Contextual Selection of Pseudo-terminology Constraints for Terminology-aware Neural Machine Translation in the IT Domain

Benjamin Pong
University of Washington
Seattle WA, USA
{benpong}@uw.edu

Abstract

This system paper describes the development of a Neural Machine Translation system that is adapted to the Information Technology (IT) domain, and is able to translate specialized IT-related terminologies. Despite the popularity of incorporating terminology constraints at training time to develop terminology-aware Neural Machine Translation engines, one of the main issues is: In the absence of terminology references for training, and with the proliferation of source-target alignments, how does one select word alignments as pseudo-terminology constraints? The system in this work uses the encoder's final hidden states as proxies for terminologies, and selects word alignments with the highest norm as pseudo-terminology constraints for inline annotation at run-time. It compares this context-based approach against a conventional statistical approach, where terminology-constraints are selected based on a low-frequency threshold. The systems were evaluated for general translation quality and Terminology Success Rates, with results that validate the effectiveness of the contextual approach.

1 Introduction

This paper describes UW-BENMT's submission[1] to WMT2025 Terminology Translation Shared Task (Semenov et al., 2025). The aim of this edition of the shared task is to evaluate how well machine translation systems can handle specialized terms in specific domains where terminology accuracy and consistency are critical. Although general machine translation systems have improved significantly, specialized terminology remains a challenge (Semenov et al., 2023; Alam et al., 2021), and this task evaluates the effectiveness of integrating terminology dictionaries into machine translation engines in the Information Technology (IT) domain. This is a highly practical task as machine translation engines that have been adapted to this domain can potentially assist in the international communication of technical APIs, manuals and DevOp guides across IT teams that operate in multilingual environments. This task involves segment-level terminology translation for three language pairs: English $\rightarrow$ {German, Spanish, Russian}. The provided data consists of 500 parallel sentences per language pairs, with reference terminology dictionaries.

2 Related Work

There are two major approaches to terminology translation, with the first being lexical constrained decoding where the target terminologies entries are forced to match the source side lexical terms as decoding-time constraints (Chatterjee et al., 2017). However, one major shortcoming of this strategy is the computational overload as the number of terminology constraints increase. To address this issue, Dinu et al. (2019) pioneered a methodology to train neural machine translations engines to recognize terminology constraints. The focus of their approach is to annotate the data with source-target terms inline as soft constraints (i.e., a form of data augmentation). The success of this approach is reflected in popularity of system submissions that adopted it at the WMT2021 and WMT2023 Terminology Translation Shared Tasks (Ailem et al., 2021; Nieminen, 2023; Bogoychev and Chen, 2023; Park et al., 2023). Different alternatives to implementing this approach have been proposed, such as masking out the source-side terms (Ailem et al., 2021; Liu et al., 2023), which have been argued to surpass simple inline term annotations.

As pointed out by Bogoychev and Chen (2023), one of the main challenges of terminology transla-

[1]Repository for system development can be found at https://github.com/Benjamin-Pong/Terminology-Neural-Machine-Translation-IT-domain.git

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 1292–1301
November 8–9, 2025.

tion is that terminology dictionaries are not readily available for training, and their proposal differs from Dinu et al. (2019) in that they devised a way to automatically create terminology constraints for existing training corpora. The development of pseudo-terminology constraints is a challenge in itself because it is difficult to select word alignments that are most representative of 'technical terminologies'. While Dinu et al. (2019) and Bogoychev and Chen (2023) used randomly selected terms, other systems construed domain-specific terms as low frequency lexical items (Semenov et al., 2023; Park et al., 2023), a common approach that is practiced in works beyond the shared task (Koehn and Knowles, 2017; Yuan et al., 2018; Bowker, 2021).

The main contribution of this system paper is to shed light on the context-driven approach to selecting pseudo-terminology constraints for IT-adapted NMT training, and present this as an alternative to the frequency-based approach. It focuses on improving the psuedo-terminology creation process in the absence of training terminology dictionaries by comparing two methods; frequency-based, context-based.

