УДК 81'322.4 https://www.doi.org/10.33910/2686-830X-2022-4-1-25-30

Pre-editing English news texts for machine translation into Russian

E. S. Kokanova ✉[1], M. V. Berendyaev [1], N. Yu. Kulikov [1]

[1] Northern (Arctic) Federal University, 17 Severnaya Dvina Emb., Arkhangelsk 163002, Russia

Authors
Elena S. Kokanova,
SPIN: 9687-2303;
Scopus AuthorID: 57218198923;
ResearcherID: ABG-9970-2020;
ORCID: 0000-0001-6623-5636;
e-mail: e.s.kokanova@narfu.ru

Maxim V. Berendyaev,
SPIN: 4399-6155;
e-mail: m.berendyaev@narfu.ru

Nikolay Yu. Kulikov,
SPIN: 4461-3126;
e-mail: n.kulikov@narfu.ru

For citation: Kokanova, E. S., Berendyaev, M. V., Kulikov, N. Yu. (2022) Pre-editing English news texts for machine translation into Russian. *Language Studies and Modern Humanities*, vol. 4, no. 1, pp. 25–30. https://www.doi.org/10.33910/2686-830X-2022-4-1-25-30

Received 31 January 2022; reviewed 25 February 2022; accepted 25 February 2022.

Funding: The study did not receive any external funding.

Abstract. The paper discusses the possible advantages of pre-editing English news texts for machine translation into Russian. Pre-editing is defined as a process of adapting source text in order to reach a better quality of machine translation. Two case studies were conducted: one in 2021 and the other one in 2022. During case studies texts from *bbc.com* were chosen, pre-edited and translated using neural machine translation systems. Analysing common pre-editing techniques and their impact on the result of machine translation in terms of certain error patterns we conclude that in most cases pre-editing helps to eliminate a number of errors, improve the overall quality of machine output and reduce the amount of time and efforts needed for post-editing machine translation. The conducted case study also showed that despite the fact, that machine translation systems are constantly developing and changing, it is possible to identify common peculiarities of machine translation regarding a certain style or type of text and certain language pair, analyse the error patterns and find the appropriate pre-editing techniques that will be applicable for the most of machine translation systems for many months and years. Pre-editing does not guarantee the high quality of the translation itself, but together with post-editing it allows reaching results similar or equal to human translation saving a translator's time and efforts.

Keywords: pre-editing, neural machine translation, news text, English, Russian

Introduction

The advantages of machine translation (MT) over human translation (e.g., translation from scratch) make it extremely attractive: high translation speed, high performance and economic benefits. There is another important point for the discussion. Is it possible to claim that there is a great potential in the field of MT? Historically, the human participation in the MT process was considered as the impossibility of using MT on a large scale (Barkhudarov, Kolshanskij 1958), while others allowed the possibility of editing MT to the necessary extent (Hays 1960). There is an active development of MT systems, in particular, the MT technology based on neural networks. Before neural machine translation (NMT) appeared, the main challenge causing much discussion was poor translation quality, which could lead to misinterpretations and misdiagnoses (Dew, Turner, Choi et al. 2018).

Currently, it is accepted that pre-editing is the analysis and verification of the source text (ST) in order to identify possible errors for NMT and their elimination or correction (Marzouk, Hansen-Schirra 2019, 186). The research shows the influence of pre-editing on the quality of translation performed by NMT. Pre-editing includes forecasting problem areas and possible errors. Pre-editing texts for NMT is the process of preparing the source text (ST) before translating. The paper considers the peculiarities of translating news texts from English into Russian using NMT systems and the possibility of improving the quality of output data with pre-editing of ST. Globally, the use of MT combined with pre-editing has become integrated in the translation industry and in translator training. There are not only translators and post-editors, but also pre-editors.

