Term Translation: Convert or Converse?

Aida Kostikova
Bielefeld University
aida.kostikova@uni-bielefeld.de

Kristin Migdisi CrossLang
Sara Szoc CrossLang
Tom Vanallemeersch CrossLang
Franklin Rooseveltlaan 348/bus 8, 9000 Gent, Belgium
[first name].[last name]@crosslang.com

Abstract

A well-known challenge of machine translation (MT) is accurately translating domain-specific terminology. While various methods have been suggested to address this challenge, they all come with limitations and increase the user's dependence on a specific MT engine. Recently, large language models (LLMs) for various natural language processing tasks, including automated translation, have gained significant attention, urging the need to investigate the potential of these models for terminology translation. Therefore, we compare ChatGPT, an LLM-based chatbot conversing with a user, to DeepL, an MT system converting sequences to sequences. We use both systems to perform translations with and without glossaries. We also combine both systems by post-editing MT output with the chatbot. Automated and manual evaluations indicate that the global translation quality of MT is better than or on par with that of the chatbot with a glossary, but that the latter system excels in terms of terminological accuracy when used for translation or for post-editing. While such post-editing avoids user dependence on a specific MT engine, it sometimes causes new translation issues, such as shifts in meaning, suggesting the need for future improvements. Our experiments focus on two language pairs, English-Russian and English-French, and on two domains (COVID-19 and legal documents).

1 Introduction

One of the most persistent and complex challenges in machine translation (MT) is the accurate handling of domain-specific terminology. This terminology is often context-specific, making its translation intricate and time-consuming. This problem has been approached with various MT techniques striving to improve the accuracy of translated terms, by enforcing a specific training procedure. However, each of these strategies faces limitations.

Lately, there has been a growing trend towards training large language models (LLMs) for specific tasks. For instance, GPT (Generative Pre-trained Transformer), when specialised for a chatbot, can also be used to request a translation. Despite this trend, the potential of LLMs for terminology-aware translation remains relatively unexplored.

In this paper, we analyse the capabilities of a chatbot (more specifically ChatGPT 4), a system that *converses* (interacts) with a user, in handling the terminology problem. We compare it with an MT system (more specifically DeepL), which *converts* sequences to sequences. We also combine both systems, through a post-editing procedure. Our experiments involve two language pairs, English-Russian and English-French, and two domains, COVID-19 and legal documents.

The performance of the systems is evaluated using four scenarios. In the first one, an MT system is presented with English sentences containing challenging terms. In the second one, a chatbot is requested (prompted) to provide translations for these English sentences. In the third scenario, we provide the chatbot with a list of terms and their corresponding translations. The

last scenario, post-editing, involves providing the chatbot with source sentences, MT output, and a glossary.

We assess the outcome of the tasks both automatically and manually. In the first case, we measure the global translation quality of a sentence using automatic metrics. In the second, we focus on terminological accuracy by assigning a sentence to an error category, where applicable.

In the subsequent sections, we describe the background of our research, the methodology, the data, and the results. Finally, we describe potential future workflow based on our findings and the challenges encountered.

2 Background

Several methods exist for incorporating terminology into neural machine translation (NMT) systems:

- Mixing training data. This approach ensures the NMT training data contains both generic and domain-specific training data. While this allows the NMT engine to produce relevant terminology (e.g. to translate *bankruptcy* with French *faillite*), there is no guarantee of getting the right translation (e.g. *bankruptcy* may also be translated with French *banqueroute*).

- Incorporating placeholders. This approach makes use of non-terminal tokens in NMT systems (such as <*term#1*>), through pre- and postprocessing (Crego et al., 2016). While the NMT model learns how to deal with terminology, information is lost: the tokens constituting terms are no longer present during training of the model and thus inflection of target terms is not handled (Michon et al., 2020); this requires specific procedures before and after applying the NMT model.

- Constrained decoding. This approach ensures that the desired term translations appear in the NMT output (Hokamp and Liu, 2017). This is achieved at the cost of higher computational demands, which slows down the translation process.

- Injecting translations in the source sentence. The NMT model learns how to incorporate terminology translations in the target sentence when they are provided inside the source sentence (Song et al., 2019; Dinu et al., 2019). The system learns to copy words from time to time. This approach lacks the power to generalize, as the injection of the target term takes place without regard for the target context.