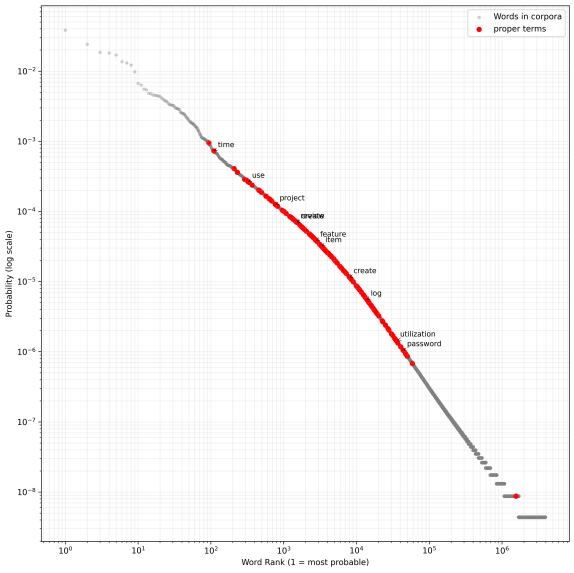

Figure 1: Rank-frequency plot showing the frequency distribution of proper terms against all words in a corpora

3 Selection of Pseudo-terminology Constraints

3.1 Proxies for Terminology Selection: Frequency vs Context

There are two main reasons why frequency may not be the best proxy for terminology selection. The first reason is motivated by the statistical frequencies of proper terms against the tokens in publicly available corpora. Consider the Rank-Frequency Distribution in Figure 1, where the frequencies of the provided data's English terminologies were compared against the frequencies of all English tokens from the Europarl and WikiMatrix data[2]. The frequencies were rescaled to a value between 0 to 1. Based on the Rank-Frequency distribution of terminologies, terminologies lie in mid-rank, contrary to conventional assumptions that terminologies are low-frequencies.

The second reason lies in the fact that lexical terms selected as proper terms are not treated as proper terms in all contexts. A vast majority of the proper terms in the development set are also treated as random terms and do not have any domain-specific meanings. Consider the way 'create', a mid-ranked frequency word based on the rank-frequency distribution, is used in following sentences taken from the provided data: (1) *With the Story Builder, you can create stories to visualize information with charts and table* and (2) *Activate or Create your Data Provider Profile.*

"Create" is treated as a specialized IT-terminology in the first sentence, but not in the second sentence. This ambiguity is difficult to resolve but one could speculate that "create" in the first instance is used in the context of a software application assisting with the creation, which renders it a specialized IT-terminology. Since it is challenging to define deterministic semantic rules for domain-specific terminologies, and neither does a frequency-based approach suffice since 'create' is not exactly a low-frequency word, an approach that is agnostic to these two conditions should be explored.

3.2 Terminology Constraint Filtering by Norm of Top Most Hidden States

Inspired by Wu et al. (2024) and Schakel and Wilson (2015) who established a theoretical and empirical connection between important-words and context for general machine translation, this paper adapts their methodology to terminology-aware machine translation. More specifically, the methodology is applied at the pseudo-terminology constraints creation stage of the end-to-end NMT pipeline, where source-target alignments are filtered based on how "important" they are in seman-

[2]There are a total of 4 million unique English words in both of these corpora.

Language Pair	Dataset	Sentences	Out-domain Training	Out-domain For CED filtering
en-es	WikiMatrix	6,452,177	30,000	8,303,595
	Europarl	1,881,418		
	Total	**8,333,595**		
en-de	WikiMatrix	6,227,188	30,000	8,025,709
	Europarl	1,828,521		
	Total	**8,055,709**		
en-ru	WikiMatrix	5,203,872	30,000	10, 551, 783
	ParaCrawl v9	5,377,911		
	Total	**10,581,783**		

Table 1: This table shows the dataset sources and statistics for each language pair. It also shows the amount of data that was used for CED filtering, from which approximately 2.2 million samples were selected for training purposes.

tic meaning. The norm of the encoder's final hidden states represents contextually important information within lexical items (Wu et al., 2024).

