Automatic evaluation of the quality of machine translation of a scientific text: the results of a five-year-long experiment

Ilya Ulitkin[1,*], *Irina* Filippova[1], *Natalia* Ivanova[1], and *Alexey* Poroykov[1]

[1]Moscow Region State University, 10A, Radio Str., 105005, Moscow, Russia

Abstract. We report on various approaches to automatic evaluation of machine translation quality and describe three widely used methods. These methods, i.e. methods based on string matching and *n*-gram models, make it possible to compare the quality of machine translation to reference translation. We employ modern metrics for automatic evaluation of machine translation quality such as BLEU, F-measure, and TER to compare translations made by Google and PROMT neural machine translation systems with translations obtained 5 years ago, when statistical machine translation and rule-based machine translation algorithms were employed by Google and PROMT, respectively, as the main translation algorithms [6]. The evaluation of the translation quality of candidate texts generated by Google and PROMT with reference translation using an automatic translation evaluation program reveal significant qualitative changes as compared with the results obtained 5 years ago, which indicate a dramatic improvement in the work of the above-mentioned online translation systems. Ways to improve the quality of machine translation are discussed. It is shown that modern systems of automatic evaluation of translation quality allow errors made by machine translation systems to be identified and systematized, which will enable the improvement of the quality of translation by these systems in the future.

1 Introduction

To date, there exist numerous machine translation systems based on different approaches to translation and computer algorithms. They exhibit different advantages and disadvantages, but a feature they all share is that they are normally used as fully automatic devices, translating a source text into a target language without human intervention. Unlike fully automated translation, human translation (or simply "translation") is an activity performed exclusively by people, perhaps with the use of computers as a word processor and electronic dictionaries as aids.

Machine translation is the translation of a text from a source language into a text in a target language using computer programs. However, professional translators may be involved in the pre-editing of the source text or post-editing of the translated text, but they do not usually intervene in the translation process itself.

* Corresponding author: ulitkin-ilya@yandex.ru

Although the concept of machine translation can be traced back to the seventeenth century, it was only in the second half of the twentieth century that the US government-funded research stimulated international interest in machine translation systems.

Initially, scientists had intended to create a fully automated high-quality machine translation system, but by the 1960s the number and difficulty of linguistic problems became so obvious that many researchers (who were not linguists, no matter how strange it seems) were disillusioned and realized that the development of fully automated systems was unrealistic at least at that level of sophistication involved, with the translation of such systems requiring human intervention (because many complex elements of a language cannot be easily programmed, such as homonyms or metaphors) [1].

The first automated translation system, which translated only 250 words from Russian and English, was publicly demonstrated in the United States in 1954. We now call the basis of this system first-generation architecture, which used only a dictionary, trying to match the words of the source language with those of the target language, that is, to translate directly. This approach was simple and cheap, but the results were meager, mimicking the syntactic structures of the original language [2]. Moreover, this approach was more suitable for pairs of languages that were structurally related. Despite the poor translation quality, the project was supported and stimulated further research in the United States and the Soviet Union.

By the mid-1960s, a new generation of machine translation systems had been developed, based on an interlingual approach to translation. Nevertheless, in 1964, the US government commissioned a report from the Automatic Language Processing Advisory Committee (ALPAC), which criticized the slowness of machine translation systems, lack of accuracy, and the cost of machine translation compared to human translations. The future of machine translation looked extremely bleak. As a result, the US government stopped investing significant funds into the development of machine translation systems.

Although automated translation systems proved unsuitable for replacing human translations, they were quite accurate in translating specific texts or texts of a particular genre. For example, the Météo system developed in Canada in 1976 was quite successful in translating weather forecasts between French and English.

By the late 1970s the interlingual approach in which a source text was transformed into a special ‘interlingua’, after which a target text was generated from this intermediate form with the help of the target-language dictionaries and grammar rules during the synthesis stage, had exhausted itself. The main problem was the impossibility of building a truly language-neutral representation that unites all possible aspects of syntax and semantics for all known languages [2]. Much has been written about this approach but to this date interlingual systems are only available as prototypes.

In the late 1970s and early 1980s research focused on the transfer approach. In this architecture, a source text was analyzed by the source language dictionary and converted into an abstract source-language representation. This representation was further transferred into its equivalent target-language representation, followed by translation into the target language. This rule-based approach was less sophisticated than interlingual and direct translation. However, scientists faced the problem that dictionaries did not contain enough knowledge to resolve ambiguities.