Recently, the use of LLMs for various natural language processing (NLP) tasks, including automated translation, has gained significant attention, urging the investigation of the potential of these models for terminology translation, especially in light of the above-mentioned drawbacks and the dependence of a user on a specific engine in terms of training data composition and training procedure.

NMT models and LLMs differ fundamentally in their training and architecture. NMT models are trained on parallel data and have an encoder-decoder architecture. In contrast, LLMs are trained on large amounts of monolingual data in one or more languages and employ a decoder-only architecture. This approach has opened new possibilities in multilingual NLP. LLMs are versatile, capable of being adapted for various tasks. For instance, they constitute the basis for chatbots such as ChatGPT, developed by OpenAI (Ouyang et al., 2022). Chatbots take user prompts and provide a response. Such prompts may also include labelled examples, which allows for in-context learning by the system (Brown et al., 2020).

The multilingual capabilities of LLMs are being investigated in comparison to NMT. For instance, Hendy et al. (2023) found that GPT models are very performant for high resource languages but have limited capabilities for low resource languages. In addition, Garcia et al.

(2023) show the usefulness of providing a limited number of example sentence pairs when using an LLM to translate.

Various translation prompts have been proposed (Jiao et al., 2023). These prompts can either include translation task information only, provide additional context domain information, or use part-of-speech tags as auxiliary information (Gao et al., 2023).

Recent developments in NMT, such as those explored by Moslem et al. (2023), have begun addressing the integration of terminology in automated translation. This study notably improves the incorporation of pre-approved terms in translations using a methodology that combines synthetic data generation and terminology-constrained post-editing with LLMs like ChatGPT. However, the accurate rendering of terminology during translation by chatbots and comparison with traditional NMT systems at this level was not the primary focus of these studies. Our research, on the other hand, is specifically aimed at examining the efficacy of a chatbot as compared to an NMT system. We investigate how chatbots, which can use terminological information as context in prompts, without requiring specialised training data compositions or procedures, may not only facilitate terminology handling but also potentially enhance the quality of translation, especially in terms of terminological accuracy. The following sections will describe our approach in conducting this comparative analysis, highlighting the distinct aspects of our methodology.

3 Methodology

We compare the output of an NMT system (DeepL) to that of a chatbot (ChatGPT 4). We restrict the scope of our investigation to these two state-of-the-art systems, leaving the investigation of other systems (for instance open-source software) for future investigation. The NMT system and the chatbot are compared as follows. Based on a translation memory or parallel corpus in a specific domain, we select challenging source terms, select an illustrative sample of source sentences containing these terms, and apply the following scenarios to these sentences:

1. Translate the sentences using NMT.
2. Request the chatbot to translate the sentences. Prompts with the structure shown in Figure 1 are entered (we include multiple sentences in the prompt to provide more context).
3. Apply the same procedure as in 2 but include a glossary in the prompt, as shown in Figure 2. This is a form of in-context learning, as opposed to the zero-shot learning in scenario 2.
4. Given the sentences which, based on the human evaluation procedure described below, are known to be translated incorrectly at the terminological level by the NMT system, provide a prompt to the chatbot requesting it to post-edit the NMT output based on the glossary, as shown in Figure 3. The prompt also includes the source sentence.

Translate these sentences from English into Russian:

U.S. older adults, including those aged ≥65 years and particularly those aged ≥85 years, also appear to be at higher risk for severe COVID-19-associated outcomes; however, data describing underlying health conditions among U.S. COVID-19 patients have not yet been reported.

In the EU/EEA, the first three confirmed cases were reported by France on 24 January 2020 in persons returning from Wuhan, Hubei Province, China.

Figure 1. Prompt for scenario 2 (translation using chatbot)

Here is the list of COVID-19-related terms in English and their equivalents in Russian:

Respiratory distress syndrome – Острый респираторный дистресс-синдром
Respiratory dysfunction – дыхательная дисфункция

Given this glossary, translate the sentences provided below from English into Russian. Make sure that the terminology translation fully adheres to the glossary I provided, the translation domain is Covid-19.

