3-бөлім	**Раздел 3**	**Section 3**
Информатика	**Информатика**	**Computer Science**

IRSTI 16.31.21 DOI: https://doi.org/10.26577/JMMCS.2021.v111.i3.11

A. Karibayeva, V. Karyukin, A. Turgynbayeva, A. Turarbek*
Al-Farabi Kazakh National University, Kazakhstan, Almaty
*e-mail: turarbek_asem@mail.ru

THE TRANSLATION QUALITY PROBLEMS OF MACHINE TRANSLATION SYSTEMS FOR THE KAZAKH LANGUAGE

Kazakh language is related to languages with complex morphology and syntax. Today, most machine translation systems consider Kazakh language too, like Google, Yandex, Prompt, etc. This article describes the errors, shortcomings, and problems of machine translation (MT) into the Kazakh language. To analyze errors in machine translation into the Kazakh language, the most popular electronic translation programs were selected. When translating from Russian and English into Kazakh (and vice versa), various errors may occur, since the Kazakh language is different from other languages, and has special characteristics. To compare the results, an empirical method was used, namely, monitoring and testing the translation results of machine translation systems. Considering the results of statistical methods, the rule-based and neural networks based methods in machine translations were also analyzed. The practical significance of the study lies in the development of recommendations for the identification and elimination of errors when editing the results of MT. The scientific significance of the study lies in the fact that for the first time errors and inaccuracies arising from machine translation of the Kazakh language have been systematized. The assessment of the quality of MT is also presented. The research carried out in this article will be used for the post-editing problem in machine translation.

Key words: Machine translation, systems of machine translation, RBMT, SMT, NMT, Kazakh language, quality of translation.

А. Кәрібаева, В. Карюкин, А. Тұрғынбаева, Ә. Тұрарбек*
Әл-Фараби атындағы Қазақ ұлттық университеті, Қазахстан, Алматы қ.
*e-mail: turarbek_asem@mail.ru

Қазақ тіліне арналған машиналық аудармa жүйелерін аудару сапасы мәселелері

Казақ тілі күрделі морфологиясы мен синтаксисі бар тілдерге жатады. Бүгінгі таңда машиналық аударма жүйелерінің көпшілігі қазақ тілін, мысалы Google, Яндекс, Prompt және т.б. қолданады. Қазақ тіліне машиналық аудармадағы (МА) қателерді анықтау үшін ең танымал электрондық аударма бағдарламалары таңдалды. Орыс және ағылшын тілдерінен қазақ тіліне (және керісінше) аударған кезде әртүрлі қателіктер туындауы мүмкін, өйткені қазақ тілі басқа тілдерден өзгеше және ерекше сипаттамаларға ие. Нәтижелерді салыстыру үшін эмпирикалық әдіс қолданылды, атап айтқанда машиналық аударма жүйелерінің аударма нәтижелерін бақылау және тестілеу. Статистикалық әдістердің нәтижелерін ескере отырып, ережеге негізделген әдістер мен машиналық аудармалардағы нейрондық желілерге негізделген әдістер де талданды. Зерттеудің практикалық маңыздылығы МА нәтижелерін өңдеу кезінде қателерді анықтау және жою бойынша ұсыныстар әзірлеу болып табылады. Зерттеудің ғылыми маңыздылығы қазақ тілін машиналық аудару кезінде туындайтын қателер мен дәлсіздіктер алғаш рет жүйелендірілгендігінде. Сондай-ақ, МА сапасын бағалау ұсынылған. Осы мақалада жүргізілген зерттеу машиналық аудармада пост-редакциялау мәселесін шешу үшін қолданылады.

Түйін сөздер: Машиналық аударма, машиналық аударма жүйелері, RBMT, SMT, NMT, қазақ тілі, аударма сапасы.

А. Карибаева, В. Карюкин, А. Тургынбаева, А. Турарбек*
Казахский национальный университет имени аль-Фараби, Казахстан, г.Алматы
*e-mail: turarbek_asem@mail.ru

Проблемы качества перевода систем машинного перевода для казахского языка

Казахский язык относится к языкам со сложной морфологией и синтаксисом. Сегодня большинство систем машинного перевода также рассматривают казахский язык, например Google, Яндекс, Prompt и т. д. В данной статье описаны ошибки, недостатки и проблемы машинного перевода (МП) на казахский язык. Для анализа ошибок в машинном переводе на казахский язык были отобраны наиболее популярные программы электронного перевода. При переводе с русского и английского языков на казахский (и наоборот) могут возникать различные ошибки, так как казахский язык отличается от других языков и имеет особые характеристики. Для сравнения результатов был использован эмпирический метод, а именно мониторинг и тестирование результатов перевода систем машинного перевода. С учетом результатов статистических методов были также проанализированы методы, основанные на правилах, и методы, основанные на нейронных сетях, в машинных переводах. Практическая значимость исследования заключается в разработке рекомендаций по выявлению и устранению ошибок при редактировании результатов МП. Научная значимость исследования заключается в том, что впервые систематизированы ошибки и неточности, возникающие при машинном переводе казахского языка. Также представлена оценка качества МП. Исследование, проведенное в этой статье, будет использовано для решения проблемы постредактирования в машинном переводе.