While the importance of context in resolving ambiguities in machine translation (Maruf et al., 2019; Post and Junczys-Dowmunt, 2024) is not entirely new, its effects on terminology translation has yet to be explored, which is the focus of this system paper. Given a set of source-target alignments for a sentence-pair, and given that the norm of the encoder's final hidden states is computed for each source word, the word-alignment pair(s) whose source-word has the highest norm will be chosen as a pseudo-terminology constraint for augmentation.

4 Neural Machine Translation Training

4.1 Training Data: IT-domain Parallel Data Selection using Cross-Entropy Difference

Due to data sparsity in the IT-domain, it is necessary to procure sufficient quality data for system developments. To do so, Cross Entropy Difference (CED) (Moore and Lewis, 2010) was applied to select IT-specific data from a larger corpus of out-of-domain or general domain content.

For CED to be effective, the assumption is that there is already some available in-domain data. Inspired by Moslem et al. (2023), LLM was used to generate synthetic monolingual (English) in-domain parallel data using the gold terminologies provided by the organizers. The data provided contains 500 parallel sentences per language pair, and each sentence-pair comes with a set of terminology mappings. Using only the source-side terminologies as inputs into Aya-expanse-8b (Dang et al., 2024) with temperature=0 and top-p sampling of 1, synthetic monolingual data was generated with the following prompt: `Please use the terms '{terms}' to generate {number_of_sents} full sentences in {source_language}-{target_language} whose content is related to information technology.`

Note that eventhough parallel data was generated, for the purposes of the task at hand, the English source sentences were used for subsequent indomain and outdomain language modeling. A total of 30K IT-related synthetic sentences were generated and treated as in-domain data. The average sentence length is 50-75 tokens.

Table 1 shows the sets of publicly-available parallel corpora that were considered for each language pair. These serve as the general-domain content. For each set of corpora, 30K English sentences were randomly sampled to train an out-domain 4-gram language model (with Laplace Smoothing). To ensure that the in-domain and out-domain models are comparable, apart from controlling for the size of training data, the sentences selected for out-domain model training also had an average of 50-75 tokens.

The remaining sentences were scored according to the cross-entropy difference between the in-domain and out-domain language models, and ranked in non-increasing order since lower entropy signifies a closer match to the in-domain data[3]. Throughout this process, several preprocessing strategies were applied to reduce noise in the publicly available data. FastText (Joulin et al., 2016) with a threshold of 0.9 was used to remove sentence-pairs where either the source or target sentence was not the desired language. Sentence-

[3]Discussion of CED is beyond the scope of this paper. I encourage the reader to refer to Moore and Lewis (2010) for details.

pairs that contained only punctuations and numbers were also removed. The top 20% was chosen as in-domain training data, and the bottom 20% as out-domain data, which amounts to 2.2 million parallel sentences per language pair. Additional checks were performed to ensure that parallel sentences chosen for training did not overlap with the provided data so that the provided data can be held-out for evaluation.

The overall effectiveness of CED for data selection was validated on downstream general translation quality, and hence the main approach to data selection for subsequent experiments for terminology translation. See Section 6 for details.

4.2 Experiments

The baseline uses the training data as they are, without any lexical constraints being augmented inline at training time. Since the chief focus of this paper is to compare the effects of using (low)frequency and encoder-based contextual scoring for pseudo-terminology selection, two more experimental configurations were designed. The first being training data where only low frequency words were selected as pseudo-terminologies for augmentation (i.e FreqTerm). The second configuration consists of data where only words with the highest norm computed from the final layer of the encoder's hidden states (i.e ContexTerm). For these two configurations where lexical constraints are augmented inline, similar to Dinu et al. (2019), only 10% of the training data is augmented, amounting to around 200K randomly selected sentence-pairs. The special token <*src*> was used to mark the start of the source word and <*tgt*> was used to mark the start of the corresponding target word.

The first part of creating the pseudo-terminology involved extracting word alignments between the source and target languages by using a state-of-the-art neural word aligner, Awesome Align (Dou and Neubig, 2021). Alignments between stopwords were removed. Given that awesome align does not account for multi-word alignment, further lightweight processing was applied to merge consecutive source words that were mapped to the same target word to produce multiword source-target word alignments.