Figure 2. Prompt for scenario 3 (translation using chatbot + glossary)

Please post-edit the following Russian sentences translated from English. The English source text is provided for your reference. The sentences are related to COVID-19, and I have noticed that the terminology used in the translations may not be accurate. Your task is to edit the sentences, replacing any incorrect or inadequate medical terms with the appropriate ones. You can refer to the list of COVID-19 terms provided below for guidance.

---------- List of COVID-19 Terms (English to Russian) ----------

…

English source:

…

Russian translation:

…

Please review each sentence carefully and make any necessary changes to ensure accurate and appropriate COVID-19 terminology is used.

Figure 3. Prompt for scenario 4 (post-edition of NMT output using chatbot + glossary)

We perform an automated evaluation of the global translation quality of the first three scenarios using the metrics BLEU, chr_f, TER and BERTScore. The first two of these calculate an n-gram match between output and reference (in terms of tokens or characters). The third one calculates post-editing effort and the fourth performs a semantic comparison of sentences using deep learning (embeddings).

We manually evaluate the translation output for all four scenarios (i.e. also the scenario for post-editing) at the terminological level, assigning one of the following error types to a sentence if applicable:

- Inaccurate translation: the translated term (i) does not precisely match the original term's meaning despite maintaining the general sense, (ii) is misleading, or (iii) is unrelated to the source term.
- Literal translation: the translated term matches the original term's meaning but has a different, unusual phrasing.
- Loss of elements: the system omits vital components of the source term in the translation.

4 Data

We apply the above methodology to two domains: COVID-19 (English-Russian) and legal-domain terminology (English-Russian and English-French).

For the first domain, we select a translation memory (TM) from the TICO-19 repository[1] and identify 49 challenging (that is, ambiguous, idiomatic, or culturally specific) terms in the English source sentences via SketchEngine.[2] We extract 90 sentences containing on average one or two of these terms from the TM, along with their Russian translations, to serve as reference translations.

For the second domain, we select 27 terms from the Rules of Court of the European Court of Human Rights, as well as from the European Convention on Human Rights, for both the English-Russian and English-French translation directions. In this case, we only perform (1) translation with the chatbot providing the glossary and (2) post-editing. We omit the use of the chatbot without the glossary, as the findings for the first domain, described in Section 5, clearly indicate a lower performance when working without a glossary.

5 Results

5.1 Automated Metrics Analysis

Table 1 shows the automated metric scores for the translation task involving COVID-19 terminology.

EN-RU COVID-19 (90 sentences)					
Metric	NMT	Chatbot	Delta with respect to NMT	Chatbot + glossary	Delta with respect to NMT
BLEU ↑	32.9	24.9	-8	29.0	-3.9
chr_f ↑	60.3	53.9	-6.4	59.8	-0.5
TER ↓	56.0	64.2	+8.2	59.0	+3
BERTScore ↑	88.2	86.1	-2.1	88.0	-0.2

Table 1. Automated metric scores for COVID-19

The results show that the NMT system outperforms the chatbot when the latter is not provided with a glossary, but that the gap shrinks when the glossary is provided. This demonstrates the efficacy of added contextual support.

Interestingly, in the legal domain (for which the chatbot is not tested without a glossary, as mentioned earlier), the chatbot with a glossary not only closes the gap with NMT but slightly outperforms it across all the evaluation metrics for both English-Russian and English-French translations (see Table 2). This indicates a pronounced effectiveness of the chatbot in handling domain-specific terminology when provided with a glossary.

(57 sentences)	EN-RU legal			EN-FR legal		
Metric	NMT	Chatbot + glossary	Delta	NMT	Chatbot + glossary	Delta
BLEU ↑	33.3	33.8	+0.5	45.2	45.6	+0.4
chr_f ↑	61.2	62.0	+0.8	69.8	71.0	+1.2
TER ↓	55.4	55.0	-0.4	41.6	40.7	-0.9
BERTScore ↑	87.1	87.8	+0.7	89.6	90.5	+0.9

Table 2. Automated metric scores for the legal domain

5.2 Human Evaluation Results

The human evaluation focuses on assessing the accuracy of terminology in the translations.[3] For the COVID-19 dataset, the use of the chatbot with a glossary markedly reduces

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

[2] https://www.sketchengine.eu

[3] The evaluation was performed by the authors. Potential future improvements consist of interannotator agreement and the involvement of domain experts.

terminological errors compared to NMT and chatbot without a glossary, as can be seen in Table 3.