Compiling machine translation dictionaries is a laborious and expensive process, because they have to contain a wealth of information to address issues such as lexical ambiguity, complex syntactic structures, idiomatic language, and anaphora in multiple languages. Austremühl [3] emphasizes that knowledge about the world is especially difficult to implement in machine translation systems, since a computer cannot make the same knowledge-based decisions as humans. If the dictionary is too small it will not have

enough information; if it is too large the computer will be less likely to choose the correct meaniings.

In the 1990s research led to the development of a third generation machine translation system—the corpus-based translation system (namely, statistical machine translation systems and example-based translation systems). In the statistical-based approach a source text is segmented into phrases and then compared to an aligned bilingual corpus using statistics and probability theorem to select the most appropriate translation. The example-based approach imitates combinations of pre-translated examples in its database. For this approach to be successful the database must contain close matches to a source code. This approach is also at the heart of translation memory tools.

All machine translation architectures work best with technical texts with limited or repetitive vocabulary. Thus, Gross [4] demonstrated that a piece of high or even medium literary value is best left to human translators, while texts with mathematical and abstract concepts fall into a category that can be translated quite well by a computer.

In the 1980s, there was a shift towards pre-editing, during which the original text was simplified according to certain rules so that it would be easier for the computer to translate it. Then the translation process was carried out by the machine. And after that the text was post-edited. The European Commission, which has been using machine translation since the 1960s, has established that "as long as the translation can be restricted in subject matter or by document type ... improvements in quality can be achieved" [5].

The development of international communications and the growth of the localization industry have made it quite obvious that translators cannot meet the huge demand for cheap, fast (even instant), often large-scale information exchange in different languages. Huge investments have been made in the development of machine translation systems for private and public use, primarily in the main languages. Hybrid systems have emerged that combine rule-based and corpus-based architectures to improve the accuracy and quality of translations.

The growth in information has changed the attitude of companies towards translation, as their goal is often a simple exchange of information. For example, European Commission workers often only need an idea of what the document is about. Common users can rely on free online machine translation systems to understand the essence of what is written on the website. That is, when there is a need for a simple understanding of the content of a text, a machine translation is often a much faster and more cost-effective solution than a human translation.

Recent developments in machine translation have led to the introduction of deep learning and neural networks to improve the accuracy of translations. Language service providers now offer customized machine translation engines where, in addition to including terminology from a specific field such as life sciences, the travel industry or IT, the user can also upload their own translation data to improve the accuracy, style and quality of machine translation.

On November 15, 2016, Google deployed neural machine translation in its Google Translate. Initially, a total of eight language pairs for English were announced, including French, German, Spanish, Portuguese, Chinese, Japanese, Korean and Turkish. The neural machine translation (NMT) system developed by Google improves translation fluency and accuracy by using an artificial neural network suitable for deep learning. By studying millions of examples, NMT generates the most appropriate translation, with the result being adapted to fit the grammar of human language.

Currently, there is a constant increase in the use of various online services for translating texts from one language to another. Many countries spend huge sums on the development of such services in the hope that in the near future automatic translation

systems will be able to meet the growing human needs for translation from one language to another.

With the development of machine translation (MT) systems, an impetus was given to a large number of investigations, thereby encouraging many researchers to seek reliable methods for automatic MT quality evaluation. Machine translation evaluation serves two purposes: the relative estimate allows one to find out whether one MT system is better than another, and the absolute estimate (having a value ranging from 0 to 1) gives a measure of efficiency (for example, the figure 1 means perfect translation).

However, the development of appropriate methods for numerical MT quality evaluation is a challenging task. In many fields of science, measurable efficiency indices exist, such as, for example, the difference between the predicted and actually observed results. Since natural languages are complicated, an assessment of translation correctness is extremely difficult. Traditionally, the bases for evaluating MT quality are adequacy (the translation conveys the same meaning as the original text) and fluency (the translation is correct from the grammatical point of view). The existing automatic metrics for assessing the quality of translation are constantly being improved and allow an objective evaluation of the quality of certain systems. In this regard, the purpose of this work is to analyze the quality of machine translation of scientific and technical text performed by the neural machine translation systems Google and PROMT, and to compare the results with the data of our research in 2016 [6].

