The Journal of Social Sciences Research
ISSN(e): 2411-9458, ISSN(p): 2413-6670
Special Issue. 1, pp: 377-381, 2018
URL: https://arpgweb.com/journal/journal/7/special_issue
DOI: https://doi.org/10.32861/jssr.spi1.377.381

Academic Research Publishing Group

Original Research Open Access

# Lingo-Stylistic Analysis of Statistical and Neural Machine Translation

**Guzel R. Eremeeva***
Kazan Federal University, Russia

**Polina V. Antonova**
Kazan Federal University, Russia

**Marat A. Yahin**
Kazan Federal University, Russia

## Abstract

The urgency of the problem under investigation is caused by the increasing popularity of machine translation for the solution of various kinds of communicative tasks. The purpose of the article is to compare statistical and neural machine translation systems. The leading approach to the study of this problem was the linguo-stylistic analysis of linguistic material using software from Microsoft Translator. The main results of the article consist in a comparative analysis of the translation of simple and complex texts through statistical and neural machine translation systems, which led to the conclusion that the greatest number of errors is associated with the translation of semantic constructions. Materials of the article can be useful to the experts working in the field of machine translation, to students and all who are connected with computer linguistics.

**Keywords:** Language; Russian; English; Computer linguistics; Research.



## 1. Introduction

Machine translation is one of the most intensively developing areas of computer linguistics, an inevitable attribute of modern interlingual communication. The term machine translation itself can be interpreted both in a narrow and broad sense. In the narrow sense, machine translation is an action performed on a computer, which is aimed at converting text in one natural language into an equivalent text on the other. For example, the input of a machine is given text in one language, and the output is text on the other, and the transformation of the text will occur without human assistance. In a broad sense, machine translation can be regarded as a science at the junction of linguistics, cybernetics and mathematics (Mukesh *et al.*, 2010). In this case, the goal will be to build a system that will allow machine translation to be implemented in a narrow sense. Until 1980, the main basis for research in the field of machine translation was the creation of linguistic rules, syntactic analysis, as well as rules for the generation and transmission of syntactic units. A major role in the development of methods and approaches in the field of machine translation was played by the works of the linguist (Chomsky, 1956). Later it led to the formation of new methods used in modern machine translation systems (Radek *et al.*, 2017; Sakaeva and Takhtarova, 2016)

Complexity and comprehensiveness of machine translation is the main stimulus to the development of works in this field. The attention of theorists in this case attracts the possibility of testing hypotheses about the structure of language levels and the effectiveness of the proposed algorithms. In consequence new theories of formalization of language data and automation of translation arise. To develop high-performance and efficient machine translation systems that could contribute to the solution of a number of communicative tasks, it is necessary to have an accurate idea of the principles and structure of machine translation systems. The aim of our study is to compare the statistical and neural machine translation systems and to present the linguistic analysis of linguistic material by means of the machine translation systems under study (Kalinink, 2017).

## 2. Methodology

As a linguistic material for the study, two different types of texts of varying degrees of complexity were taken. As the first source, a passage was selected, consisting of 8 short sentences and 69 words. This passage was a description of the contents of the user's guide. The second source was a passage from a scientific text. The passage considered included 4 sentences and 72 words. The sentences used in this article were longer than in the first case, the structure of the sentences was more complicated, and for these reasons, there were more moments in the translation process, where the machine translation systems being compared could make mistakes. The tool for comparing selected machine translation systems was software from Microsoft Translator. In order to assess the quality of the translation of the two texts proposed, the quality of the transfer of the information of the sentences was considered by comparison with the reference translations (He *et al.*, 2016; Tastemirova *et al.*, 2018).

*Corresponding Author

## 3. Results

The results of the comparison of linguistic material through statistical and neural machine translation systems are presented below.

Russian reference translation:

Типичное руководство пользователя содержит:

- Введение, содержащее ссылки на связанные документы и информацию о том, как ориентироваться в руководстве пользователя;
- Страницу содержания;
- Главы, описывающие, как использовать, по крайней мере, наиболее важные функции системы;
- Главу, описывающую возможные проблемы и пути их решения;
- Часто задаваемые вопросы и ответы на них;
- Глоссарий и, в больших документах, предметный указатель.

