Coming to Terms with Glossary Enforcement: A Study of Three Approaches to Enforcing Terminology in NMT

Fred Bane **Anna Zaretskaya** **Tània Blanch Miró** **Celia Soler Uguet** **João Torres**
TransPerfect
{fbane,azaretskaya,tblanch,csuguet,joao.torres}@translations.com

Abstract

Enforcing terminology constraints is less straight-forward in neural machine translation (NMT) than statistical machine translation. Current methods, such as alignment-based insertion or the use of factors or special tokens, each have their strengths and drawbacks. We describe the current state of research on terminology enforcement in transformer-based NMT models, and present the results of our investigation into the performance of three different approaches. In addition to reference based quality metrics, we also evaluate the linguistic quality of the translations thus produced. Our results show that each approach is effective, though a negative impact on translation fluency remains evident.

1 Introduction

Ensuring translations use the preferred term can be business-critical for commercial translation providers. While there are existing methods to ensure the correct translation of specified terms, the impact of these methods on translation quality merits closer inspection. Typically, they have been evaluated in terms of general translation metrics such as BLEU, in addition to the accuracy of the terminology translation. However, there is a dearth of more detailed linguistic analysis of the performance of different techniques; for example, how often do the terms agree morphologically with the rest of the sentence? What are the potential issues when unruly, real-world, client glossaries are applied to models trained in more controlled laboratory conditions, and what steps can be taken to mitigate these issues?

In the present work we implement three approaches to glossary/terminology enforcement in two language pairs (English-Russian and Japanese-English) and compare their performance on the terminology enforcement task. In particular, we investigate two methods based on interventions in the training data and one post-processing method which uses the model's attention mechanism to identify the tokens representing the translation of the input term in the output and replaces these tokens with the translation from the glossary. In addition to automated evaluation (chrF, COMET, and accuracy), we also engaged professional linguists to design a test set of edge cases from their particular language pairs, and evaluate the performance of each approach using this bespoke test set.

The ultimate objective of this research is to inform the implementation of a glossary feature for use by machine translation project managers and end users, and thus we must anticipate that the feature will be applied in a multitude of unexpected ways. For a guide to what our feature may be subjected to, we turned to a database of historical glossary enforcement requests kept by our company. These requests were created by a mixture of linguists, clients, and project managers in translation projects. The contents of these glossaries are very noisy and diverse, including nouns, adjectives, verbs, prepositions, numbers, and acronyms, and ranging in length from single characters to entire sentences. This resource served both as the source material to annotate our training data for the methods using data intervention, and the inspiration for our test cases.

In addition to the practical motivation of our research, we hope to provide the MT community with an insight on the linguistic effects that each of these methods have on the translation output. Below we share our methodology and the results of our experiments.

2 Related Work

The first approaches to introducing terminology enforcement in NMT were quite limited in terms of handling languages with inflections. For example, in one approach, a special placeholder token was used to mask the term in the source sentence, and then replaced with the correct term after the translation (Crego et al., 2016). In the more sophisticated *alignment* method, one of the attention heads of the transformer is trained with statistical word alignments, and the output of this attention head at translation time is used to identify the tokens in the translation that correspond to the source term, and replace this token by the translation from the glossary. While this method provides an improvement, it still poses a problem for languages with inflections, since the target term is inserted in its glossary form, and dependencies may be produced in the wrong form.

In the *constrained decoding* method, the NMT decoder is guided to produce translation candidates that include the specified translation of a given source term that is present in the input sentence (Chatterjee et al., 2017; Hasler et al., 2018; Hokamp and Liu, 2017). This method, while certainly producing more fluent translations, adds a significant computational overhead (Post and Vilar, 2018). Since our applications of MT include several time-sensitive use cases, such as chat and instant website translation, we did not consider the constrained decoding method for our experiments.

Later, Dinu et al. (2019) proposed a method where intervention was made in the training data: they insert the target term directly in the source sentence and use factors to signal which tokens are actual source text and which are target translations. Factor embeddings are concatinated to the token embeddings and the two are learned in parallel. Through training, the model learns to essentially copy the input tokens marked as translations. More information on the practical implications of implementing this approach in a real-life production setting can be found in Exel et al. (2020) and Bergmanis & Pinnis (2021) address the application of this method to morphologically-rich languages.