Methods and material

The two news texts from bbc.com with a total number of words 300 were chosen for the case study. The source text 1 is "Reaching 130 million girls with no access to school" (Coughlan 2020). The source text 2 is "Why so many young Swedes live alone" (Savage 2019). First, the texts were pre-edited. Second, the source texts were translated from English into Russian before and after pre-editing using different MT systems based on neural networks. They are Google Translate, Amazon Translate and PROMT Professional Neural. After that, there was a comparison of six target texts.

In order to evaluate the quality of MT and the effectiveness of pre-editing, translation errors were identified both in source and pre-edited texts MT outputs. All errors were divided into following groups: 1) lexical errors, including mistranslations, omissions / additions, wrong word choice, errors of collocation, etc. and 2) grammatical errors, including wrong word order, incorrect word forms, agreement errors, etc.

The number of words in each source text is about 150 before pre-editing. In one NMT, the number of words in raw output texts is about 265 before pre-editing. The number of words in raw output texts is about 262 after pre-editing.

In the study neural machine translation systems are called NMT 1, NMT 2 and NMT 3 instead of NMT systems' trade names in order to avoid comparisons and judgements. The purpose of the study is to reveal some patterns rather that evaluate the particular NMT system.

There was one more text, the source text 3 "Why procrastination is about managing emotions, not time" (Jarrett 2020). At the stage of the first reading the decision was made not to pre-edit it. There were too many problem areas related to diversity of tenses, changing person forms of verbs, polysemantic words, phrasal verbs, variety of grammatical structures, etc. At the stage of translating before pre-editing the assumption was confirmed. The poor (in comparison to other texts) quality of NMT 1, NMT 2 and NMT 3 outputs showed that pre-editing would have taken too much time and efforts, and it presumably would not have improved the MT output quality significantly.

In January 2022 (5 months after the first case study) we repeated the experiment using a new source text from bbc.com "Two-thirds with Omicron say they have had Covid before" (Roberts 2022). The new study showed no significant difference from the previous one in terms of translation quality, patterns of errors and effectiveness of pre-editing techniques. This means, that despite the rapid changes in NMT systems and in the news texts themselves, there are a number of common mistakes to be prevented through pre-editing and a number of pre-editing techniques that work in different circumstances. To illustrate this fact the examples from new source text (ST-2022) will be given in the "Discussion" part.

Discussion

Often, the effect of pre-editing on the quality of NMT is unpredictable, but some orientation of changes in NMT output has been revealed, which depends on the received pre-editing (Mercader-Alarcón, Sánchez-Matínez 2016; Miyata, Fujita 2021, 1541). Let us list some methods of pre-editing texts for MT:

1. Break complex sentences into simple short sentences and medium-length sentences (5 to 20 words).
2. If there are two verbs in the sentence that convey two thoughts, divide this sentence into two.
3. Make sure all text is written in the same language.
4. Replace the infinitive, present participle, and past participle forms of verbs at the beginning of sentences with less ambiguous words.
5. Put verbs into active voice when possible.
6. Put verbs into simple tenses if the target language does not have the same system, e.g. there are no continuous and perfect tenses in Russian.
7. Eliminate idioms, slang, and jargon.
8. Check that punctuation is correct.
9. Insert, where necessary, missing words, articles and demonstrative pronouns before nouns (that, which, etc.).
10. Replace abbreviations and acronyms with full words, names or phrases, etc. (Kokanova, Berendyaev, Kulikov 2019; Machine translation tips 2016).

https://www.doi.org/10.33910/2686-830X-2022-4-1-25-30

Pre-editing techniques: lexical aspect

The results of translation texts before pre-editing show that the NMT systems do not cope well with translating polysemantic words, which determines the importance of pre-editing, finding and eliminating such problematic places. After pre-editing the systems do cope with the translation, but still require editing the output of the NMT, that is post-editing (PEMT).