2 Research method and materials

The quality of translation (by this term we mean the level of the quality of the performed translation or interpretation, evaluated on the basis of a certain set of objective and subjective criteria adopted for evaluating this type of translation or interpretation) depends on a number of objective and subjective factors. It is determined, first of all, by the quality of the source text or the translated speech, as well as the professional qualifications of the translator and his or her readiness for the implementation of this particular act of translation.

Considering the methods for evaluating the quality of translation, we distinguish between expert evaluation of translation (manual evaluation of the quality of translation) and automatic.

2.1 Expert evaluation of machine translation quality

Translation is an intellectual challenge, and, therefore, skepticism about the possibility of using a computer for automated translation is quite natural. However, the creators of MT systems have managed to endow their systems with a form of understanding, and machine translation now belongs to a class of artificial intelligence programs.

Expert (also manual and professional) evaluation of translation quality is very subjective. A specific text and translation task largely determine the criteria and, as a consequence, the result of the evaluation [7].

The parameters that experts employ to evaluate machine translation quality may vary from project to project. For example, according to Mitrenina [8], the key parameters in the expert evaluation should be adequacy, which evaluates the accuracy of translation, and fluency, which is responsible for the correctness of the translation. However, expert evaluation cannot be limited to adequacy and fluency only. Mitrenina [8] also notes that an expert can also evaluate translation in terms of the time and effort spent on translation, i.e. to evaluate the translation according to the resources spent by the translator/editor on correction and revision.

The methods of translation quality evaluation, developed by the Automatic Language Processing Advisory Committee (ALPAC) and the Advanced Research Projects Agency (ARPA), are considered to be the first and primary expert techniques.

ALPAC was established in April 1964 to evaluate the progress in computational linguistics in general and in machine translation in particular. Perhaps the committee's most notorious report was issued in 1966. The Committee was quite skeptical of the results in the field of machine translation, and emphasized the need for basic research in computational linguistics. In addition, it recommended that government funding in this field be cut off [9]. Nevertheless, ALPAC developed practical methods to compare machine translation of texts from Russian into English with a reference human translation, with two major measuring characteristics being intelligibility (all ratings were made on a scale of 1 to 9) and fidelity (all rating were made on a scale of 0 to 9) [9]. Thus, point 1 on the intelligibility scale was described as hopelessly unintelligible, whereas point 9 meant a perfectly clear and intelligible translation without stylistic infelicities. Fidelity was a measure of how much information the translated sentence retained compared to the original. Being measured on a scale of 0 to 9, each point was associated with a textual description. Thus, point 9 on the fidelity scale was described as highly informative, whereas point 0 meant that the original contains less information than the translation.

ALPAC made the following conclusions: Firstly, speaking of expert evaluation, one should take into account that the characteristics of intelligibility and fidelity turned out to be highly interrelated. Secondly, it became clear that the minimum number of experts should be four. And, thirdly, experts must know the subject area and the original language in order to successfully evaluate the translation [9].

The Defense Advanced Research Projects Agency (DARPA), originally known as the Advanced Research Projects Agency (ARPA), was established in February 1958. In 1991, DARPA compared rule-based statistical translation systems and human-aided translation systems and worked out successful methods that are presently included in standard evaluation programs.

It can be concluded that an expert evaluation of human-aided and machine translation systems requires a great deal of work and a large number of people, because experts involved in the evaluation of translation can only assess a limited amount of sentences and texts. Moreover, it remains unclear how to evaluate translations and which criteria to apply.

2.2 Automatic evaluation of machine translation quality

Measuring translation quality is a challenging task, primarily due to the lack of definition of an ‘absolutely correct’ translation. The most common technique of translation quality evaluation is to compare the output of automated and human translations of the same document. But this is not as simple as may seem: One translator's translation may differ from that of another translator. This inconsistency between different reference translations presents a serious problem, especially when different reference translations are used to assess the quality of automated translation solutions.

A document translated by specially designed automated software can have a 60% match with the translation done by one translator and a 40% match with that of another translator. Although both professional translations are technically correct (they are grammatically correct, they convey the same meaning, etc.), 60% overlap of words is a sign of higher MT quality. Thus, although reference translations are used for comparison, they cannot be a completely objective and consistent measurement of the MT quality.

Automatic evaluation of machine translation quality using reference texts is based on the use of various metrics that make it possible to simplify and reduce the cost of quality evaluation.