Ailem et al. (2021) propose another approach to manipulate the training data: instead of using the source factors, they use special tokens to mark the source and target terms inserted in the source sentence. In addition, the authors apply token masking, which helps the model generalize better on unseen terms, and adapt the weighted cross-entropy loss to bias the model towards generating constraint tokens, resulting in improved translation quality and correctly generated constraint terms. This approach also accounts for different morphological variations of terms both on the source and on the target side by applying string-based approximate matching.

Until recently, most works only evaluated their results in terms of BLEU scores and accuracy of the terminology enforcement. However, they did not provide any insight into how well the term fits in the sentence, if the surrounding translations are correct, etc. For this reason, Alam et al. (2021a) proposed new metrics that can reflect correctness of terminology. In particular, they suggest to look at the tokens surrounding the term and compare them to the reference translation (*Window Overlap*) and to compute terminology-focused TER (Snover et al., 2006). These metrics are designed to complement the exact-match accuracy and the holistic MT quality metrics and were subsequently used in the first shared task dedicated to terminology in NMT (Alam et al., 2021b).

Since the experiments described above demonstrate that terminology constraints can be successfully applied in NMT without a significant overall performance loss and computational overhead, we choose two methods that are most suitable for our production settings, as well as a baseline method (replacing target tokens by the correct term translation based on the word alignments) to analyse each method's advantages. Our goal is not only to measure terminology accuracy and overall model performance, but also to get insight on how naturally the terms are incorporated into the target sentence.

3 Materials and Methods

We implemented three approaches to glossary enforcement: alignment-based replacement, annotation with special tokens as per Ailem et al. (2021), and factorization as per Dinu et al. (2019). As a control, we also obtain translation from a model trained with the same data without any terminology intervention.

3.1 Glossaries

Both the *annotation* and *factors* method rely on a glossary to prepare the training data. Glossaries can be compiled in multiple ways, such as using existent bilingual dictionaries, or learning dictionaries in an unsupervised manner. We chose to use data from historical translation projects as our glossaries, assuming that these may be the best approximation of the distribution of inputs our glossary feature will see in production.

As these data were extremely noisy, some filtering was required. We filtered terms containing no alphabetic, hiragana, katakana, or kanji characters, pairs with very unusual length ratios for the language pair (many terms contained lists of possible translations in the target field), pairs containing more than five whitespace-separated tokens, etc. For English-Russian, our database contained around 223k unique terminology pairs, of which 78k were retained after heuristic filtering. For Japanese English, the database contained approximately 240k unique pairs, of which 156k were retained after filtering. Many of these retained pairs were near duplicates, such as varying US/UK dialectical forms, pairs differing only in capitalization, or terms in their singular and plural forms. Of these terms, some 24k term pairs were actually found in the English-Russian training data, and 64k were found in the Japanese-English training data. We defer to later work a more in-depth investigation of the effects of different glossaries on model capabilities.

3.2 Data Resources

The training data were comprised of data from CC Matrix (Schwenk et al., 2019) and internal data resources, containing approximately 122 million sentence pairs for the English-Russian direction, and 56 million for the Japanese-English direction. The data were filtered with hand-crafted heuristics (for example very long or very short inputs, sentence pairs with unusual length ratios, sentence pairs with excessive punctuation or no detectable linguist content, etc.) and cross-entropy scores from an NMT model. For the *annotation* and *factors* methods, sentences from these corpora containing source and target glossary pairs included in our glossaries were identified and prepared as required for these techniques. The original versions of these sentences were retained in the corpora, to ensure that the model would still learn to translate these terms in the absence of guidance at inference time, and the modified versions were appended. Thus, the corpora increased in size by approximately 10 million and 7.7 million sentence pairs, respectively.

We elected to perform such modification only where the source and target terms appeared in exactly the same form as in the glossary, surrounded by word boundaries on either side for the English and Russian corpora (as Japanese does not separate words with white space, this constraint was not applicable for this language). Though lemmatization has been productively used to match other word forms not in the glossary – which appears to increase the ability of the model to adapt the term appropriately to the translation (Bergmanis and Pinnis, 2021) – we chose to use only exact matches for our benchmarking experiment to maximize the clarity of the training signal.