EN-RU COVID-19 (90 sentences)					
Translation error type	NMT	Chatbot	Delta with respect to NMT	Chatbot + glossary	Delta with respect to NMT
Literal translation	16 (18%)	23 (26%)	+7 (8%)	0	-16 (18%)
Inaccurate translation	30 (33%)	46 (51%)	+16 (18%)	5 (5%)	-25 (28%)
Loss of elements	4 (4%)	5 (5%)	+1 (1%)	5 (5%)	+1 (1%)
All	50 (55%)	74 (82%)	+24 (27%)	10 (11%)	-40 (44%)

Table 3. Distribution of most common terminological error types in translations, COVID-19

Inaccurate translations drop dramatically from 33% with NMT and 51% with the chatbot without a glossary to just 5% with the chatbot plus a glossary. In the example below, the chatbot without a glossary translates *shortness of breath* as *одышка* instead of *затруднение дыхания* (the first Russian translation being a less severe term implying temporary breathlessness). The chatbot, when provided with the term list, provides the correct translation:

- Source text: *Common symptoms include fever, cough and **shortness of breath**.*
- Glossary: *shortness of breath → затрудненное дыхание*
- NMT output: *Общие симптомы включают лихорадку, кашель и **одышку**. (literally: Common symptoms include fever, cough and **dyspnea**.)*
- Translation by chatbot without glossary: *Распространенные симптомы включают лихорадку, кашель и **одышку**. (literally: Widespread symptoms include fever, cough and **dyspnea**.)*
- Translation by chatbot with glossary: *Общие симптомы включают лихорадку, кашель и **затруднение дыхания**. (literally: Common symptoms include fever, cough and **shortness of breath**.)*

Similarly, literal translation errors are reduced to 0% with the chatbot plus a glossary, from 18% with NMT and 26% with the chatbot without a glossary. For instance, when translating the sentence provided below, the chatbot with a glossary chooses the more contextually appropriate term *самоизоляция*:

- Source text: *... try to stay indoors for **self-quarantine** and limit contact with potentially infected individuals.*
- Glossary: *self-quarantine → самоизоляция*
- NMT output: *... помещениях для **самокарантина** и ограничить контакты с потенциально инфицированными людьми. (literal translation: term not commonly used)*
- Translation by chatbot with glossary: *... помещении для **самоизоляции** и ограничивать контакт с потенциально инфицированными лицами. (more appropriate term in this context)*

A similar pattern can be observed in the legal domain, where the chatbot with a glossary outperforms NMT in terms of terminological correctness (see Table 4), especially for English-French: errors of all types present in the NMT output are absent from chatbot output. Examples of sentences in the legal domain for both language pairs and their NMT and chatbot output are shown in Appendix A.

(57 sentences)	EN-RU legal			EN-FR legal		
Translation error type	Scenario 1: NMT	Scenario 3: chatbot+glossary	Delta	Scenario 1: NMT	Scenario 3: chatbot+glossary	Delta
Literal translation	3 (5%)	4 (7%)	+1 (2%)	16 (28%)	0	-16 (28%)
Inaccurate translation	21 (37%)	16 (28%)	-5 (9%)	5 (9%)	0	-5 (9%)
Loss of elements	1 (2%)	2 (4%)	+1 (2%)	0	0	0
Total	25 (44%)	22 (39%)	-3 (5%)	21 (37%)	0	-21 (37%)

Table 4. Distribution of most common error types in translations, legal domain

5.3 Evaluation of Post-editing Task

We select all problematic NMT translations (50 sentences for the COVID-19 domain; 25 and 21 sentences for the legal texts for the English-Russian and English-French translation directions, respectively) and prompt the chatbot to post-edit them using the prompt specified in Figure 3. Table 5 shows the results for the post-editing task in both domains.