The first metric available for the Russian ↔ English pair is BLEU (Bilingual Evaluation Understudy). Note that this metric is one of the most popular and available at the moment, which makes it possible to calculate the BLEU score using tools such as MT-ComparEval, Interactive BLEU score evaluator, and many others.

The main idea behind this metric is as follows: the closer the machine translation (MT) to a translation made by a professional translator, the better. In order to evaluate the quality of machine translation, the degree of closeness of MT to one or more human translations is measured using a numerical metric. Thus a system for evaluating MT should have two components: 1) a numerical metric by which the proximity of translations is calculated and 2) examples (corpus) of good quality translations made by translators [10]. The BLEU metrics compares *n*-grams of the candidate translation with the reference translation, and the number of matches is calculated. The higher the number of matches, the better the quality of the candidate translation. It can be argued that the BLEU metric relies on two characteristics. The first is precision. In order to assess precision, the number of the words (unigrams) from the candidate translation that occur in any of the standard translations is calculated. Unfortunately, machine translation systems can in some cases generate too many 'necessary' words (which, for example, may result in the appearance of the repeating article 'the the the' in the translation), which, in turn, can lead to precision that is too high. In order to avoid this problem, the maximum number of words from the candidate translation that are in one of the reference translations is calculated. Then the total number of words of each candidate translation is reduced to the maximum number of the same (matched) words in the reference translation and is divided by the total (unlimited) number of words in the candidate translation [10]. It should be noted that such a calculation takes place not only for unigrams, but also for *n*-grams. This calculation of precision gives an idea of two aspects of translation: adequacy and fluency. Translation using the same words (unigrams) as in the reference translation tends to correspond to an adequate translation. Longer matches of *n*-grams indicate fluency in translation [10]. Precision is evaluated by multiplying all *n*-grams and extracting the fourth root from the product, and so the geometric mean is obtained. The second component of the BLEU metric is brevity penalty (BP). The penalty is scored on a scale from 0 to 1: BP is equal to unity if the length of the candidate translation is greater than the length of the reference translation. BP is less than unity if the length of the candidate translation is equal to or less than the length of the reference translation [10]. Thus, the closer the score to unity, the greater the overlap with the reference translation and, therefore, the better the MT system. The BLEU metric measures how many words coincide in the same line, with the best score given not to matching words but to word sequences. For example, a string of four words in the translation that matches the human reference translation (in the same order) will have a positive impact on the BLEU score and is weighted more heavily (and scored higher) than a one- or two-word match. Note that the peculiarity of the BLEU metric is that it is based on the exact match of word forms. It can be argued that this metric best suits analytical languages, where the forms may coincide in many cases. In addition, it is also important to emphasize that BLEU does not take into account syntax and word order (but it does take into account longer matching *n*-grams).

Another measure available for the Russian ↔ English language pair is TER (Translation Edit Rate).

TER calculates the number of corrections needed to make the resulting translation semantically similar to the reference one. The TER score measures the amount of editing that a translator would have to perform to change a translation so it exactly matches a reference translation. By repeating this analysis on a large number of sample translations, it is possible to estimate the post-editing effort required for a project. TER scores also range from 0 to 1. However unlike the other scores, with TER a higher score is a sign of more

post-editing effort and so the lower the score the better, as this indicates less post-editing is required [11].

Deletions, shifts, replacement of individual words, etc. are possible changes estimated by TER. The shift occurs within the translation by moving an adjacent sequence of words. All changes, including shifts of any number of words by any distance, have the same 'cost.' In addition, punctuation marks are considered as ordinary words, and incorrect use of the case is considered a change [11].

One more measure available for the Russian ↔ English language pair is the F-measure. The F-measure developers claim that their metric shows the best match with the human evaluation [12]. However, this is not always the case. The F-measure metric does not work very well with small segments [13].

The F-Measure score measures how precise the MT system operates when retrieving words and how many words it can retrieve or recall during translation. This is why it is commonly referred to as a Recall and Precision measurement. By expressing these two measurements as a ratio, it is a good indicator as to the performance of the translation system and its ability to translate content. F-Measure scores range from 0 to 1, the closer to unity the score, the better the recall and precision of the translations. The F-Measure gives an indication as to the quality of the translations that a translation system will output.

