sources such as Europarlv10, ParaCrawlv7.1, News Commentary v16, UN Parallel Corpus V1.0, CommonCrawl corpus, and 10^9French-English corpus. We combine these datasets, remove duplicates, and tokenise the text using Moses (Koehn et al., 2007)[3] tokeniser scripts. The resulting dataset consists of 44M unique sentence pairs. The terminologies for French-to-English

[2]https://www.statmt.org/wmt21/terminology-task.html

[3]https://github.com/moses-smt/mosesdecoder

translation were obtained from the TICO-19 project by Anastasopoulos et al. (2020),[4] focusing on the COVID-19 domain. There are 595 unique domain-specific terms, and the test set comprises a total of 2100 sentences.

4 Methodology

4.1 Domain adaptation using terminology-aware mining

Terms or phrases appearing in domain-specific data may encode meanings or usages different to those when they appear in generic data. In order to obtain correct translations for terms or phrases of a domain text, Translation Service Providers (TSPs) usually use domain-specific terminology or glossaries. Obtaining such terminological resources is challenging as this process can be very expensive in terms of both cost and time. Automatically identifying and extracting domain-specific terminology from training data or external resources and integrating them into industrial translation workflows can partly alleviate this problem (Haque et al., 2018; Mouratidis et al., 2022). A notable obstacle to these approaches could be the training itself. Since the NMT training process is a highly time-consuming task, integrating terminology at training or fine-tuning from scratch is not a feasible solution. In fact, this is unimaginable in an industrial setting where terminologies are often needed to be updated for translating newly arrived documents with particular styles. We could have certain situations where the training time may not be a concern, and the entire terminology is available at the training. However, an NMT system trained with added terminology or that uses terminology during inference does not guarantee to generate translations with expected terms. Adapting a generic NMT system to a specific domain and obtaining accurate translations for the domain-specific terms can be more challenging when one does not have domain-specific data. In this study, we investigate this specific scenario (i.e. unavailability of domain text) and systematically make use of large general-domain data in order to fine-tune our MT systems. First, we extract terms from the source sentence to be translated based on the named tags provided in the test data. Then we mine parallel sentences from the general domain parallel data based on the frequency of occurring the extracted domain-specific terms in the parallel sentences. The extracted sentences are then used to fine-tune our NMT models. Note that the entire process (term extraction from the test sentence to be translated and mining parallel sentences from large generic data) is characterised as *on-the-fly* instance-based adaptation by Farajian et al. (2017).

Algorithm 1 Algorithm for Instance-Based Adaptation Using Terminology-Aware Mining

for *src_sent* in *tst_set* **do**
  D_{Trm} = **Extract_trm**(*src_sent*)
  R_{Sent}= **Retrieve**(*max_trm*(Data,D_{Trm}))
  F_{MT}=**Finetune**(G_{MT},R_{Sent})
  Translate(F_{MT},*src_sent*)
end for

In Algorithm 1, we present our approach for instance-based adaptation using terminology-aware mining. The algorithm leverages domain-specific terminology to adapt the NMT system by fine-tuning it on relevant instances from the general-domain parallel data.

The algorithm picks a source sentence (*src_sent*) from the test set (*tst_set*) and performs the following steps:

- **Extract** domain-specific terminology (D_{Trm}) from the source sentence to be translated

[4]https://tico-19.github.io/

using the **Extract_trm** function. This function identifies terms that are specific to the given domain within the source text using the annotated tags provided in the test data.

- The **Retrieve** function, used with the max_trm parameter, mines sentences (R_{Sent}) matching most domain-specific terms D_{Trm} from general-domain data.
- Fine-tune the general-domain MT system (G_{MT},R_{Sent}) using the retrieved sentence (R_{Sent}). The **Finetune** function updates the model parameters based on the domain-specific instance, resulting in a fine-tuned MT system (F_{MT}).
- **Translate** the source text (*src_sent*) using the fine-tuned MT system (F_{MT}) to generate a domain-adapted translation.

5 Experimental Setup

5.1 NMT Model

The mBART (Multilingual BART) (Liu et al., 2020) model is a multilingual extension of the BART (Bidirectional and Auto-Regressive Transformers) (Lewis et al., 2020) model, a sequence-to-sequence pre-training framework for natural language understanding and generation tasks. mBART uses a standard sequence-to-sequence Transformer architecture with 12 layers of the encoder and 12 layers of the decoder, where each layer has 16 heads and a model dimension of 1024. The model is trained on large-scale multilingual data, enabling it to perform well across various languages and tasks. mBART is pre-trained using a combination of denoising auto-encoding and masked language modeling, involving reconstructing corrupted text or predicting masked tokens. One key feature of mBART is its shared vocabulary across languages, making it easier to fine-tune the model for downstream tasks, such as MT, summarisation, or sentiment analysis. Leveraging its pre-trained knowledge, mBART achieves state-of-the-art performance on various NLP tasks and languages.