3.3 Training

Aside from the settings required for each approach, all models used identical standard transformer (base) configurations (Vaswani et al., 2017). We allowed models to train for 50 epochs or until perplexity failed to improve for ten consecutive validation checkpoints. Models were trained using the Marian framework (Junczys-Dowmunt et al., 2018) on eight Quadro RTX 6000 GPUs. Each model was trained twice and the best performing model was used for the experiment.

4 Evaluation

Human and automated evaluation methods were used to judge the performance of each approach. For the human evaluation, we worked with linguists to design test sets covering different morphological forms and specific edge cases identified for their languages. The morphological forms covered included adjectives, verbs, simple nouns in nominative, plural, and genitive forms, phrasal nouns and verbs, and entire clauses. For example, the ENRU test set contained, among regular nouns and noun phrases, terms like *men's*, *go back*, *turned off*. These terms are usually not recommended to be applied in the MT context, but they are often found in client glossaries, so we wanted to understand the behavior of different terminology enforcement methods in these scenarios. Among the edge cases tested were the Japanese elision of the subject and other cases where grammatical differences between the languages create ambiguity. In total, there were

27 terms in the ENRU test set and 26 terms in the JAEN test set. Once we had the test sets created, we requested native linguists in the target language to provide two different translations for each selected term. Then, we found sentences that contained the source terms amoung our internal datasets or asked the linguists to artificially create them. These sentences were used for the human evaluation.

During the human evaluation stage, evaluators were presented with translations of these sentences from the four different systems: the control system with no glossary enforcement, the system trained with the *annotation* approach, the system trained with the *factors* approach, and the system where the target term is inserted based on the alignments. For each source sentence, we first enforced the first translation of the term and then the second one.

The linguists were asked the following questions about each of the translations: (a) Is the term present in the translation? (b) Is the term in the correct grammatical form? (c) Are the grammatical dependencies on the term in the correct form? (d) Does the term assume a non-existent form? (e) Are there any duplicated words? (f) Rate the overall accuracy of the translation from 1 to 10. (g) Rate the overall fluency of the translation from 1 to 10. As the size of these bespoke test sets was necessarily quite small, the statistical significance of the results was not calculated and only the raw results are presented.

For the automated evaluation, we used publicly available corpora for comparability. For the English-Russian language pair, data from the WMT shared task on terminology enforcement were used. Due to the lack of a public corpus designed for terminology enforcement in the Japanese-English language pair, the Bilingual Corpus of Wikipedia's Kyoto Articles[1] and its accompanying lexicon were adapted. We selected terms without non-letter characters that were identified as organizations, proper names, or works of art using Spacy's NER function. Finally, we filtered both corpora to remove any sentences that did not contain terms to be enforced. For terms with multiple glossary translations, the form used in the reference translation was enforced.

Translations were scored with COMET and chrF, and the number of exact and fuzzy matches were counted. Exact match was defined as a 100% substring match with word boundaries on either side, and a fuzzy match was defined as at least 80% sub-string match. The threshold for statistical significance was established as p <0.01.

5 Results

5.1 Human Evaluation

The results of the human evaluation for each language pair are shown in Tables 1 and 2. We provide counts of each of the parameters we evaluated for each of the term translations (Term 1 and Term 2). The only exception is the *No glossary* approach, where we did not explicitly provide any instructions to the MT engine, so we provide cumulative numbers. We find it useful, however, to show which of the two term translations was preferred by the engine.

Overall, the *alignment* method had the best performance when it comes to including the term in the translation, which is expected by design. In the English-to-Russian language pair, this method also predictably was the worst when it comes to the morphological agreement (of the term itself and of the surrounding words). However, this was not the case for Japanese into English, where all the methods performed similarly well in this aspect. This suggests that this limitation of the *alignment* method may be more evident in morphologically rich target languages.