To address the proliferation of word-alignments, the next step is to implement the frequency and contextual approaches for selecting word-alignments to be used as pseudo-terminology constraints.

The former was straightforwardly implemented by first computing a frequency distribution (normalized to values between 0 and 1) for the source-side words using the source-side's training data. Only source-target aligned pairs whose source words' probabilities were at most 10^{-5}, were selected. This low-frequency threshold is not arbitrary but instead, motivated by the aforementioned frequency distribution of terminologies (See Section 3.1 and Figure 1). For a multiword source word, if a subword met the threshold, the entire multiword source word was chosen as a suitable candidate for term annotation. Furthermore, since each sentence may have multiple candidate word-aligned pairs that meet this threshold, randomly chosen aligned word-pairs were chosen for augmentation per sentence. Note that the number of word aligned pairs chosen for augmentation can be adjusted by the engineer.

As for the context-scoring configuration, a pre-trained encoder-decoder NMT model (Ng et al., 2019) was used to extract the final layer's hidden representation per token. Only the source-side (English) needs to be encoded for all language pairs. For each multiword source word, max-pooling is applied to compute a unified final hidden representation, followed by a computation of the norm of this hidden representation. These words were ranked in non-increasing order, with the source word with the highest norm being selected, and consequently, its corresponding source-target alignment, was selected as a terminology constraint.

To ensure a fair comparison between FreqTerm and ContexTerm, two factors were considered; first, the number of terminology constraints selected per sentence for augmentation, and the number of sentences that contain the constraints. The latter was standardized across both experiments by choosing only 10% of the data to be augmented. The former was enforced by enforcing identical thresholds on hyperparameter *k*; k-randomly selected low frequency word(s) and the top-k highest norm(s).

The training data per language pair were tokenized using Moses (Koehn et al., 2007). Byte Pair Encoding (BPE) for subword segmentations (Sennrich et al., 2016) were independently applied to both the source and target languages with 32000 merge operations. 500,000 sentences from the training data were selected to generate BPE codes.

4.3 Model Architecture

All systems were Transformer-based networks, trained (Vaswani et al., 2023) using *fairseq* (Ott et al., 2019). Training configurations were specif-

ically optimized for each language pair. See Appendix A for the hyperparameters. En-es and en-ru were trained for minimum of 35 epochs and a maximum of 50 epochs with early stopping, while en-de was trained for a minimum of 50 epochs and a maximum of 100 epochs with early stopping. Uniformly across all models, the last 7 checkpoints were averaged and used for decoding, with a beam size of 5.

5 Evaluation

Systems were evaluated on the provided data, which was not used for training purposes, and have been verified to not overlap with the selected training data. Although examples were taken from the data to illustrate the point in Section 3.1, note that the approach in this paper does not rely on defined patterns or rules from the data, resulting in minimal risk of evaluation bias. Furthermore, the provided data is the only source of gold references for quality evaluation. With these careful considerations, the provided data was safely treated as held-out test data.

5.1 Evaluation Metrics

The translations provided by the trained models were scored against the gold references using BLEU (Papineni et al., 2002), Chrff++ (Popović, 2017) and COMET-DA (Rei et al., 2020) to assess general translation quality. Terminology Success Rate (TSR) (Semenov et al., 2023) was also used to assess the degree of occurrences of the term translations, with and without lemmatization to account for morphological complexities across the different languages, and also fuzzysearch with a threshold of 90% to account for orthographic deviations. Another reason for selecting this mode of evaluation instead of exact matching (Alam et al., 2021) is to capture adequate term translations, and also to consider the effect of syntactic contexts on the morphological shape of the terms.

5.2 Modes of Evaluation

The NMT systems were evaluated under three modes of incorporating terminology constraints at inference time, with the first mode being no terminologies (i.e., *no term*) where the source sentence is freely translated into the target language. The second mode requires the *proper* terminologies to be incorporated, and the third mode incorporates *random* source-target word alignments.