In this study, we wanted to see how our proposed domain adaptation method of terminology-aware fine-tuning would work on mBART. We placed particular emphasis on terminology translation (cf. Section 4). Our experiment used mBart-50-many-to-many[5] MT, a strong checkpoint based on mBart, as our baseline model. In our experiment, we utilised the following hyperparameters: a learning rate of 2e-5, a weight decay of 0.01, a training batch size of 32, and an evaluation batch size of 32.

We apply the instance-based adaptation on mBART (see Algorithm 1). We expect that our terminology-aware mining techniques will be able to help adapt the baseline so that the model can correctly translate a larger number of domain-specific terms. In order to thoroughly assess how our proposed terminology-aware adaptation process works on terminology translation, we carried out experiments with a different number of instances (one, three, and five) and epochs (one, three, and five) for fine-tuning. By examining the impact of varying numbers of sentence and epoch combinations on the model's performance and its handling of domain-specific terms, we aimed to gain a deeper understanding of the potential benefits and limitations of the proposed approach.

6 Experiments and Results

We evaluated our MT systems using BLEU (Papineni et al., 2002), COMET (Rei et al., 2020), NIST (Przybocki et al., 2010) and Term Count as our evaluation metrics. Term Count (TC) measures the number of occurrences of domain-specific terms accurately translated by the MT system. Table 1 shows the results that we obtained through our experiments. It displays BLEU,

[5]https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

TC, NIST and COMET scores for each of the test scenarios described in Section 5. We can see from the table that TC improves in two cases over the baseline. In both cases, the improvement occurs for a single sentence with three and five epochs. We conducted statistical significance tests for two system comparisons using bootstrap resampling (Koehn, 2004) and found that the differences in scores were statistically significant.

Furthermore, the improvement in TC over the baseline MT system suggests that the proposed adaptation method effectively improves the generic NMT system's ability to handle domain-specific terminology. In order to further understand the results in Table 1, we visualise the results in Figures 1, 2, 3 and 4.

Sentence	Epoch	BLEU	Term Count	COMET	NIST
Base		27.63	2175	0.844	10.80
1	1	26.60	2155	0.825	09.75
1	3	27.21	2191	0.826	09.92
1	5	27.68	2190	0.822	10.06
3	1	26.12	2115	0.829	09.81
3	3	26.24	2111	0.832	10.08
3	5	26.43	2094	0.832	10.21
5	1	25.39	2119	0.817	09.38
5	3	26.18	2109	0.833	10.11
5	5	26.30	2089	0.835	10.23

Table 1: Results of instance-based adaptation using terminology-aware mining.

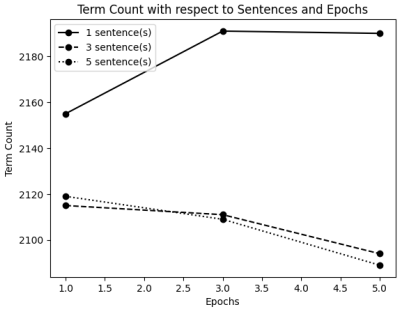

Figure 1: Term count scores in relation to the number of sentences and epochs used in the adapted model.

In Figure 1, we show the performance of our adapted MT systems for the French-to-English translation task using TC scores. The graph presents the results for different combinations of sentences (one, three, and five) and epochs (one, three, and five) in the fine-tuning process. The x-axis represents the number of epochs, and the y-axis represents TC. The lines with varying markers correspond to the different epoch combinations. In Figure 1, we observe that increasing the number of sentences used for fine-tuning does not contribute significantly to the improvement of terminology translation performance. Rather, we find that increasing the number of epochs for a single sentence is more beneficial. This finding suggests that the model may benefit from more focused training, concentrating its learning efforts on a smaller number of sentences for a longer period of time (i.e., more epochs). By doing so, the model can potentially gain a deeper understanding of the specific domain terminology, which in turn can lead to better translation performance with respect to the domain-specific terms.