When the glossary term was a correct translation but not in the appropriate form for the sentence, the *annotation* and *factors* models sometimes modified the term into the appropriate form (examples of this are provided in Table 3 below and Table 7 in Appendix A), and sometimes modified the sentence structure in order to use the glossary form of the term in an appropriate way. In these cases, the *factors* approach was most likely to modify the term to an appropriate form, but the translations without glossary enforcement were judged to be of the best quality. The *alignment* method maintained the term exactly in its glossary form and often produced ungrammatical sentences in response to such inputs. Analysis of the evaluation results grouped by part of speech showed no clear pattern. Thus, we see no indication that any part of speech is more difficult than any other, nor that any approach more or less capable of applying the glossary constraints depending on their part of speech.

Other limitations of the *alignment* method were much more common in the Japanese-English language pair. Namely, we observed a higher number

[1] https://github.com/venali/BilingualCorpus

	No glossary		Annotation		Factors		Alignment	
	Term 1	Term 2	Term 1	Term 2	Term 1	Term 2	Term 1	Term 2
Term is present	14 (+1)	3 (+1)	23	20	23	13 (+2)	**24**	**23**
Correct form	19		**19**	**15**	17	12	10	11
Correct dependencies	19		**23**	**19**	21	15	18	12
Non-existent form	1		1	0	0	0	0	0
Duplicated words	0		0	0	0	0	2	2
Average accuracy	**9.4**		8.9	8.3	8.9	8.4	8.8	8
Average fluency	**9.6**		8.9	8.8	8.9	8.5	8.1	7.5

Table 1: English-Russian human evaluation results. When the term is present only partially (i.e. the term consists of multiple tokens and only one of them is present), its count is indicated in parentheses. The highest scores are marked in bold and are considered separately for terms 1 and 2. The total number of source sentences was 27.

	No glossary		Annotation		Factors		Alignment	
	Term 1	Term 2	Term 1	Term 2	Term 1	Term 2	Term 1	Term 2
Term is present	9 (+4)	3 (+1)	20 (+4)	20 (+4)	16 (+7)	16 (+6)	**24 (+2)**	**22 (+3)**
Correct form	17		**24**	22	21	21	23	**23**
Correct dependencies	17		**24**	22	21	21	23	**23**
Non-existent form	1		0	1	2	3	3	2
Duplicated words	0		0	0	0	0	1	1
Average accuracy	6.9		7.1	**8.8**	**7.6**	7.6	6.8	6.9
Average fluency	8.6		8.6	8.4	**9.1**	**9**	8.1	8.1

Table 2: Japanese-English human evaluation results. When the term is present only partially (i.e. the term consists of multiple tokens and only one of them is present), it is shown in parentheses. The highest scores are marked in bold and are considered separately for terms 1 and 2. The total number of source sentences was 26.

of non-existent grammatical form and duplicated words. The latter is typically due to the failure of the alignment mechanism in cases when a term corresponds to multiple target words, which may not be contiguous.

When it comes to the general translation quality, in the English-Russian language pair the model with no glossary enforcement achieved the best scores, even though its translation did not necessarily contain the required terms. Out of the three terminology enforcement methods, *annotation* and *factors* methods were the best with the *annotation* method slightly outperforming in fluency. The Japanese-English language pair paints a slightly different picture, with the *annotation* and *factors* models sharing the first positions in accuracy and fluency.

The results show significantly more partial matches in the Japanese-English language pair. Many of these correspond to terms that were verb phrases where a pronoun in the glossary translation was replaced by the subject of the sentence in the MT output (see examples in Table 6 in Appendix A).

Overall, based on the results of the human evaluation for English-Russian, it seems like the most optimal terminology approach is the *annotation* one. It has relatively good term accuracy as well as the general translation quality, and is the best in maintaining morphological agreement within the sentence. In the Japanese-English direction, morphological agreement plays a less significant role, so these results are more even across the different approaches. The *alignment* method has the highest term accuracy, but at the same time is more prone to producing errors such as duplicated words and non-existent forms. The *factors* method has the highest position in the overall translation quality but underperforms in terminology accuracy. The *annotation* method shows the most balanced scores overall.