ST 1: Why so many young Swedes live alone
ST 1 after pre-editing: Why so many young Swedes live independently

ST 2: girls with no access to school
ST 2 after pre-editing: girls without education

ST 2: to stress the sense of urgency
ST 2 after pre-editing: to stress the urgency of the problem

The new case study showed the same issue, and the same pre-editing technique helped to prevent the error:

ST-2022: Two shots offer little protection against catching Omicron
ST-2022 after pre-editing: Two injections offer little protection against catching Omicron

Also, NMT systems do not recognize whether the meaning of the verb is reflexive or non-reflexive if both variants are possible for the particular verb. It is necessary to identify such cases and pre-edit using verbs with clearer meaning. Pre-editing helps to avoid the potential mistake.

ST 1: I'd always wanted to move out of home and I'd always felt ready
ST 1 after pre-editing: I'd always wanted to move out of home and I'd always been ready

Another weakness of NMT systems is related to the fact, that they often do not "understand" idiomatic and colloquial language. As a result, when source text contains idioms or collocations, which is typical for news tests, the MT output looks more like word-for-word translation than transferring the original meaning. Therefore, pre-editing can include replacing idioms and collocations for more neutral and clear lexical items.

ST 1: In Sweden it's a different story
ST 1 after pre-editing: It's different in Sweden

ST-2022: 'Get boosted'
ST-2022 after pre-editing: 'Get the booster dose'

English phrasal verbs can be one more source of errors in NMT from English into Russian. Combining polysemy and colloquial meaning, they often lead to mistranslations. Pre-editing helps NMT cope with translation.

ST 2: girls who are completely missing out on school
ST 2 after pre-editing: girls who do not go to school at all

ST-2022: let in the fresh air
ST-2022 after pre-editing: air the room

There are some risks of errors to occur when the text contains words with abstract meaning. In order to reach understandable translation, those words can be replaced with more precise ones.

ST 2: Julia Gillard, former Australian prime minister, is campaigning for the right of girls to stay in education
ST 2 after pre-editing: Julia Gillard, former Australian prime minister, is campaigning for the right of girls to go to school

ST-2022: More work is needed
ST-2022 after pre-editing: More research is needed

Pre-editing techniques: grammatical aspect

In terms of grammar, the language typology should be taken into consideration. For instance, it is typical for English sentence to have difference between logical and syntactic subject, but sometimes that is not the case for Russian language. NMT systems do not change the subject and the error occurs. It is important to bring the syntactic structure closer to the logical one when pre-editing the sentences.

ST 1: A 2019 study found
ST 1 after pre-editing: The findings of a 2019 study found

ST-2022: Coronavirus infections have slowed recently
ST-2022 after pre-editing: The rates of coronavirus infections have slowed recently

In some cases grammatical structures themselves (e. g. structures with non-finite verbs) lead to omission of information necessary for the translated text. While pre-editing the structures can be changed in order to add the necessary information, but the result has to be checked at the stage of post-editing. For instance, in the following example the quality of some NMT raw outputs even worsened, because NMT systems confused gerund and participle V-ing forms.

ST 1: The most common age to leave home

ST 1 after pre-editing: The most common age of young people leaving home

In the following example pre-editing had predictably positive effect.

ST-2022: they had already previously tested positive for Covid

ST-2022 after pre-editing: they had already got the positive tests for Covid

The opposite situation is also possible. The grammatical structure may contain unnecessary words playing purely syntactic role. When translating into languages with different syntax they become unnecessary. To pre-edit in such cases means only to remove those words.

ST 2: Reaching 130 million girls

ST 2 after pre-editing: 130 million girls

Interestingly, using news texts as research material, we noticed fluctuating effectiveness of pre-editing texts for the following NMT systems: Amazon Translate, Google Translate and PROMT Professional Neural. Only an understanding of how a particular NMT system works, as well as the ability to predict the impact of pre-editing texts for a particular NMT system, allows you to determine the best methods for pre-editing, for example, news texts for NMT from English to Russian. Thus, some pre-editing techniques show the expected results, which means that their application is possible. However, several cases of negative impact of pre-editing on the quality of NMT have also been identified. Some techniques of pre-editing news texts minimize the occurrence of errors, but the need for post-editing of NMT output remains.