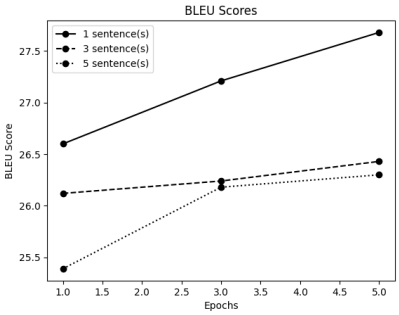

Figure 2: BLEU scores in relation to the number of sentences and epochs used in the adapted model.

In Figure 2, we have plotted the performance of our adapted MT systems using BLEU scores to analyze the relationship between the number of sentences, the number of epochs, and the translation quality. The x-axis represents the number of epochs, and the y-axis represents the BLEU scores. The lines with varying markers correspond to different sentence combinations. We observe that increasing the number of sentences does not proportionally improve the translation quality. This pattern resembles the findings in terms of TC (as in Figure 1), where adding more sentences offered no improvement. This suggests that adding more sentences to the fine-tuning data may not guarantee better translation outcomes.

While the graphs for TC and BLEU display a similar trend, it is crucial to understand that an increase in the BLEU score does not necessarily indicate an improvement in terminology. In fact, alterations made to the adapted model might have led to improvement in the meta-language without directly translating to substantial improvements in the translation of domain-specific terms.

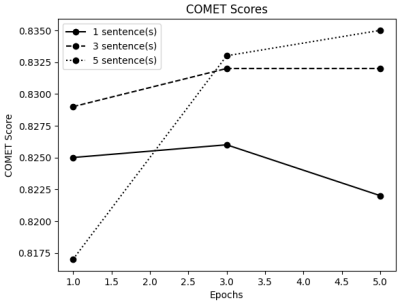

Figure 3: COMET scores in relation to the number of sentences and epochs used in the adapted model.

In Figure 3, we plotted our MT systems' performance based on COMET scores to analyse the relationship between the number of instances used for training and epochs. The x-axis represents the number of epochs, and the y-axis represents the COMET scores. The lines with varying markers correspond to different sentence combinations. We see that the COMET scores exhibit a different trend. When the number of sentences is increased, the translation quality measured by COMET scores appears to improve. This contrasts with the trends observed in terms of TC and BLEU. We also see that increasing the number of sentences did not consistently lead to betterment in translation quality.

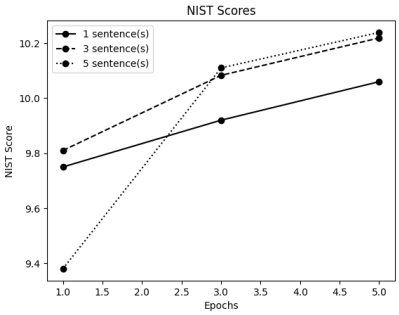

Figure 4: NIST scores in relation to the number of sentences and epochs used in the adapted model.

Similarly, in Figure 4, we plotted the performance of our MT systems based on NIST scores to analyse the relationship between the number of instances used for training and the number of epochs. We observe that increasing the number of epochs appears to benefit the quality of translation. Furthermore, we observed that training a model with more instances and epochs yields better results.

The discrepancy between the trends observed for three metrics (COMET, BLEU, NIST and TC) could be attributed to the differences in the evaluation metrics. While TC, NIST, and BLEU scores focus on specific aspects of translation quality, such as the handling of domain-specific terminology and *n*-gram overlaps between the reference and the translation, the COMET metric is designed to provide a more holistic assessment of translation quality by considering factors such as fluency, adequacy, and style.

6.1 Analysis of Terminology Improvements

Table 1 presents the results of our experiments aimed at improving terminology translation using instance-based adaptation. We discovered that the TC scores for the adapted MT system are found to be high in two cases (i.e. setup: a single sentence using three and five epochs). As for analysing translations produced by the MT systems, we choose the best-performing adapted MT system (i.e., one sentence and three epochs).

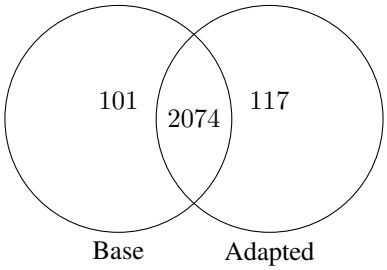

Figure 5: Venn diagram comparing terminology translation counts of the baseline and best domain-adapted MT system.