5.2 Automated Evaluation

The results of the automated evaluation, shown in Table 4 below, are similar to the results of the human evaluation. The *factors* method obtained the best COMET and chrF scores in the Japenese-English direction, while in the English-Russian di-

Source	I’m going for a run.	I see him run.	Run!!!!!
No glossary	Я собираюсь а пробежку.	Я вижу, как он бежит.	Бегите!!!!!
Annotation	Я собираюсь бегать.	Я вижу, как он бегает.	Выполнить бегать!!!!!
Factors	Я иду на бегать.	Я вижу, как он бегает.	Бегать!!!
Alignment	Я еду на бегать.	Я вижу, как он бегать.	бегать!!!!

Table 3: Translations when the glossary form is a correct translation but not in the appropriate morphological form for the sentence. In this case, our glossary pair was ’run’: ’бегать’.

rection the *annotation* model showed the best performance. The *alignment* method achieved competitive results in all categories, and was clearly the most consistent in its adherence to the imposed glossary constraints. The performance of all models was quite poor on the Japanese-English automated test data, we speculate this is due to the significant domain gap between the training and test data. The English-Russian automated test data was COVID-related, and thus more in-domain, which we believe explains the superior performance in this language pair.

6 Discussion

Our results show that each method of enforcing terminology tested, which we have referred to in this paper as *alignment*, *annotation*, and *factors*, is effective in promoting the use of the requested translation. In both languages the approaches outperformed the baseline in this regard. The approaches did well in a wide variety of test cases, even test cases that may strain credulity. The benefit of giving this sort of guidance to the model seems to be more significant for input content that is out-of-domain for the training data, but this improvement in terminology use does little to mitigate the quality drop observed in such translation scenarios. The alignment method seemed to have a larger negative impact on translation quality, as measured by accuracy, fluency, and morphological agreement, but was also the most likely to have the correct term present in the sentence.

Additionally, our results show that the use of noisy source material for glossary creation is viable. Some intervention may still be required to retain only good quality term pairs. It remains to be seen how well this glossary actually approximates the distribution of input terms in production.

Contrary to the fears of Bergmanis and Pinnis (2021), using only exact matches in data preparation does not limit the model to simple copying behavior. However, a tendency to restructure the output sentence so as to properly use the exact term provided is noticeable. Users of glossary features should be guided on how best to work with polysemous terms in NMT.

None of the methods emerged as clearly superior, with different models performing better in different tasks and different language pairs. We believe that this suggests that each approach can be viable, but must be carefully adapted for the specific language pair and usage scenario. A solution combining the *annotation* or *factors* method with the *alignment* method may present a good option. In such a solution, input data would be prepared according to the requirements for the former method, and alignment-based insertion can be used as a fallback, when the model does not produce the expected term. The use of lemmatization in this fallback method may help reduce the incidence of false positives for cases where the model has used the term correctly but in a morphological form different to that of the glossary term.

7 Future Work

This research suggests multiple potential paths for future research. Firstly, our assumption that historical terminology enforcement requests approximate the distribution at inference time calls for proper scrutiny. Research comparing the effects of using different glossaries to prepare training data under controlled conditions can show if there is any significant downstream effect in the translation task.

Furthermore, there are many avenues of investigation stemming from the data preparation procedure. What is the appropriate ratio of samples with and without glossary enforcement signals in the dataset? What are the effects of lemmatization or fuzzy matching of glossary pairs in the dataset? What would be the effect of adding the glossary signal at the start or end of the sequence instead of at the location where the source term occurs? Should there be a limit to how many times a particular term appears? The frequency distribution

Model	chrF	COMET	Exact match %	Fuzzy match %
JAEN No glossary	33.2	-0.54	27.62	33.56
JAEN Annotation	35.1*	-0.44*	91.7*	94.24*
JAEN Factors	**36.1***	**-0.4***	90.36*	95.21*
JAEN Alignment	35.3*	-0.48*	**100***	**100***
ENRU No glossary	60.7	**0.7**	68.95	85.9
ENRU Annotation	**61.2***	**0.7**	76.19*	95.05*
ENRU Factors	60	0.65	68.17	88.38
ENRU Alignment	61.1*	0.62	**98.28***	**99.81***

Table 4: Automated evaluation metrics for the Japanese-English (JAEN) and English-Russian (ENRU) language pairs. The highest scores for each language pair are marked in bold, * indicates a statistically significant (p <0.01) improvement over the translation without glossary constraints.

of terms in our datasets showed roughly an inverse rank-frequency curve (Zipf's law), with some terms appearing with great frequency and a long tail of terms appearing only once.