	EN-RU COVID-19	EN-RU legal	EN-FR legal
Sentences in NMT output containing error	50 (55% of 90)	25 (44% of 57)	21 (37% of 57)
Sentences with desired correction by chatbot	→ 47 (94%)	→ 15 (60%)	→ 20 (95%)

Table 5. Post-editing task results for COVID-19 and legal domains

For COVID-19, the chatbot successfully post-edits the terms in 47 sentences. An exemplary case of successful post-editing in this domain involves the term *transmissibility* in the following sentence:

- Source text: *On one hand, the **transmissibility** of SARS-CoV-2 is at least as high as that of community-acquired HCoVs.*
- Glossary: *transmissibility → передаваемость*
- NMT output: *С одной стороны, **трансмиссивность** SARS-CoV-2, по крайней мере, так же высока, как и у HCoV, передающихся через сообщества. (literally: On one hand, the **transferability** of SARS-CoV-2 is at least as high as that of community-acquired HCoVs.)*
- Translation by chatbot with glossary: *С одной стороны, **передаваемость** SARS-CoV-2, по крайней мере, так же высока, как и у HCoV, передающихся через сообщества. (literally: On one hand, the **transmissibility** of SARS-CoV-2 is at least as high as that of community-acquired HCoVs.)*

The NMT system translates the term as *трансмиссивность*, which is then correctly post-edited by the chatbot to *передаваемость*, adhering to the glossary's guidance.

The chatbot is able to handle inflections appropriately, ensuring that the translations are not only terminologically accurate but also grammatically coherent. In the example below, the chatbot not only replaces the term with the more accurate term *вспомогательных белков* from the glossary but also correctly adjusts the inflection to match the plural form used in the sentence:

- Source text: *A number of lineage-specific **accessory proteins** are also encoded by different lineages of CoVs.*
- Glossary: *accessory protein→ вспомогательный белок*

- NMT output: *Различные линии CoVs также кодируют ряд специфических **белков-аксессуаров**. (literally: Different lines of CoVs also code a number of specific **proteins-accessories**.)*
- Post-edited with chatbot: *Различные линии CoVs также кодируют ряд специфических **вспомогательных белков**. (literally: Different lines of CoVs also code a number of specific **accessory proteins**.)*

The errors in the three incorrectly post-edited sentences involve an unchanged term, terms where information is lost, and a term where information is added. Apart from changes at the terminological level, the chatbot occasionally has agreement errors, loses important sentence elements, or changes sentence meaning, as in the following text, where it mistranslates *HCoV* as *hepatitis C*:

- Source text: *It is also of particularly great interest to see whether SARS-CoV-2 might exhibit seasonality as in the cases of community-acquired **HCoVs**.*
- Post-edited with chatbot: *Особый интерес представляет также вопрос о том, может ли SARS-CoV-2 проявлять сезонность, как в случае с внебольничным вирусом **гепатита С**. (literally: It is also of particularly great interest to see whether SARS-CoV-2 might exhibit seasonality as in the cases of community-acquired **hepatitis C**.)*

Moreover, as can be seen from Table 5, the success rate of post-editing varies across domains. More specifically, the lowest success rate is observed in the English-Russian legal domain, where the chatbot successfully post-edits 60% of the sentences. In contrast, the chatbot achieves its highest success rate in the English-French legal domain, successfully correcting errors in 95% of the sentences, with only one error remaining.

Appendix A lists automatically post-edited sentences in the legal domain for both language pairs. Appendix B shows automatic metrics and confidence intervals for the various types of output (NMT, chatbot with glossary, chatbot post-editing) for this domain and these language pairs.

6 Conclusions and Future Work

In order to improve the performance of NMT engines in the area of term translation, various strategies have been developed for training data composition and training setup. As these requirements can lead to a dependence of the user on a specific engine, we explored to what extent the multilingual capabilities of a chatbot, a system which can be provided with various prompts in a user-friendly way, are useful for improving terminological accuracy. To this end, we compared the output of DeepL, an NMT engine, to that of a chatbot, ChatGPT.

Our comparative analysis was structured through four distinct scenarios, designed to evaluate translation outputs in terms of global translation quality and terminological accuracy: (i) translation using NMT without any additional input or modification; (ii) translation by the chatbot without the aid of a glossary; (iii) translation by the chatbot with the help of a glossary; (iv) post-editing of NMT output using the chatbot. To assess the performance across these scenarios, we employed both automated evaluation metrics and human evaluation methods.

On the one hand, NMT offers better translation quality (in the COVID-19 domain for English-Russian) or on-par output (for the legal domain in English-Russian and English-French) compared to the chatbot with a glossary. On the other hand, the latter system excels in terminological accuracy when requested to translate or to post-edit NMT output; this is especially the case for the COVID-19 domain and for the English-French legal-domain text. However, post-editing also carries the risk of introducing noise.