Results

In order to measure the amount of post-editing needed for NMT texts with and without pre-editing, we post-edited NMTs of ST 1, ST 1 after pre-editing, ST 2, ST 2 after pre-editing, ST-2022 and ST 2022 after pre-editing. We used TAUS guidelines (MT Post-Editing Guidelines 2010) for achieving quality similar or equal to human translation: corrected grammatical and lexical errors, ensured that no information was accidentally added or omitted, applied basic punctuation and spelling rules, etc.

In terms of human quality assessment, while post-editing we noticed, that pre-editing reduced the number of mistakes and made them easier to correct. It also reduced the time spent on post-editing. On average, pre-edited texts required 30–40% less time than texts without pre-editing.

Finally, we used the Memsource Post-editing Analysis tool (Analysis Overview 2021) in order to calculate the edit distance between raw NMT outputs and post-edited texts. The percentage of post-edited texts similarity to raw outputs was analysed given the amount of time and efforts spent on pre-editing and post-editing.

The pre-edited news texts are more than 90% similar to raw machine output. The news texts without pre-editing are more than 70% similar to raw machine output. Considering these facts, it is more appropriate to mention stability and predictability of NMT rather than higher or lower performance when applied to news texts. The pre-editing time is also significantly important. It takes not much time to pre-edit if there are enough skills to predict the impact of pre-editing news texts for a particular NMT system.

After the quantitative analysis, the following patterns of translation errors were extracted:

1. mistranslation of polysemantic words remains one of the most frequent NMT error in terms of lexis;
2. omission of necessary information as well as addition of unnecessary items is also typical for NMT systems;
3. machine translation systems do not “understand” some idioms, collocations and phrasal verbs. It can be argued, that dealing with metaphorical meaning is completely outside the current capabilities of NMT;
4. in terms of grammar, NMT often struggles with word order and cannot restructure the sentence even if the target language system requires it;
5. NMT cannot translate the headings properly as it cannot cope with incomplete sentences.

A number of issues mentioned can be tackled using the combination of pre-editing and post-editing. Polysemantic words can be replaced with synonyms. Complicated phrases and constructions can be eliminated as well as idioms and strong collocations. Sentences can be restructured and clarified. But in less frequent cases the effect of pre-editing can be unpredictable including lack of positive impact and even prevalence of negative one.

There is some dependence of the pre-editing rules on the target language to which the English news texts are translated, most of them would also be useful when the target language of the translation is not Russian. This is, for instance, the rule to expand acronyms to help the MT systems determine the term, to eliminate polysemy, or to divide too long sentences into shorter ones.

Another utterly important question is whether we should pre-edit or not. On the example of source text 3 we can see that in some circumstances pre-editing becomes inefficient due to amount of time and efforts needed. It is necessary to consider the fact that post-editing is inevitable in any case.

Conclusion

While NMT systems are constantly improving, there are still some issues that reduce the quality of raw machine output. By predicting those issues and eliminating problematic places through pre-editing human remains key figure in human-machine interaction. On the whole, pre-editing improves the quality of NMT from English into Russian, but in order to pre-edit effectively, it is necessary to learn the methods of pre-editing texts for NMT. Several cases of negative impact of pre-editing on the quality of NMT have been identified, that indicates the necessity to understand how the NMT works. Some methods of pre-editing news texts minimize the number of errors, but the need for PEMT output remains.

Knowing and using pre-editing and post-editing techniques reduce time and efforts needed for reaching comprehensible, accurate and stylistically fine translation quality.

Conflict of Interest

The author declares that there is no conflict of interest, either existing or potential.

Author Contributions

Authors contributed equally to the submission.