We produce a Venn diagram to visually compare and better understand how terms are translated by the baseline and the best domain-adapted MT systems. We show the Venn diagram in Figure 5. The diagram has two overlapping circles, showing the separate terminology counts produced by each of the MT models. The left circle, labeled “Base”, represents the baseline

MT system and contains 101 terms. This area represents the unique terminology translation counts from the baseline model. The right circle, labelled "Adapted", represents the best domain-adapted model and contains 117 unique domain-specific terms. The area representing the overlap between both circles contains 2074 terms. This is shared terminology translation counts from both models.

Table 2: Example: adapted MT system correctly translates terminology.

Source	dans environ 14 % des cas , la COVID-19 entraîne une atteinte plus sévère nécessitant une hospitalisation , tandis que les 6% de cas restants développent une forme grave de la **maladie** nécessitant des soins intensifs .
Reference	in ca 14% cases , covid-19 develops into a more severe disease requiring hospitalisation while the remaining 6% cases experience critical **illness** requiring intensive care .
Baseline MT	in about 14% of cases, covid-19 causes a more severe condition requiring hospitalization, while the remaining 6% develop a serious form of the **disease** requiring intensive care.
Adapted MT	in about 14% of the cases, covid-19 leads to more severe illness requiring hospitalization, while 6% of the remaining cases develop a serious form of serious **illness** requiring intensive care

To further understand how the two models differ when it comes to the quality of terminology translation, we select an example sentence from the test set. In Table 2, we present translations of the sentence we picked by the baseline and adapted MT systems. We can see from the table that the adapted MT system demonstrates improvement over the baseline MT system, where the domain term "maladie" in the source sentence is accurately translated as "illness" by the adapted MT system. In contrast, the baseline system incorrectly translates it as "disease". However, it is essential to note that the baseline system still provides a decent translation. While it may not capture the exact terminology, the overall semantic content of the sentence is preserved, demonstrating the robustness of the baseline system.

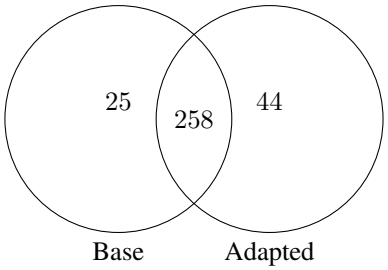

Figure 6: Venn diagram comparing multi-word terminology counts of the baseline and best domain-adapted MT system.

We observed that the adapted MT system better handles the translation of multi-word terms. We show a Venn diagram in Figure 6 where the left circle labelled as "Base" represents the baseline MT system and contains 25 multi-word terms. This indicates the unique terminology translation counts from the baseline model. The right circle labelled as "Adapted" represents the best domain-adapted model and contains 44 unique multi-word domain-specific terms. The area

Table 3: Example: adapted MT system correctly translates multi-word terminology.

Source	la ventilation mécanique devient plus complexe avec le développement du **syndrome de détresse respiratoire aiguë** (SDRA) au cours de la COVID-19 et l' oxygénation devient plus difficile .
Reference	mechanical ventilation becomes more complex as **acute respiratory distress syndrome** (ards) develops in covid-19 and oxygenation becomes increasingly difficult .
Baseline MT	mechanical ventilation becomes more complex with the development of **acute respiratory disorder syndrome** (sdra) during covid-19 and oxygenation becomes more difficult.
Adapted MT	mechanical ventilation becomes increasingly complex as **acute respiratory distress syndrome** (ards) develops in covid-19 and oxygenation becomes increasingly difficult.

representing the intersection between both circles contains 251 terms. This is shared terminology translation counts by both MT models. In Table 3, we show another example translation. This time, we chose a source sentence that contains a multi-word term. We see from the table that the adapted MT system shows improvement over the baseline MT system where the multi-word term "syndrome de détresse respiratoire aiguë" in the source sentence is accurately translated as "acute respiratory distress syndrome " by the adapted MT system. In contrast, the baseline system incorrectly translates it as "acute respiratory disorder syndrome".

7 Conclusion and Future Work

This study presents a terminology-aware instance-based domain adaptation method. We tested our method for English-to-French translation. Our results demonstrate that the proposed approach helps improve terminology translation. Furthermore, we discover that increasing the number of sentences used for fine-tuning does not significantly impact the improvement of terminology translation performance. Instead, a more efficient strategy appears to be one that considers a high number of epochs for a single sentence. This observation suggests that the model may benefit from more focused training, concentrating its learning efforts on a single sentence over an extended period (i.e., more epochs). We evaluated our MT systems using BLEU, NIST and COMET evaluation metrics. We observe that the BLEU metric correlates with the correct TC, while the COMET metric shows improvements for the adapted model with an increased number of sentences. NIST metric shows improvement for a higher number of instances and epochs. We also found that the adapted model outperformed the baseline when it comes to translating multi-word terms. Our current proposed approach fine-tunes all instances, irrespective of whether a test instance requires fine-tuning or not, which may lead to the deterioration of translation quality for some sentences. In the future, we plan to identify those sentences that require fine-tuning and adapt only to them.