3 Automatic translation quality evaluation of Google and PROMT translation systems

In 2016, we selected for analysis 500 sentences from scientific papers of the journal 'Quantum Electronics' (http://www.quantum-electron.ru/) and their translations into English made by professional translators. In the same year (2016) Russian sentences were translated by the Google and PROMT MT systems, and these translations were compared with the reference translation (see *Bulletin of Moscow Region State University*. *Series: Linguistics*. 2016. no. 4. pp. 174-182). The same Russian sentences were translated into English in 2021 using the Google and PROMT MT systems and their translations were analyzed again to evaluate their quality.

For the automatic translation quality evaluation, we used the Language Studio™ Lite program from the website (http://www.languagestudio.com), which is free and allows one to evaluate the MT quality using such popular metrics as BLEU, F-Measure, and TER [14].

3.1 Automatic evaluation of machine translation quality using n-gram metrics

First, we compared the reference text (the translations made by the human translators) and the Google and PROMT candidate texts (the translations made by MT systems) using the *n*-gram metric. The results obtained for Google and Prompt in 2016 and 2021 are presented in Table 1 and 2 below.

Table 1. Analysis of translations of Google Translate for 2016 and 2021, based on the *n*-gram model.

Translation Evaluation Summary

Job Start Date:	12/29/2015 10:20 AM	2/9/2021 11:40 AM
Job End Date:	12/29/2015 10:20 AM	2/9/2021 11:41 AM
Job Duration:	0 min(s) 12 sec(s)	0 min(s) 17 sec(s)
Reference File:	science_reference_corrected.txt	science_reference_corrected.txt
Candidate File:	science_google_corrected.txt	google_translation_2021.txt
Evaluation Lines:	500	500
Tokenization Language:	EN	EN
Results Summary:	**46.147**	**50.194**

Table 2. Analysis of translations of PROMT translation service for 2016 and 2021, based on the *n*-gram model.

Translation Evaluation Summary

Job Start Date:	12/29/2015 10:21 AM	2/9/2021 11:43 AM
Job End Date:	12/29/2015 10:21 AM	2/9/2021 11:43 AM
Job Duration:	0 min(s) 12 sec(s)	0 min(s) 13 sec(s)
Reference File:	science_reference_corrected.txt	science_reference_corrected.txt
Candidate File:	science_PROMT_corrected.txt	prompt_translation_2021.txt
Evaluation Lines:	500	500
Tokenization Language:	EN	EN
Results Summary:	**30.791**	**44.420**

The results show that over the past 5 years, the Google MT system has improved its performance by about four percentage points (50.19% matches in 2021 compared to 46.14% in 2016), which allows us to conclude that there is an undoubted improvement in the quality of machine translation based on neural learning. When translating scientific texts, the PROMT MT system showed an almost 14 percentage point improvement in translation quality (44.42% matches in 2021 compared to 30.79% in 2016), which is not surprising, since in 2016 the PROMT MT engine was developed for rule-based translation rather than for statistical translation based on *n*-grams.

Based on the data obtained, we can conclude that the Google and Prompt MT systems translate scientific and technical texts well where terminology and simple sentences prevail. Thus, the threshold of full matches at a level of 75% for both MT systems is more than 100 out of 500 sentences. This percentage of matches ensures minimal costs for post-editing machine translation.

Analysis of the obtained data shows that in sentences with 100% to 70% matches we encounter one or two mistakes in translation, whereas in sentences with 69% to 50% matches sentences contain three or more mistakes. At the same time, both MT systems demonstrate high-quality translations of simple unextended and extended sentences, as well as of complex sentences. The systems demonstrate 'good knowledge' of scientific terminology.

The main mistakes that we have found when comparing reference sentences and machine-translated sentences are incorrect translations of abbreviations. Machine translation systems are not always able to correctly translate specialized abbreviations, which, in turn, are easily translated by a professional translator who specializes in a particular field. All this suggests that in order to achieve the best MT quality it is necessary to decipher all abbreviations throughout the entire text. Another example of mistakes in machine-translated sentences is incorrect word order although, it should be noted that such mistakes are becoming less and less frequent in MT systems after their corresponding training.

Thus a comparison of reference translations with machine translations makes it possible to single out the following repeating mistakes in both MT systems: the absence of articles, errors in the meaning and choice of a word, and violation of the order of words in sentences.

The findings indicate that Google and Prompt MT systems are constantly being improved. The latter suggests that the potential of neural machine translation systems will improve over time.

