LINGUISTICS AND LANGUAGES (UDC 81)

DOI: https://doi.org/10.18454/RULB.2020.22.2.33

ПРИМЕНИМОСТЬ МАШИННОГО ПЕРЕВОДА К ПЕРЕВОДУ ТЕКСТОВ

Научная статья

Страхова Д.А.*

ФГОУ ВПО Тульский государственный университет, Тула, Россия

* Корреспондирующий автор (dariya.strakhova[at]mail.ru)

Аннотация

В исследовании оценивается использование показателей сложности текста в качестве метрики его пригодности для машинного перевода (МТ). Эксперимент проводился на корпусе текстов научных статей, переведенных в паре Ru-En как профессиональными переводчиками, так и системой МТ DeepL. Показано, что существующие метрики сложности текста непригодны для прогнозирования качества МТ, так как оно оказывается недопустимо низким даже при обработке исходных текстов малой сложности. Кроме того, показано, что МТ непригоден для перевода практически 50% представленного корпуса текстов.

Ключевые слова: качество машинного перевода, показатели сложности текста, расстояние редактирования, CAT-инструменты, применимость машинного перевода.

INVESTIGATION OF MACHINE TRANSLATION APPLICABILITY

Research article

Strakhova D.A.*

Tula State University, Tula , Russia

* Corresponding author (dariya.strakhova[at]mail.ru)

Abstract

The study evaluates the text complexity metrics applicability in assessing text fitness for machine translation (MT). Possible approaches to such assessments are discussed. The experiment used a corpus of research papers translated from Russian into English by both a professional translator and the DeepL MT engine. It is shown that the existing metrics cannot be used to forecast the MT quality since the MT output may be unusable even for low-complexity source texts. It is also proved that MT is unsuitable for almost 50% of the specialized text to be translated.

Keywords: machine translation quality, text complexity metrics, editing distance, CAT tools, machine translation applicability.

Introduction

Widespread adoption of machine translation (MT) engines has been a controversial issue [5]. There have been numerous cases [6] of MT overuse resulting in gross translation errors. In such sensitive fields as engineering, medicine, etc., MT errors may have tragic consequences. Unfortunately, the public often perceives MT as a “one-fits-all” solution without making any reservations for the MT deficiencies. There is an urgent need for a tool to estimate whether MT can be applied to a specific text to be translated.

The paper is structured as follows: first we introduce the problem statement and review the available publications on the subject matter; then we cover the experimental research methodology, present and visualize its results; in the end, we discuss the results and show that in most the MT applicability is greatly overestimated.

Problem Statement

The study objective is improving translation quality by analyzing whether there are any correlations between available text complexity metrics and the PEMT editing distance. To achieve this goal, we analyzed large parallel Ru-En texts translated by professional translators and by the best available free MT engine, estimated the complexity metrics, compared the translation results, and checked if any correlations exist.

Literature Review

The MT quality problem has been extensively discussed in available references [1], [2], [3]. The researchers proposed to use various quality metrics [1], [2] and neural networks [4] to assess MT output quality. The common drawback of these solutions is that they are retroactive, not proactive: the assessment is applied after the MT output is generated, not before it. A better strategy would be evaluating the source text for its fitness for MT before translating. With such a metric, a potential LSP customer would determine whether MT can be used at all to translate the specific content. Such an approach is closely associated with the source text quality issues [7]. Unfortunately, the defective source text problem is not sufficiently studied so far.

Methodology

The study intended to identify common features of texts poorly translated by MT engines and to suggest a way to predict whether a certain text is suitable for MT or not. So far the only feasible MT quality strategy is comparing MT output to a professional translator output. The commonly used metric is the Levenshtein distance [8]. In [9], a similar methodology known as referential translation is used: “we introduce referential translation machines (RTM) for quality estimation of translation outputs, which is a computational model for identifying the acts of translation for translating between any given two data sets with respect to a reference corpus selected in the same domain.” We propose a slightly different strategy: to apply the text string matching algorithm available in any computer-aided translation (CAT) tool. It is commonly known that a CAT tool

looks up the translation memory (TM) for fuzzy matches. In most systems, the match value is indicated (usually in %.) Suppose we have three text arrays: source S, human-translated HS, and machine-translated MT. Now, MT is uploaded into a TM as the source text. MT is saved as a source text in a bilingual file. Now, the bilingual file is pre-translated. The CAT tool would match each HT segment stored in the TM to each MT segment stored in the bilingual file. As a result, for each segment the match value MV would be indicated. It is a reverse metric of the editing distance (ED): the higher the match, the lower the distance: ED=100-MV. Most CAT tools do not indicate matches lower than 40%, which is acceptable for our research. After that, the MV values are assigned to the source text segments to identify any similarities between the segments poorly translated by MT (ED>40%.) One option is to check the correlation with the source segment length.