Our findings suggest opportunities for further research and development:

- To reduce the risk of noise introduction, we could vary prompt phrasing, the number of sentences included in a prompt, and the type and size of context. For instance, we may vary the number of glossary terms and include example sentence pairs for terms with multiple translations in a domain.
- Given the variety of NMT systems and chatbots (variants of ChatGPT, open-source models, etc.) and the fast-paced evolution in the area of chatbots, we could extend the scope of the study by increasing the number of systems, as well as the number of domains, languages, and metrics (e.g. additional deep learning metrics). It is also important to investigate whether and when "hallucinations" (nonsensical output) appear, which are regularly observed in complex deep learning based systems.
- While our study aimed to reduce dependence on specific MT engines, it would be interesting to compare the performance of a chatbot-based approach to that of uploading a user-specified glossary in a commercial system like DeepL.

Acknowledgements

This work is the result of an internship of the main author at CrossLang in the first half of 2023. She would like to express her gratitude for having received the opportunity to investigate the topic described in this paper.

References

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-shot Learners. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1877–1901.

Crego, Josep, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, et al. 2016. Systran's Pure Neural Machine Translation Systems. arXiv:1610.05540.

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.

Gao Yuan, Ruili Wang, and Feng Hou. 2023. How to Design Translation Prompts for ChatGPT: An Empirical Study. arXiv:2304.02182.

Garcia, Xavier, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The Unreasonable Effectiveness of Few-shot Learning for Machine Translation. In *Proceedings of the International Conference on Machine Learning*, pages 10867–10878.

Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good are GPT models at Machine Translation? A Comprehensive Evaluation. arXiv:2302.09210.

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.

Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is ChatGPT a Good Translator? Yes with GPT-4 as the engine. arXiv:2301.08745.

Michon, Elise, Josep Crego, and Jean Senellart. 2020. Integrating Domain Terminology into Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937.

Moslem, Yasmin, Gianfranco Romani, Mahdi Molaei, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Domain Terminology Integration into Machine Translation: Leveraging Large Language Models. arXiv:2310.14451.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Proceedings of Advances in Neural Information Processing Systems*, pages 27730–27744.

Song, Kai, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 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.

Appendix A: Examples for EN-RU and EN-FR (Legal Documents)

EN-RU :

Glossary	Source	Chatbot translation	NMT	Chatbot post-editing
Removal – Выдворение	The Court must be notified of any change to those details (date and time of **removal**, address etc.) as soon as possible.	Суд должен быть немедленно уведомлен обо всех изменениях этих данных (дата и время **выдворения**, адрес и т.д.). (literally: The Court must be notified of any change to those details (date and time of **removal**, address etc.) as soon as possible.)	Суд должен быть уведомлен о любых изменениях в этих данных (дата и время **удаления**, адрес и т.д.) в кратчайшие сроки. (literally: The Court must be notified of any change to those details (date and time of **deletion**, address etc.) as soon as possible.)	Суд должен быть уведомлен о любых изменениях в этих данных (дата и время **выдворения**, адрес и т.д.) как можно скорее. (literally: The Court must be notified of any change to those details (date and time of **removal**, address etc.) as soon as possible.)
Amendment – Поправка	Any proposal for **amendment** shall be communicated by the Secretary General of the Council of Europe to the Parties.	Любое предложение о **поправке** должно быть сообщено Генеральным секретарем Совета Европы Сторонам. (literally: Any proposal for **amendment** shall be communicated by the Secretary General of the Council of Europe to the Parties.)	Любое предложение о внесении **изменений** доводится Генеральным секретарем Совета Европы до сведения Сторон. (literally: Any proposal for **change** shall be communicated by the Secretary General of the Council of Europe to the Parties.)	Любое предложение о **Поправке** доводится Генеральным секретарем Совета Европы до сведения Сторон. (literally: Any proposal for **Amendment** shall be communicated by the Secretary General of the Council of Europe to the Parties.)
Judgment – постановление	The final **judgment** shall be published.	Окончательное **суждение** будет опубликовано. (literally: The final **opinion** shall be published.)	Окончательное **решение** должно быть опубликовано. (literally: The final **decision** shall be published.)	Окончательное **постановление** должно быть опубликовано. (literally: The final **judgment** shall be published.)
Application – постановление	The Court shall reject any **application** which it considers inadmissible under this Article.	Суд отклонит любую **жалобу**, которую он считает недопустимой в соответствии с этой статьёй. (literally: The Court shall reject any **complaint** which it considers inadmissible under this article.)	Суд отклоняет любое **заявление**, которое он считает неприемлемым в соответствии с настоящей статьей. (literally: The Court shall reject any **application** which it considers inadmissible under this article.)	Суд отклоняет любое **заявление**, которое он считает неприемлемым в соответствии с настоящей статьей. (literally: The Court shall reject any **application** which it considers inadmissible under this article.)