3.2 Assessment of translation quality using comparative metrics BLEU, F-measure and TER

The second analysis was performed using metrics such as BLEU, F-measure, and Translation Error Rate (TER). A comparison was made of two candidate texts with a reference translation. The results of research for 2016 and 2021 are presented in Tables 3 and 4 below.

Table 3. Evaluation of the quality of translations made by Google and Prompt in 2016 using BLEU, F-measure and TER metrics [6].

Translation Evaluation Summary

Job Start Date:	12/29/2015 10:17 AM
Job End Date:	12/29/2015 10:18 AM
Job Duration:	0 min(s) 44 sec(s)
Number of Reference Files:	1
Number of Candidate Files:	2
Evaluation Lines:	500
Tokenization Language:	EN
Evaluation Metrics:	BLEU, F-Measure, TER (Inverted Score)

Results Summary

Candidate File:	**1**	**2**
BLEU Case Sensitive	**24.54**	**42.10**
BLEU Case Insensitive	**25.98**	**43.62**
F-Measure Case Sensitive	**60.01**	**72.26**
F-Measure Case Insensitive	**61.35**	**73.24**
TER Case Sensitive	**38.07**	**54.43**
TER Case Insensitive	**38.70**	**54.94**

Candidate Files:

1 : science_PROMT_corrected.txt
2 : science_google_corrected.txt

Reference Files:

1 : science_reference_corrected.txt

-- Report End –

Table 4. Evaluation of the quality of translations made by Google and Prompt in 2021 using BLEU, F-measure and TER metrics.

Translation Evaluation Summary

Job Start Date:	2/9/2021 11:40 AM
Job End Date:	2/9/2021 11:41 AM
Job Duration:	0 min(s) 54 sec(s)
Number of Reference Files:	1
Number of Candidate Files:	2
Evaluation Lines:	500
Tokenization Language:	EN
Evaluation Metrics:	BLEU, F-Measure, TER (Inverted Score)

Results Summary

Candidate File:	**1**	**2**
BLEU Case Sensitive	**40.25**	**45.79**
BLEU Case Insensitive	**41.53**	**47.35**
F-Measure Case Sensitive	**71.41**	**75.05**
F-Measure Case Insensitive	**72.20**	**75.79**
TER Case Sensitive	**53.03**	**56.82**
TER Case Insensitive	**53.42**	**57.21**

Candidate Files:

1 : prompt_translation_2021.txt
2 : google_translation_2021.txt

Reference Files:

1 : science_reference_corrected.txt

-- Report End –

As in the previous Section, the Google MT system shows a slight increase in translation quality in 2021 when comparing the results of the analysis of translations performed by the MT statistical system and the MT neural system. At the same time, the MT Prompt system shows a significant improvement in its performance compared to 2016, which is explained by the transition from rule-based translation to translation based on neural learning.

The results of comparing data for 2021 show that there is an alignment of the scores of the Google and Prompt MT systems, which is associated primarily with the use of neural learning of MT systems in translation in both systems.

MT systems based on neural learning are trained on the huge corpora of existing translations into different language pairs. Unlike the statistical approach to translation, the search algorithms of which intuitively prefer to use the sequences of words that are the most likely translations of the original ones, neural machine translation systems do not just search for matches to a word and phrases, but carefully study the relationship between two languages. Analyzing each segment of texts allows modern systems to understand its context by determining the meaning of each word in the segment that needs to be

translated. As a result of such an analysis, neural machine translation systems select the necessary grammatical structures, correctly restoring the semantics and structure of the translation text.

Thus in our study we found the following trend. Modern neural machine translation systems demonstrate approximately the same results, which is explained by the use of similar algorithms used to generate translation.

Based on the search for the maximum number of matches between MT systems and reference translations, that is, the ratio between the total number of matching words to the length of the translation and the reference text, the F-measure metric shows the best results. This suggests that for the most part the number of words in reference texts and candidate texts is close (more than 70% for scientific texts when using the Google and PROMT MT systems). In addition, the matches are found not only at the level of the number of words, but also at the level of vocabulary, which is also quite important, since the less the editor has to edit the text, the better.

The TER metric based on measuring the amount of editing showed worse results: for scientific and technical texts more than 50% when using Google and PROMT MT systems.