English source text:

A typical user manual contains:

- An introduction that contains links to related documents and information on how to navigate the user guide;
- The contents page;
- Chapters describing how to use at least the most important functions of the system;
- A chapter describing possible problems and ways to solve them;
- Frequently asked questions and answers to them;
- Glossary and, in large documents, an index.

The results of comparison of translations of plain text using statistical and neural machine translation systems are given in Table 1:

**Table-1.** Comparative analysis of translation results in case of plain text

| Statistical machine translation | Neural machine translation |
|---|---|
| Типичная инструкция содержит:<br>• Введение, содержит ссылки на соответствующие документы и информацию о том, как ориентироваться в руководстве пользователя;<br>• Содержимое страницы;<br>• Разделы, описывающие, как использовать по крайней мере наиболее важных функций системы;<br>• Глава, описывающие возможные проблемы и способы их решения;<br>• Часто задаваемые вопросы и ответы на них;<br>• Глоссарий и в больших документов, индекс. | Типичное руководство пользователя содержит:<br>• Введение, содержащее ссылки на соответствующие документы и информацию о том, как ориентироваться в руководстве пользователя;<br>• Страница содержания;<br>• Главы, описывающие, как использовать по крайней мере наиболее важные функции системы;<br>• Глава, описывающая возможные проблемы и способы их решения;<br>• Часто задаваемые вопросы и ответы на них;<br>• Глоссарий и, в больших документах, индекс. |

It can be noted that most of the sentences in the translation made by the neural machine translation express all the necessary information, while in the translation performed by the statistical machine translation system a number of syntactic errors arise. Table 2 shows the number of mistakes made when translating the user's manual by the neural and statistical systems (Koehn, 2014).

**Table-2.** The number of mistakes in statistical and neural machine translation systems relative to general categories in the case of plain text

| Mistakes of the statistical machine translation system | Mistakes of the neural machine translation system |
|---|---|
| • Semantic: 66,7%<br>• Syntactic: 33,3 % | • Semantic:100%<br>• Syntactic: 0% |

Consider now a more complex text.

Russian reference translation:

Первый День Земли был отмечен в США. Он был основан сенатором Соединенных Штатов Гэйлордом Нельсоном как диспут-семинар – вид общеобразовательного форума или семинара. Это было 22 апреля 1970 года. Хотя этот первый День Земли был сосредоточен на Соединенных Штатах, организация, «запущенная» Дэнисом Хейесом, который был первым национальным координатором в 1970 году, сделала его международным в 1990 году, и организовала мероприятия в 141 стране. English source text : The first Earth Day was held in the USA. It was founded by United States Senator Gaylord Nelson as an environmental teach-in – a sort of general educational forum or seminar. That was on April 22, 1970. While this first Earth Day was focused on the United States, an organization launched by Denis Hayes, who was the original national coordinator in 1970, took it international in 1990 and organized events in 141 nations. The results of comparing complex text translations through statistical and neural machine translation systems are given in Table 3:

**Table-3.** Comparative analysis of translation results in case of complex text

| Statistical machine translation | Neural machine translation |
|---|---|
| В США был проведен первый день земли. Он был основан Соединенных Штатов Сенатор Гейлорд Нельсон как экологической– своего рода общий образовательный форум или семинар. Это было 22 апреля 1970 года. Хотя этот первый день земли было сосредоточено на Соединенные Штаты, Организация запустила Денис Хайес, который был первоначально Национальный координатор в 1970 году, взял его международных в 1990 году и организованных мероприятий в 141 Наций. | Первый день земли состоялся в США. Он был основан сенатором Соединенных Штатов Америки, как экологическое обучение-своего рода общий образовательный форум или семинар. Это было 22 апреля 1970. Хотя этот первый день земли был сосредоточен на Соединенных Штатах, организация, запущенная Денис Хейс, который был первоначальным национальным координатором в 1970, взял его международным в 1990 и организовал события в 141 Наций. |

The main type of mistakes made by the statistical machine translation system is associated with a misinterpretation by the system of automatic translation of relations within sentences and word combinations, with violations of the rules of the Russian language. Grammar mistakes in translation can be divided into several types: syntactic (incorrect construction of sentences), incorrect use of word forms, mistakes in the matching of cases. The system of neural machine translation allowed much fewer errors in the order of words, 20% less lexical and 50% grammatical errors. Table 4 shows the number of mistakes made in the translation of a complex text by the neural and statistical systems (Nikitin, 2012).