Glossary (continued)	**Source**	**Chatbot translation**	**NMT**	**Chatbot post-editing**
Domestic court – внутригосударственный Суд	The Court does not hear appeals against decisions of **domestic courts**, and applicants in expulsion or extradition cases should pursue **domestic remedies** [...]	Суд не рассматривает апелляции по решениям **национальных судов**, и заявители в делах об исключении или экстрадиции должны искать **национальные средства правовой защиты** [...] (literally: The Court does not hear appeals against decisions of **national courts**, and applicants in exclusion or extradition cases should pursue **national means of legal protection** [...])	Суд не рассматривает апелляции на решения **национальных судов**, и заявители по делам о высылке или экстрадиции должны использовать **внутренние средства правовой защиты** [...] (literally: The Court does not hear appeals against decisions of **national courts**, and applicants in expulsion or extradition cases should pursue **internal means of legal protection** [...])	Суд не рассматривает апелляции на решения **национальных судов**, и заявители по делам о высылке иностранцев или экстрадиции должны использовать Внутригосударственные Средства Защиты [...] (literally: The Court does not hear appeals against decisions of **national courts**, and applicants in expulsion or extradition cases should pursue **domestic means of legal protection** [...])

EN-FR :

Glossary	**Source**	**Chatbot translation**	**NMT**	**Chatbot post-editing**
Applicant – réquerant	In all cases before a Chamber or the Grand Chamber, a High Contracting Party one of whose nationals is an **applicant** shall have the right to submit written comments and to take part in hearings.	Dans tous les cas devant une Chambre ou la Grande Chambre, une Haute Partie Contractante dont un national est **réquerant** a le droit de soumettre des observations écrites et de participer aux audiences. (literally: In all cases before a Chamber or the Grand Chamber, a High Contracting Party of which a national is an **applicant** has the right to submit written comments and to take part in hearings.)	Dans toutes les affaires portées devant une chambre ou la Grande Chambre, une Haute Partie contractante dont l'un des ressortissants est **demandeur** a le droit de présenter des observations écrites et de participer aux audiences. (literally : In all cases brought before a Chamber or the Grand Chamber, a High contracting Party of which one of the nationals is **applicant** has the right to present written comments and to take part in hearings.)	Dans toutes les affaires devant une chambre ou la Grande Chambre, une Haute Partie contractante dont l'un des requérants est **ressortissant** a le droit de soumettre des observations écrites et de participer aux audiences. (literally : In all cases before a Chamber or the Grand Chamber, a High contracting Party of which one of the applicants is **national** has the right to submit written comments and to take part in hearings.)