The experimental dataset was a corpus of research papers translated by highly reputable professional translators from Russian into English. The dataset included 19,038 words after filtration. The source segments were filtered to remove numbers, empty entries, tags, and duplicates. The segment sequence was preserved to keep the context intact. The corpus was analyzed for several text complexity metrics. The average number of words per sentence was 14.6 while the average word length was 6.7 characters (shorter than the average 7.2-word length for Russian). The results were as follows:

Flesch-Kincaid level	4.81
Automated Readability Index	3.5
Coleman-Liau index	4.27
Gunning fog index	7.17

Surprisingly for research papers, the general original text readability metrics (language-agnostic) were high. The resulting dataset was translated by DeepL, arguably the best MT engine available today. It should be noted that the entire papers were uploaded for translation, so the MT engine could use all the context available. Then the texts were uploaded to SDL Trados Studio 2019 as indicated above, and the match values for each segment were obtained. The results were processed in an electronic spreadsheet.

Results

The first result was the distribution of the editing distances vs. the number of words (Fig. 1). The distances are normalized to the 0…100 range.

The diagram clearly shows that the MT engine failed to properly translate an overwhelming majority of words. Only 26.29% of the words were translated with up to 20% editing distance, while for 40.09% the editing distance exceeded 50%. Such a poor performance clearly shows the MT deficiency further enhanced by the language pair (inflected to analytic language) and the subject matter (complicated research texts in engineering, biology, neural networks, simulation, etc.).

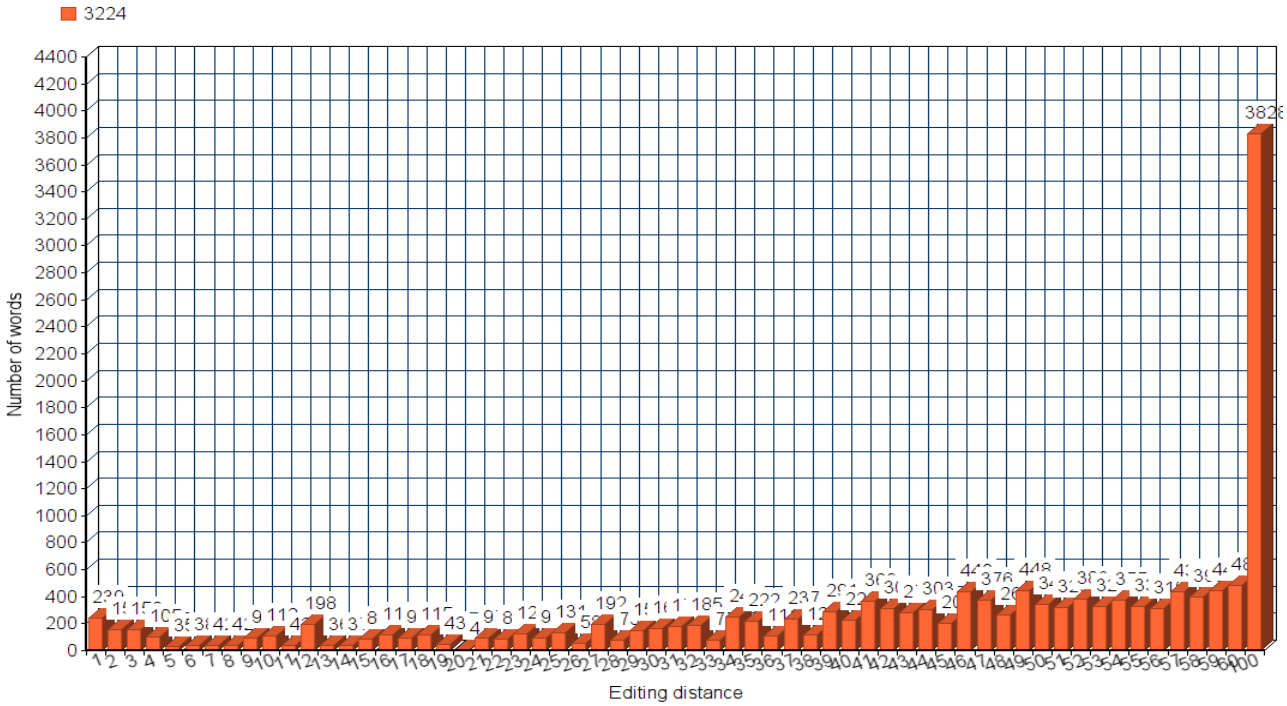

Figure 1 – Distribution of editing distances

One hypothesis to be verified is that MT poorly handles long, complicated text segments. We checked the correlation between the sentence length in words and the editing distance (Fig. 2).