Abbreviations

MT — machine translation
NMT — neural machine translation
PEMT — post-editing machine translation
ST — source text

Sources

Coughlan, S. (2020) Reaching 130 million girls with no access to school. *BBC News*, 08.03.2020. [Online]. Available at: https://www.bbc.com/news/education-51769845 (accessed 29.08.2021). (In English)

Jarrett, Ch. (2020) Why procrastination is about managing emotions, not time. *BBC Worklife*, 14.05.2020. [Online]. Available at: https://www.bbc.com/worklife/article/20200121-why-procrastination-is-about-managing-emotions-not-time (accessed 29.08.2021). (In English)

Roberts, M. (2022) Two-thirds with Omicron say they have had Covid before. *BBC News*, 26.01.2022. [Online]. Available at: https://www.bbc.com/news/health-60132096 (accessed 28.01.2022). (In English)

Savage, M. (2019) Swedes typically stop living with their parents earlier than anywhere else in Europe. But can leaving home at a young age have a dark side? *BBC Worklife*, 22.08.2019. [Online]. Available at: https://www.bbc.com/worklife/article/20190821-why-so-many-young-swedes-live-alone (accessed 29.08.2021). (In English)

References

Analysis Overview. (2021) *Memsource Help Center*. [Online]. Available at: http://help.memsource.com/hc/en-us/articles/360013675760 (accessed 28.08.2021). (In English)

Barkhudarov, L. S., Kolshanskij, G. V. (1958) K voprosu o vozmozhnostyakh mashinnogo perevoda [On the possibilities of machine translation]. *Voprosy Yazykoznaniya*, vol. 1, pp. 129–133. (In Russian)

Dew, K. N., Turner, A. M, Choi, Yo. K et al. (2018) Development of machine translation technology for assisting health communication: A systematic review. *Journal of Biomedical Informatics*, vol. 85, pp. 56–67. https://doi.org/10.1016/j.jbi.2018.07.018 (In English)

Hays, D. G. (1960) Linguistic research at the RAND corporation. In: *Proceedings of the National Symposium on Machine Translation (2–5 February, 1960)*. Los Angeles: Englewood Cliffs Publ.; N. J. Prentice-Hall Publ., pp. 13–25. (In English)

Kokanova, E. S., Berendyaev, M. V., Kulikov, N. Yu. (2019) Tipy oshibok pri nejronnom mashinnom perevode tekstov ob arkticheskikh konvoyakh [Types of errors in neural machine translation texts about Arctic convoys]. In L. Yu. Shchipitsina (ed.). *Razvitie severo-arkticheskogo regiona: problemy i resheniya v gumanitarnoj sfere. Materialy Vserossijskoj nauchno-prakticheskoj konferentsii (25–27 aprelya 2019) [Development of the North-Arctic region: Problems and solutions in the Human Studies. Proceedings of the All-Russian scientific and practical conference (April 25–27, 2019)]*. Arkhangelsk: Northern (Arctic) Federal University Publ., pp. 80–84. (In Russian)

Machine translation tips. (2016) *IBM Cloud Docs*. [Online]. Available at: https://cloud.ibm.com/docs/GlobalizationPipeline?topic=GlobalizationPipeline-globalizationpipeline_tips (accessed 28.08.2021). (In English)

Marzouk, S., Hansen-Schirra, S. (2019) Evaluation of the impact of controlled language on neural machine translation compared to other MT architectures. *Machine Translation*, vol. 33, no. 3, pp. 179–203. https://doi.org/10.1007/s10590-019-09233-w (In English)

Mercader-Alarcón, J., Sánchez-Matínez, F. (2016). Analysis of translation errors and evaluation of pre-editing rules for the translation of English news texts into Spanish with Lucy LT. *Revista Tradumàtica: Tecnologies de la Traducció*, no. 14, pp. 172–186. http://dx.doi.org/10.5565/rev/tradumatica.164 (In English)

Miyata, R., Fujita, A. (2021). Understanding pre-editing for black-box neural machine translation. In: *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics (EACL)*. [S. l.]: Association for Computational Linguistics Publ., pp. 1539–1550. [Online]. Available at: https://doi.org/10.48550/arXiv.2102.02955 (accessed 28.08.2021). (In English)

MT Post-Editing Guidelines. (2010) *TAUS: The Language Data Network*. [Online]. Available at: https://www.taus.net/academy/best-practices/postedit-best-practices/machine-translation-post-editing-guidelines (accessed 30.09.2021). (In English)