Lastly, more research into interventions in the decoding algorithm is warranted. Techniques such as adaptive MT and constrained decoding, or some yet undiscovered technique may still prove to be superior to the methods investigated in this work. While progress thus far has been remarkable, the issue of terminology enforcement is far from solved, so close attention to new research is necessary.

References

Ailem, Melissa, Jingshu Liu, and Raheel Qader. 2021. Encouraging neural machine translation to satisfy terminology constraints. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1450–1455, Online, August. Association for Computational Linguistics.

Alam, Md Mahfuz ibn, Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, and Vassilina Nikoulina. 2021a. On the evaluation of machine translation for terminology consistency. arXiv.

Alam, Md Mahfuz ibn, Ivana Kvapilíko'a, Besacier Laurent Anastasopoulos, Antonios, Georgiana Dinu, Marcello Federico, Matthias Gallé, Philipp Koehn, Vassilina Nikoulina, and Kweon Woo Jung. 2021b. Findings of the wmt shared task on machine translation using terminologies. In *Proceedings of the 6th Conference on Machine Translation (WMT21)*, Online, November. Association for Computational Linguistics.

Bergmanis, Toms and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online, April. Association for Computational Linguistics.

Chatterjee, Rajen, 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, September. Association for Computational Linguistics.

Crego, Josep, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, and Peter Zoldan. 2016. Systran's pure neural machine translation systems. 10.

Dinu, Georgiana, 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, July. Association for Computational Linguistics.

Exel, Miriam, Bianka Buschbeck, Lauritz Brandt, and Simona Doneva. 2020. Terminology-constrained neural machine translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal, November. European Association for Machine Translation.

Hasler, Eva, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana, June. Association for Computational Linguistics.

Hokamp, Chris and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1535–1546, Vancouver, Canada, July. Association for Computational Linguistics.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Post, Matt and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana, June. Association for Computational Linguistics.

Schwenk, Holger, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the web.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12. Association for Machine Translation in the Americas.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Guyon, I., U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Appendix A. Supplementary Materials

Source sentence	あなたが許可を取り消した場合、あなたや赤ちゃんの身元を特定する情報を新たに収集することはありません。
Translation without glossary enforcement	If you withdraw your permission, no new information that identifies you or your baby **will be collected**.
Annotation 1	あなたが許可を取り消した場合、あなたや赤ちゃんの身元を特定する情報を新たに<S><C>we </C>収集することはありません。
Annotation 1 translation	If you withdraw your permission, **we** will not collect any new information that identifies you or your baby.
Annotation 2	あなたが許可を取り消した場合、あなたや赤ちゃんの身元を特定する情報を新たに<S><C>the research center </C>収集することはありません。
Annotation 2 translation	If you withdraw your permission, no new information identifying you or your baby will be collected by **the research center**.

Table 5: Example language-specific edge case. In the Japanese source, the subject is elided, as it may be inferred from context. Without glossary guidance, the model chooses a passive voice. With glossary guidance, an active voice can be induced. As no source term exists, we added the annotation with an empty source field where the subject would appear. Boldface for emphasis.

Source term	Target term	Source sentence	Target sentence (*annotation* method)
言い続けて	They keep saying	これは死亡が宣告された日から遺族がずっと言い続けてきたことだ。	This is because **the surviving family** has always **kept saying**, starting from the day the death was declared.
戻って来た	They have returned	市職員や住民、観光客らがそのうちの何頭かを引きずり、なんとか沖へ帰したものの、その多くが戻って来たという。	City officials, residents, and tourists dragged some of them, and they somehow returned to the offshore, but many of them said **they had returned**.

Table 6: Japanese-English examples of partial term matches. Boldface for emphasis.

Source term	Target term	Source sentence	Original translation	Annotation method
subject	пациент	One subject experienced an SAE (pneumonia) during study treatment with FSC.	У одного пациента развилось СНЯ (пневмония) во время исследуемого лечения КФС.	Один пациент перенес СНЯ (пневмонию) во время исследуемого лечения КФС.

Table 7: Sentence adaptation to match the glossary form of the term in English-Russian.