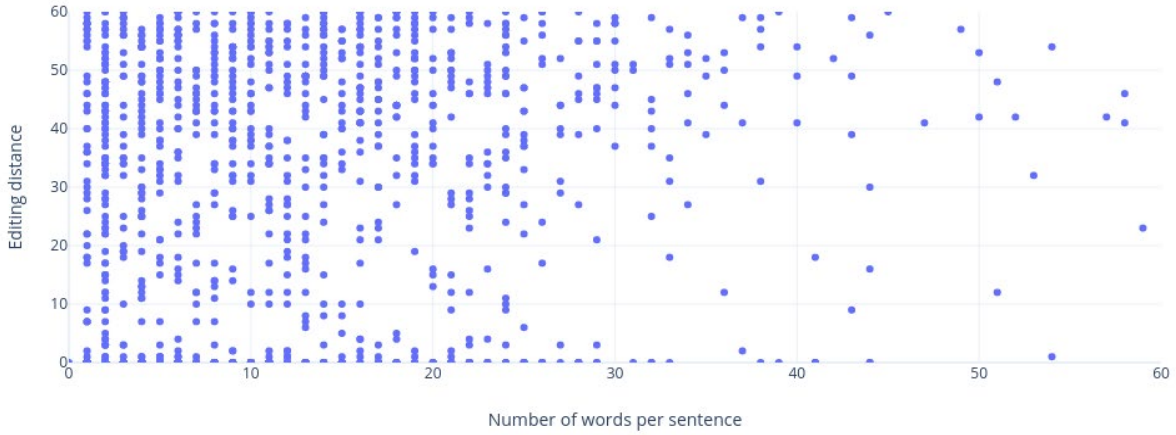

Figure 2 – Editing distance vs. number of words per sentence scattering graph

Discussion

The texts in Russian that were presented for MT translations were not too complicated in terms of grammar, sentence length, etc. Still, about half of the MT output happened to be unusable as the editing distances over 60% mean the translator has to re-write the entire sentence. Using MT+PEMT is not worth the effort since the regular translation by a human translator would take less time. Our study showed that the commonly used language-agnostic text complexity metrics cannot be used as an indicator of the source text fitness for MT since MT makes gross errors regardless of the source text complexity. Another point is that the MT value/usability seems to be greatly exaggerated since it fails to translate 40…50% the texts with at least acceptable quality.

Conclusion

The experimental results showed that MT is unsuitable for about half of the text segments presented. It confirms the hypothesis about a rather limited MT applicability to specialized (scientific, technical) texts. Another result is that text complexity metrics cannot be used to identify what text is more suitable for MT translation. Further research will be aimed at developing a new metric to assess text suitability for machine translation. One option is to use neural networks to find possible similarities between text/segments that MT is unable to handle properly. A metric estimating the share of specialized terms would also be inefficient since the text corpus used in this research has a rather low Flesch-Kincaid level. Since segment length/complexity is also irrelevant, more in-depth text analysis is required.

Конфликт интересов

Не указан.

Conflict of Interest

None declared.

Список литературы / References

1. Pooja Gupta. Quality Estimation of Machine Translation Outputs Through Stemming / Pooja Gupta et al. // International Journal on Computational Sciences & Applications (IJCSA). 2014. Vol .4, No. 3, June. pp. 15-24.

2. Mariano Felice. Linguistic Indicators for Quality Estimation of Machine Translations / Mariano Felice. Master's thesis, University of Wolverhampton, UK. – 2012

3. Aaron Li-Feng Han. Quality Estimation for Machine Translation Using the Joint Method of Evaluation Criteria and Statistical Modeling / Aaron Li-Feng Han et al. // Proceedings of the Eighth Workshop on Statistical Machine Translation. 2013. August, pp. 365–372.

4. Zhiming Chen. Improving Machine Translation Quality Estimation with Neural Network Features / Zhiming Chen et al. // Proceedings of the Second Conference on Machine Translation. pp. 551-555.

5. Joao Graca. Why Quality Estimation Is the Missing Link for Machine Translation Adoption / Joao Graca // Forbes Technology Council, Jan 24, 2019.

6. Vilar, David. Error Analysis of Machine Translation Output / Vilar, David & Xu, Jia & D'Haro, Luis et al. // International Conference on Language Resources and Evaluation, May, 2006. pp. 697-702.

7. Molnár O. Source Text Quality in the Translation Process / Molnár O. // Tradition and Trends in Trans-Language Communication. Palacký University Olomouc, 2013. pp. 59-86.

8. Li Yujian. A Normalized Levenshtein Distance Metric / Li Yujian, Liu Bo // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007. Vol. 29, No. 6, June. pp. 1091-1095.

9. Ergun Bic¸ici. Referential translation machines for quality estimation / Ergun Bic¸ici // Proceedings of the 8th Workshop on Statistical Machine Translation, pp. 343–351, 2012, Sofia, Bulgaria.