Results from NMT engines are compared between these three modes. The purpose of incorporating random source-target word alignments is to ensure that any improvements over the *no term* setting brought about by incorporating proper terminologies are not by-products of superior general translation quality (Semenov et al., 2023).

6 Results

Table 2 shows the results of all three systems (baseline, freqTerm and ContexTerm) for all language pairs across the three terminology modes. Table 3 shows the validation results for CED-selected data without any terminology incorporation for training or inference.

6.1 Translation quality

From the results in Table 2, the highest scores for translation quality metrics tend to skew toward the ContexTerm system, with the exception of en-es language pair, where FreqTerm achieved the highest BLEU and ChrF++ scores. Notably, none of the baselines achieved the highest scores.

With regards to how the terminology modes affect overall translation quality, based on the results, the highest scores tend to skew towards systems with *proper* terminology settings. However, there are several exceptions. For instance, the *no term* setting for ContexTerm has the highest BLEU score for en-es. In a similar vein, among the baseline systems, incorporating random terms or proper terms at inference time tends to result in a drop in translation quality.

6.2 Terminology Success Rate

The incorporation of soft constraints at training time increased the Terminology Success Rate (TSR) by a huge margin compared to the baselines. This is demonstrated by the fact that for all language pairs, the *proper* terminology setting achieved the highest TSR compared to the *random* and *no term*. The latter two settings tend to perform comparably low. This pattern is observed for both FreqTerm and ContexTerm, regardless of whether lemmatization was employed. These results are expected and aligns with the findings of previous works mentioned in Section 2.

Crucially, using *proper* terminology setting as a basis for comparison between FreqTerm system and ContexTerm system, the latter has the highest TSR.

Language	System	Modes	BLEU	ChrF++	COMET	TSR^{-L}	TSR^{+L}
en → es							
	Baseline	No Term	27.79	56.60	0.731	0.219	0.233
		Random	18.18	47.93	0.474	0.189	0.198
		Proper	22.35	50.78	0.482	0.421	0.423
	FreqTerm	No Term	27.42	55.79	0.732	0.222	0.232
		Random	26.08	55.20	0.720	0.227	0.238
		Proper	**29.83**	**57.94**	0.713	0.512	0.525
	ContexTerm	No Term	27.62	56.13	0.700	0.233	0.250
		Random	26.08	56.22	0.726	0.229	0.240
		Proper	27.75	57.84	**0.734**	**0.663**	**0.678**
en → ru							
	Baseline	No Term	21.02	54.53	**0.830**	0.251	0.329
		Random	19.98	52.51	0.812	0.232	0.309
		Proper	20.37	53.28	0.810	0.237	0.326
	FreqTerm	No Term	23.46	53.20	0.807	0.248	0.327
		Random	22.98	53.56	0.810	0.249	0.337
		Proper	22.63	53.91	0.811	0.489	0.571
	ContexTerm	No Term	**24.39**	53.34	0.814	0.248	0.327
		Random	22.66	53.93	0.812	0.248	0.325
		Proper	23.22	**55.34**	0.820	**0.574**	**0.654**
en → de							
	Baseline	No Term	13.91	41.96	0.656	0.151	0.142
		Random	13.21	40.0	0.41	0.150	0.141
		Proper	14.09	42.00	0.661	0.151	0.147
	FreqTerm	No Term	13.62	41.48	0.665	0.153	0.147
		Random	13.00	42.67	0.653	0.154	0.145
		Proper	13.56	44.05	**0.672**	0.654	0.645
	ContexTerm	No Term	13.10	41.04	0.66	0.152	0.150
		Random	14.43	45.23	0.670	0.177	0.171
		Proper	**14.62**	**45.98**	0.641	**0.774**	**0.755**

Table 2: Evaluation of systems by language direction, augmentation approaches and terminology modes across BLEU, ChrF++, COMET and Terminology Success Rates(TSR). TSR^{-L} refers to non-lemmatized setting while TSR^{+L} refers to lemmatized setting. Best performing systems and with their corresponding terminology settings are bolded.