**Table-4.** The number of mistakes of statistical and neural machine translation systems in relation to general categories in case of complex text

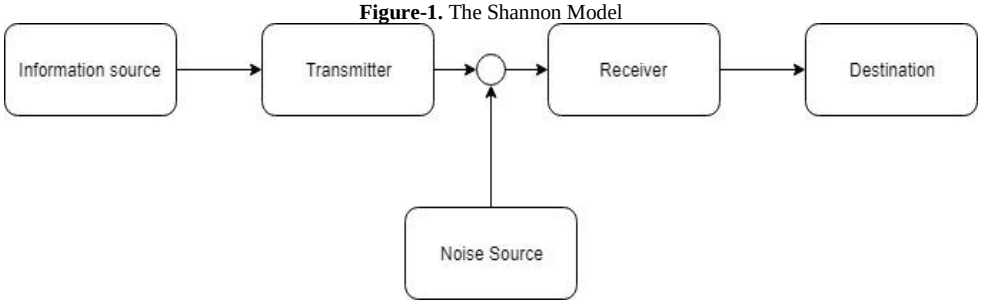
| Mistakes of the statistical machine translation system | Mistakes of the neural machine translation system |
|---|---|
| • Semantic: 28,6%<br>• Syntactic: 71,4 % | • Semantic:25%<br>• Syntactic: 75% |

After considering the translation texts of both systems, it turned out that when translating the text, which was a description of the contents of the manual, the system of neural machine translation made only 1 error, while the statistical machine translation system allowed 3 errors. In order to compare the systems, 1 and 3 were equated to 100 percent. In the case of a more complex text, the neural machine translation system made 4 errors, and the statistical machine translation system allowed 7 errors. Similarly, to the first case, 4 and 7 were equal to 100 percent. In the context of the translation of both texts, it turned out that of the two machine translation systems presented, the statistical machine translation system allowed more errors in both proposed texts (Aubakirova *et al.*, 2018).

## 4. Discussion

Statistical machine translation is a machine translation technique that uses a comparison of a large volume of language pairs, just like machine translation, which is based on examples. The basis of statistical machine translation is the search for the most probable translation of the expression, using data from bilingual corpus of texts. Thus, statistical machine translation does not operate with linguistic algorithms, but calculates the probability of applying an expression. An expression that has an optimal probability is considered most appropriate for the translation of the input text. In statistical machine translation, in contrast to the systems discussed above, the task of translating text is not put. In this case, the task is to decipher the input text. It is assumed that the text written in English is the text in Russian, but the text itself is encrypted. Shannon proposed a mathematical model of connection. The meaning of this model was as follows (Fig. 1).

**Figure-1.** The Shannon Model




The model itself consists of 5 elements, shown in Figure 1, which are arranged linearly. The transmitter encodes the information it receives from the source and transmits it to the channel, and the data is sent to the receiver through the channel, where the final destination is decoded and transmitted. The channel is affected by noise - interference, distorting the information. According to the Shannon model, creating redundant information, it is possible to restore the original data with a high probability. To translate the input text, it is necessary to find a decoding method that uses redundancy, because of which encoding should be probabilistic. The decoding task will be to find the original message, which corresponds to the highest probability. To do this, you need to find for two messages the conditional probability that the translation of the message, passing through the noisy channel, is converted to the original

message. In this case it is necessary to have a model of the language (source) and model of translation (channel). Thus, in the case of translation from English into Russian, the translation is a phrase search process that will maximize the product of the unconditional probability of the expression in Russian and the probability of the English original expression, provided this expression is in Russian. Mathematically, the foregoing is as follows:

$$\max_{f_e} P(f_e|f_r) = \max_{f_e}\big(P(f_e)P(f_r|f_e)\big),$$

Where $f_e$ – the phrase of translation (Russian), $f_r$ – the phrase of the original (English).

It should also be noted that statistical machine translation has the property of self-learning. The more the language pairs are at the disposal of the system, the better the translation result will be. Training pairs can be found in the materials of the Workshop on Statistical Machine Translation or the International Workshop on Spoken Language Translation. Advantages of statistical machine translation are the following:

• Self-learning;
• Smoothness of the translation;
• Quick setup (compared to a rule-based translation system);

The systems of statistical machine translation also have a number of shortcomings, namely:

- A large number of grammatical errors;
- Instability and unpredictability of translation;
- The need for parallel large-volume casings.