The worst result of the three metrics was obtained using the BLEU metrics based on *n*-grams. The BLEU metric determines how many words match in a line, and the best result is given not just by matching words, but by a sequence of words. For scientific and technical texts, the result was more than 45% using the Google system and more than 40% using the PROMT system.

4 Conclusion

The paper provides an overview of the most commonly used MT quality evaluation metrics. As a rule, these metrics show a good correlation of candidate translations with reference ones. One of the important disadvantages of all these metrics is that they cannot provide an evaluation of the MT quality at the level of meaning. However, they are presently the only systems for automatic evaluation of the MT quality.

The results obtained in 2021 show a noticeable improvement in the quality of neural machine translation of Google and PROMT systems compared to 2016, which is quite justified, since the new MT technology demonstrate noticeable advantages over the statistical model, and even more so over the rule-based machine translation systems.

The comparison of the reference translation with Google and PROMT translations carried out in this work allows us to conclude that most mistakes are at the level of semantics, i.e. machine understanding of the source code. All this suggests that at the present time there are no necessary databases of semantic constructions that would enable the repetition of such mistakes to be avoided. It is also worth noting that MT systems experience considerable difficulties in translating complex grammatical, syntactic and lexical structures. In this regard, it is necessary to understand that an adequate and complete automatic evaluation of the quality of translations makes it possible to identify and systematize not only errors in MT systems, but also the shortcomings of existing MT programs, which will help solve these problems in the future.

Analysis of the MT quality of candidate texts made by Google and PROMT with reference translation, carried out using the *n*-gram model and various metrics, shows that Google translation demonstrates the best correspondence with the reference translation at the vocabulary level, which is quite expected, since the training of the system is carried out on huge parallel corpora of texts, while there is an improvement in translation at the syntactic level, which is most likely associated with an improvement in translation technology. Comparison of the results of 2016 and 2021 for PROMT revealed the most noticeable growth in all scores, which is associated with the transition to neural training of

this online service. The lag behind its rival Google can be explained by the fact that PROMT is a hybrid system that takes advantage of neural learning and rule-based translation. However, all the advantages of this system are revealed most fully only with active PROMT training on large bilingual corpora (from 50,000 segments), which is not always easy to implement in practice.

The development of effective and reliable metrics for evaluating the MT quality has been actively studied in recent years. One of the most important tasks is to go beyond *n*-gram statistics while continuing to use fully automatic mode. The need for fully automatic metrics cannot be underestimated, since they provide the greatest speed of development and progress of MT systems.

Acknowledgements

The authors wish to thank their colleague Stephen Garratt (England) for his helpful suggestions on manuscript editing and polishing the language of the paper.

References

1. J. Hutchins, *Encyclopedia of Language and Linguistics* (Oxford, Elsevier, 2006) http://www.hutchinsweb.me.uk/EncLangLing-2006.pdf
2. C.K. Quah, *Translation and Technology* (Basingstoke, Palgrave, 2006)
3. F. Austermühl, *Electronic Tools for Translators* (Manchester, St Jerome, 2001)
4. A. Gross, *Computers and Translation* (London, Routledge, 1992)
5. J. Hutchins, *Current Commercial Machine Translation Systems and Computer-Based Translation Tools: System Types and Their Uses*, http://www.hutchinsweb.me.uk/IJT-2005.pdf
6. I.A. Ulitkin, Bulletin of Moscow State Region University. Series: Linguistics **4**, 174 (2016)
7. V.N. Komissarov, A.L. Koralova, *Practical Work On Translation From English Into Russian* (Moscow, Vysschaya Shokla Publ., 1990)
8. O.V Mitrenina, A*pplied and Computer Linguistics* (Moscow, URSS Publ., 2016)
9. *ALPAC Report 1966*, https://www.nap.edu/read/9547/chapter/1#vii
10. K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Proc. ACL'02, 311–318 (2002)
11. M. Snover, B. Dorr, R. Schwartz, L. Micciulla, *A Study of Translation Edit Rate with Targeted Human Annotation* http://www.cs.umd.edu/~snover/pub/amta06/ter_amta.pdf
12. I.D. Melamed, R. Green, J.P. Turian, *Precision and Recall of Machine Translation HLT-03*, 61–63 (2003)
13. J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, N. Ueffing, *Proceedings of COLING* (Geneva, 2004)
14. I.A. Ulitkin, *Human Translation vs. Machine Translation: Rise of the Machines,* http://translationjournal.net/journal/63mtquality.htm