Language	Domain	BLEU	ChrF++	COMET	TSR^{-L}	TSR^{+L}
en → es						
	in	27.79	56.60	0.731	0.219	0.233
	out	24.92	52.464	0.767	0.200	0.212
en → ru						
	in	21.02	54.53	0.830	0.251	0.329
	out	20.69	53.57	0.826	0.248	0.317
en → de						
	in	13.91	41.96	0.656	0.151	0.142
	out	15.50	45.68	0.69	0.172	0.194

Table 3: This table shows the evaluation of 'IT-domain' CED-selected data vs out-domain data without terminology constraints incorporated for training or inference.

With respect to the effect of lemmatization, TSR tends to be better with lemmatization for en-es and en-ru. Such improvements are consistent across all terminology modes as well, and across both systems. However, the degree of improvement is language-dependent. For instance, in the context of en-ru (*proper*), the *lemmatized* setting surpassed the *non-lemmatized* setting by at least 0.1 points. This is in stark contrast with en-es (*proper*) where the *lemmatized* setting only surpassed the *non-lemmatized* setting by 0.02 points. Interestingly, for en-de, we observe an opposite effect, where the *non-lemmatized* setting surpasses the *lemmatized* setting by a small margin. This is consistent across both systems, and across terminology modes. See Section 7 for possible explanations and implications of evaluating the adequacy of terminology translation with or without lemmatization.

6.3 Data Selection by Cross-Entropy Difference

Table 3 shows that CED-selected data scored higher in general domain translation tasks. In addition, even without terminology incorporation, using in-domain data achieved relatively higher TSR. However, German is an exception, which will be discussed in Section 7.

7 Discussion

7.1 Contextual Selection of Terminology Constraints

The comparisons between FreqTerm and ContexTerm clearly show the effectiveness of using contextually-based scoring to distill quality word-alignments to be used as terminology constraints or inline term annotations. ContexTerm surpasses FreqTerm by 0.08-0.15 points. This suggests that terminologies are construed in terms of the way they are being used in an utterance (i.e the 'create' example from Section 3), and not completely dependent on frequencies.

7.2 Effect of Contextually-selected Terminology Constraints on Translation Quality

The use of contextually-selected constraints seem to have a trickle-down effect on translation quality as well. As noted previously, systems with terminology constraints at training time tend to have higher translation quality. While this is expected (Dinu et al., 2019), the fact that BLEU , ChrF++ and COMET scores tend to be higher in ContexTerm compared to FreqTerm across all terminology modes suggest that these translation engines are learning domain-specific lexical alignments, which reduces variability, and thereby improving the translation outputs.

7.3 Terminology Success Rate: Lemmatization

The asymmetrical results between en-es and en-ru against en-de with regards to lemmatization could be attributed to the morphological properties of the languages. German, compared to Spanish has been argued to have more complex inflectional paradigms, and that simplifying them improves word alignment (Axelrod et al., 2008). This suggests that lemmatization should correlate with improved scores, but this is not the case. Perhaps the lemmatization oversimplified certain lexical items, resulting in noise. This is an interesting case and should be left for future work through error analysis.

7.4 Data Selection by Cross-Entropy Difference

The fact that CED-selected data showed marginally higher general translation and TSR for en-es and en-ru may indeed suggest that this is an effective approach to source quality parallel data for a less resourced domain. Furthermore, the higher TSR scores suggest that training on data closer to the actual domain can boost terminologies, without the incorporation of constraints at train time. However, the converse results for en-de, although disappointing, may be due to actual noise in the data. This investigation will be left for future work.

8 Conclusion

This paper addresses a challenge in terminology translation engines that incorporate term annotations at run-time: with the proliferation of source-target word alignments, how does one select the optimal subset of alignments as pseudo-terminology constraints? Results from this work show that the encoder's hidden representations serve as useful estimations of terminology constraints. It also highlights the robustness of a main-stay data selection approach in the absence of curated IT-related parallel data.

9 Limitations

However, there are some limitations, which open avenues for future work.

The approach to ContexTerm relies on pretrained models to extract the encoders' final hidden states from the source language (or target language). While this makes practical use of existing models, it assumes that such a model exists but this may not be so, especially when there are low resource languages that are under-served.