Neural machine translation is a new approach to solving the problems of machine translation, and has become widespread in recent years. This approach is based on the use of neural networks, which in their structure are similar to the structure of the human brain, in which the signal propagates through successive layers of elements that mimic neurons. The main advantage of this system, similarly to the system of statistical machine translation, is the possibility of self-learning. To create systems of neural machine translation a model of the codec-decoder type is used. The goal of neural machine translation is to create a fully trainee model, where each component of the model will be tuned on the basis of training buildings, in order to improve the quality of translation. Currently, most of the systems of neural machine translation is based on a similar structure: an encoder, a decoder and a tracking mechanism. In this structure, the encoder converts each word into a context vector that, with the help of the decoder, turns into the word of the target language. The purpose of the tracking mechanism is to monitor the accuracy of the translation. The main problem of neural machine translation is the inability to translate rare words. At the moment, to improve the accuracy of neural machine translation several approaches are proposed (Dorofeeva, 2013).

## 5. Summary

This study showed that when translating selected texts, the proposed systems do not always manage to extract the information necessary for correct translation of the text. Thus, it can be concluded that two popular machine translation systems do not yet allow satisfactory results. Obviously, to improve the quality of machine translation, it is necessary to carry out more extensive non-automatic assessments, involving professional translators to inform system developers about existing shortcomings. There is an opinion that machine translation systems, combining both empirical systems and rules-based systems (hybrid systems), make it possible to provide the most qualitative translation. However, this topic deserves a much deeper and more detailed consideration and is not considered in this paper (Malpica et al., 2016).

## 6. Conclusions

Natural languages are incredibly complex for machine processing. In the course of processing at the word level, the machine translation system may encounter the problem of synonyms; at the syntactic level, the system does not always succeed in determining the relationship between lexical units in the text. With the development of statistical and neural machine translation systems, the quality of machine translation has improved noticeably, and new opportunities have emerged for more precise linguistic research.

## Acknowledgements

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

## References

Aubakirova, A., Ongarbayeva, A., Sebepova, R., Karimova, B. and Mirzabekova, M. (2018). Synergetic approach in trilingua education of the republic of Kazakhstan. *Opción,*, 34(85).

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory,* 2(2): 113-24.

Dorofeeva, I. V., 2013. "Shannon-weaver model and its significance for the development of the theory of communication. Language discourse in social practice." In *Collection of scientific papers of the International Scientific and Practical Conference. Federal State Budget Educational Institution of Higher Professional Education. Tver State University.* . pp. 49-53.

He, W., Wu, H. and Wang, H., 2016. "Improved neural machine translation with SMT features." In *Thirtieth AAAI conference on artificial intelligence.* pp. 151–57.

Kalinink, S. (2017). Review of modern approaches to improving the accuracy of neural machine translation. Rem. 70-79.

Koehn, P. (2014). *Neural machine translation. Computation and language.* Cornell University Library. 1-9.

Malpica, p. H., villalobos, j. V., zamorano, m. M. and juvinao, j. M. (2016). Racionalismo emergente en la gerencia universitaria: Factor de humanización en universidades de colombia, venezuela y méxico. Revista espacios| 37(30).

Mukesh, G. S., Vatsa, N. J. and Sumit, G. (2010). Statistical Machine Translation. *DESIDOC Journal of Library & Information Technology,* 30(4): 25-32.

Nikitin, I. (2012). *Distributed software and information support of the statistical model of the translation of natural languages.* Moscow Aviation Institute. 105.

Radek, S., Roman, S. and Zuzana, K., 2017. In *Cybernetics and Mathematics Applications in Intelligent Systems Proceedings of the 6th Computer Science On-line Conference 2017 (CSOC2017). 2.*

Sakaeva, L. R. and Takhtarova, S. S., 2016. "Features of the term translation (based on the material of English, German and Russian languages)." In *Materials of the international scientific and practical conference Institutionalization of innovative educational environment of higher educational school. Kazakhstan.* pp. 405-10.

Tastemirova, G., Naraliyeva, R., Begalieva, A., Tileuzhanova, G., Sapayeva, G., Ibragimova, M. and Dosanov, B. (2018). Literary criticism of turkish literature. *Astra Salvensis*.