Acknowledgments

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224 and Microsoft Research Ireland.

References

Anastasopoulos, A., Cattelan, A., Dou, Z.-Y., Federico, M., Federmann, C., Genzel, D., Guzmán, F., Hu, J., Hughes, M., Koehn, P., Lazar, R., Lewis, W., Neubig, G., Niu, M., Öktem, A., Paquin, E., Tang, G., and Tur, S. (2020). TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Ao, S. and Acharya, X. (2021). Learning ULMFiT and self-distillation with calibration for medical dialogue system. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 196–203, Trento, Italy. Association for Computational Linguistics.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chen, G., Chen, Y., Wang, Y., and Li, V. O. (2020). Lexical-constraint-aware neural machine translation via data augmentation. In Bessiere, C., editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, (IJCAI 2020a)*, pages 3587–3593. ijcai.org.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dinu, G., Mathur, P., Federico, M., and Al-Onaizan, Y. (2019). Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Dougal, D. K. and Lonsdale, D. (2020). Improving NMT quality using terminology injection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4820–4827, Marseille, France. European Language Resources Association.

Farajian, M. A., Turchi, M., Negri, M., and Federico, M. (2017). Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137, Copenhagen, Denmark. Association for Computational Linguistics.

Fernando, A., Ranathunga, S., and Dias, G. (2020). Data augmentation and terminology integration for domain-specific sinhala-english-tamil statistical machine translation. *ArXiv*, abs/2011.02821.

Haque, R., Moslem, Y., and Way, A. (2020). Terminology-aware sentence mining for NMT domain adaptation: ADAPT's submission to the adap-MT 2020 English-to-Hindi AI translation shared task. In *Proceedings of the 17th International Conference on Natural Language*

Processing (ICON): Adap-MT 2020 Shared Task, pages 17–23, Patna, India. NLP Association of India (NLPAI).

Haque, R., Penkale, S., and Way, A. (2018). Termfinder: log-likelihood comparison and phrase-based statistical machine translation models for bilingual terminology extraction. *Language Resources and Evaluation*, 52:365–400.

Koehn, P. (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.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (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.

Lee, G., Yang, S., and Choi, E. (2021). Improving lexically constrained neural machine translation with source-conditioned masked span prediction. 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 743–753, Online. Association for Computational Linguistics.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (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.

Li, X., Zhang, J., and Zong, C. (2018). One sentence one model for neural machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Luong, M.-T. and Manning, C. (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.

Michon, E., Crego, J., and Senellart, J. (2020). Integrating domain terminology into neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mouratidis, D., Mathe, E., Voutos, Y., Stamou, K., Kermanidis, K. L., Mylonas, P., and Kanavos, A. (2022). Domain-specific term extraction: A case study on greek maritime legal texts. In *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*, SETN '22, New York, NY, USA. Association for Computing Machinery.

Nayak, P., Haque, R., and Way, A. (2020). The ADAPT's submissions to the WMT20 biomedical translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 841–848, Online. Association for Computational Linguistics.

Niehues, J. (2021). Continuous learning in neural machine translation using bilingual dictionaries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 830–840, Online. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Przybocki, M., Peterson, K., Bronsart, P., and Sanders, G. (2010). The nist 2008 metrics for machine translation challenge - overview, methodology, metrics, and results.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (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.

Sato, S., Sakuma, J., Yoshinaga, N., Toyoda, M., and Kitsuregawa, M. (2020). Vocabulary adaptation for domain adaptation in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4269–4279, Online. Association for Computational Linguistics.

Scansani, R. and Dugast, L. (2021). Glossary functionality in commercial machine translation: does it help? a first step to identify best practices for a language service provider. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 78–88, Virtual. Association for Machine Translation in the Americas.

Song, K., Zhang, Y., Yu, H., Luo, W., Wang, K., and Zhang, M. (2019). Code-switching for enhancing NMT with pre-specified 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 449–459, Minneapolis, Minnesota. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, pages 5998–6008, Long Beach, CA, USA. Curran Associates Inc.