The current system also does not consider post-processing by lexical constrained decoding. Its interaction with ContexTerm might result in overall translation quality and Terminology Success Rates.

The general translation quality is still not up to par with state-of-the-art neural machine translation engines, and this could be due to severe domain mismatches. As mentioned previously, carefully curated IT-related parallel data is under-resourced. Although CED was used to select domain-specific data from a large pool of domain-agnostic data, this approach may not actually capture actual human-curated IT-related data. This is especially true for German, whose general translation quality falls below 20.

More IT-related parallel data is needed and it would be interesting to see how the contextual approach would fair against NMT systems that have been trained on quality IT-related data.

References

Melissa Ailem, Jingshu Liu, and Raheel Qader. 2021. Lingua custodia's participation at the WMT 2021 machine translation using terminologies shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 799–803, Online. Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online. Association for Computational Linguistics.

Amittai Axelrod, Mei Yang, Kevin Duh, and Katrin Kirchhoff. 2008. The university of washington machine translation system for acl wmt 2008. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, page 123–126, USA. Association for Computational Linguistics.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

Lynne Bowker. 2021. Machine translation literacy instruction for non-translators: A comparison of five delivery formats. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 25–36, Held Online. INCOMA Ltd.

Rajen Chatterjee, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain. 2017. Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Second Conference on Machine Translation*, pages 157–168, Copenhagen, Denmark. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 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.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

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.

Jingshu Liu, Mariam Nakhlé, Gaëtan Caillout, and Raheel Qadar. 2023. Lingua custodia's participation at the WMT 2023 terminology shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 897–901, Singapore. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. *Preprint*, arXiv:1903.08788.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. *Preprint*, arXiv:1907.06616.

Tommi Nieminen. 2023. OPUS-CAT terminology systems for the WMT23 terminology shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 912–918, Singapore. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Geon Woo Park, Junghwa Lee, Meiying Ren, Allison Shindell, and Yeonsoo Lee. 2023. VARCO-MT: NCSOFT's WMT'23 terminology shared task submission. In *Proceedings of the Eighth Conference on Machine Translation*, pages 919–925, Singapore. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post and Marcin Junczys-Dowmunt. 2024. Evaluation and large-scale training for contextual machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1125–1139, Miami, Florida, USA. 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.

Adriaan M. J. Schakel and Benjamin J. Wilson. 2015. Measuring word significance using distributed representations of words. *Preprint*, arXiv:1508.02297.

Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. Findings of the WMT25 Terminology Translation Task: Terminology is Useful Especially for Good MTs. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 shared task on machine translation with terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. 2024. Importance-aware data augmentation for document-level neural machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 740–752, St. Julian's, Malta. Association for Computational Linguistics.

Yu Yuan, Yuze Gao, Yue Zhang, and Serge Sharoff. 2018. Cross-lingual terminology extraction for translation quality estimation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A NMT Training Configurations

Hyperparameters	**en-de**	**en-es**	**en-ru**
Encoder Embedding Size	512	512	512
Decoder Embedding Size	512	512	512
Encoder FFN Embedding Size	2048	2048	2048
Decoder FFN Embedding Size	2048	2048	2048
Encoder Layers	6	2	6
Decoder Layers	6	2	6
Encoder Attention Heads	16	16	16
Decoder Attention Heads	16	16	16
Learning Rate	3e-4	5e-4	3e-4
lr scheduler	Inverse Sqrt	Inverse Sqrt	Inverse Sqrt
Optimizer	Adam	Adam	Adam
Max Tokens	10000	10000	10000
Attention Dropout	0.0	0.0	0.0
Dropout	0.2	0.3	0.3
Criterion	Label-Smoothed CE	Label-Smoothed CE	Label-Smoothed CE
Warmup Steps	8000	6000	8000
Warmup init lr	1e-8	1e-8	1e-8
Clip Norm	1.0	1.0	1.0
Label Smoothing	0.1	0.1	0.1
Max source-target positions	4096	4096	4096

Table 4: Language-pair-specific hyperparameters that were used for NMT training