Glossary (continued)	**Source**	**Chatbot translation**	**NMT**	**Chatbot post-editing**
Removal – renvoi	The Court must be notified of any change to those details (date and time of **removal**, address etc.) as soon as possible.	Le tribunal doit être informé de tout changement de ces détails (date et heure de **renvoi**, adresse, etc.) dès que possible. (literally : The Court must be informed of any change to those details (date and hour of **removal**, address etc.) as soon as possible.)	Toute modification de ces données (date et heure du **déménagement**, adresse, etc.) doit être notifiée à la Cour dans les plus brefs délais. (literally : Each modification of these data (date and hour of the **relocation**, address etc.) must be notified to the Court in the shortest delays.)	Toute modification de ces informations (date et heure du **renvoi**, adresse, etc.) doit être notifiée à la Cour dans les plus brefs délais. (literally : Each modification of this information (date and hour of the **removal**, address etc.) must be notified to the Court in the shortest delays.)
Lawful detention – détention régulière	the **lawful detention** of persons for the prevention of the spreading of infectious diseases, of persons of unsound mind, alcoholics or drug addicts or vagrants;	la **détention régulière** de personnes pour la prévention de la propagation de maladies infectieuses, de personnes d'esprit dérangé, d'alcooliques ou de toxicomanes ou de vagabonds; (literally : the **lawful detention** of persons for the prevention of the spreading of infectious diseases, of persons of disturbed mind, of alcoholics or of drug addicts or of vagrants;)	la **détention légale** de personnes en vue de prévenir la propagation de maladies infectieuses, d'aliénés, d'alcooliques, de toxicomanes ou de vagabonds ; (literally : the **lawful detention** of persons in order to prevent the spreading of infectious diseases, of lunatics, of alcoholics, of drug addicts or of vagrants;)	la **détention régulière** de personnes pour prévenir la propagation de maladies infectieuses, de personnes d'esprit dérangé, d'alcooliques, de toxicomanes ou de vagabonds ; (literally : the **lawful detention** of persons for preventing the spreading of infectious diseases, of persons of disturbed mind, of alcoholics, of drug addicts or of vagrants;)
Written pleadings – observations écrites	Rule 38 – **Written pleadings**	Règle 38 – **Observations écrites** (literally : Rule 38 – **Written comments**)	Article 38 - **Plaidoiries écrites** (literally : Article 38 – **Written pleadings**)	Règle 38 - **Observations écrites** (literally : Rule 38 – **Written comments**)
Expenditure on the court – frais de fonctionnement de la Cour	Article 50 – **Expenditure on the Court** The **expenditure on the Court** shall be borne by the Council of Europe.	Article 50 – **Frais de fonctionnement de la Cour** Les **frais de fonctionnement de la Cour** sont à la charge du Conseil de l'Europe. (literally : Article 50 – **Costs for functioning of the Court** The **costs for functioning of the Court** should be borne by the Council of Europe.)	Article 50 - **Dépenses de la Cour** Les **dépenses de la Cour** sont supportées par le Conseil de l'Europe. (literally : Article 50 – **Expenses of the Court** The **expenses of the Court** are supported by the Council of Europe.)	Article 50 - **Frais de fonctionnement de la Cour** Les **frais de fonctionnement de la Cour** sont supportés par le Conseil de l'Europe. (literally : Article 50 – **Costs for functioning of the Court** The **costs for functioning of the Court** are supported by the Council of Europe.)

Appendix B: Automatic Metrics for EN-RU and EN-FR (Legal Documents)

EN-RU:

System	bertscore (μ ± 95% CI)	BLEU (μ ± 95% CI)	chrF2 (μ ± 95% CI)	TER (μ ± 95% CI)
NMT	87.3 (87.3 ± 1.5)	33.6 (33.8 ± 5.7)	61.4 (61.5 ± 3.3)	54.9 (54.7 ± 6.1)
Chatbot translation with glossary	85.3 (85.3 ± 1.3) (p = 0.0020)*	25.8 (25.9 ± 4.5) (p = 0.0010)*	55.4 (55.4 ± 3.1) (p = 0.0010)*	63.2 (63.1 ± 5.5) (p = 0.0010)*
Chatbot post-editing	87.8 (87.8 ± 1.4) (p = 0.1179)	33.8 (34.0 ± 5.9) (p = 0.3117)	62.0 (62.1 ± 3.0) (p = 0.1988)	55.0 (54.8 ± 5.9) (p = 0.3796)

EN-FR:

System	bertscore (μ ± 95% CI)	BLEU (μ ± 95% CI)	chrF2 (μ ± 95% CI)	TER (μ ± 95% CI)
NMT	89.6 (89.6 ± 1.7)	89.6 (89.6 ± 1.7)	89.6 (89.6 ± 1.7)	89.6 (89.6 ± 1.7)
Chatbot translation with glossary	45.2 (45.1 ± 5.3)	45.2 (45.1 ± 5.3)	45.2 (45.1 ± 5.3)	45.2 (45.1 ± 5.3)
Chatbot post-editing	69.8 (69.8 ± 3.4)	69.8 (69.8 ± 3.4)	69.8 (69.8 ± 3.4)	69.8 (69.8 ± 3.4)