Ключевые слова: Машинный перевод, системы машинного перевода, RBMT, SMT, NMT, казахский язык, качество перевода.

1 Introduction

The modern world and our future are completely dependent on applied intelligent systems, as new technologies are developing every day. One of the tasks of intelligent systems is machine translation from one natural language to another. Machine translation (MT) allows people to communicate regardless of language differences, as it removes the language barrier and opens up new languages for communication. Machine translation is a new technology, a special step in human development. This type of translation can help when you need to quickly understand what your interlocutor wrote or said in a letter. Of course, the quality of such a translation is very low (for some groups of languages), but in most cases the main meaning can be understood.

The Kazakh language is an agglutinative language with a complex nominative (morphological and syntactic) participation of polysyntheticism. Due to the development of our country at the global level from year to year and the growth of external relations, various translation programs are widely used when translating into Kazakh or from Kazakh into other languages. Various MT systems still cannot translate completely correctly and there are translation errors, but the field of machine translation is much more developed than in previous years. Analysis of the results of machine translation into the Kazakh language, where errors and inaccuracies are analyzed, is an actual task for the task of natural language processing.

Errors and inaccuracies should be distinguished in the results of the machine translation. Inaccuracies are associated with the stylistic incorrectness of the translated sentence. They do not interfere with its understanding, however, they require editing when creating the text.

Errors occur in the case of incorrect definition of grammatical forms, in turn, they impede the understanding of the text, and for their editing it is necessary to analyze the original sentence again.

Methods for identifying and correcting errors are individual for each type of machine translation(MT). Based on the identification of basic errors and inaccuracies in a representative sample of texts, changes can be made to the algorithms of the MT system. Establishing errors for machine translations of Kazakh language and making recommendations to the editor allows to speed up and simplify the post-editing process in MT.

In this paper, we will consider most problems of translation Kazakh texts and evaluate MT translation quality.

2 Related works

The world's first automated translation was carried out in the USA in 1954. The first automatic translation systems allowed the translation of around 300 words between the Russian and English languages. The dictionary-based direct translation approach was used in those systems where each word from the source language matched the corresponding word in the target language. Although that approach was straightforward and computationally cheap, the output results were inferior.

In the 1970s and 1980s, the research focused on the rule-based machine translation approach relying on a large number of built-in linguistic rules and thousands of bilingual dictionaries. In addition, this approach requires lexicons with morphological, syntactic, and semantic information. The translation systems implement all these complex rules to transfer the source language into the target one.

In the 1990s, a significantly more efficient approach that uses the corpus-based architecture started its development. The statistical translation system utilizes the probabilities to choose the most appropriate translation from their comparison to the aligned bilingual corpus, breaking down the source text into segments.

The fast development of the Internet and communication systems increased the volume of information in various languages, significantly increasing qualitative translations' demand. Human translators were unable to deal with this enormous stream of data. New huge investments were made to the development of machine translation systems for global and private organizations. New hybrid systems combining rule-based and statistical architectures were widely introduced. The goal of these systems was to increase the accuracy of machine translation essentially.

The recent developments in machine translation incorporated a deep neural network approach to improve machine translation quality further. The service providers offer new customized machine translation engines that can analyze texts in specific scientific domains, such as engineering, information technologies, life science, economics, etc. Some translation systems formed into widespread online translators such as Google Translate, Yandex, Tridentsoftware, etc.

Therefore, the quality of online translations improves every year. However, machine translation in many languages still has many problems during the translation of complex sentences. In this regard, for many years, scientists from different industries have been working to improve the quality of MT. The Kazakh language was added to Google Translate in 2014.

Nevertheless, today the problem of MT for the Kazakh language is very relevant. Indeed, new terms and phrases regularly appear in Kazakh dictionaries since borrowed words, including international ones, are translated into the Kazakh language. It creates additional problems in translation, displaying errors or not translating indefinite words at all, so any translation needs mandatory editing by another translator or a specialist in this field.

The formal grammatical models of simple sentences and the first version of the MT program from Kazakh into English were considered in (Zhumanov and Tukeyev, 2009; Tukeyev and et., 2010). In addition, a multivalued method for translating morphologically complex natural languages such as Russian and Kazakh (Tukeyev and Rakhimova, 2012; Tukeyev and et., 2013; Tukeyev, 2014) has been developed. The work to research and develop a neural machine translation (NMT) system for the Kazakh language has been underway since 2018. Over the past three to four years, the theory and practice of MT have expanded significantly, and a new direction of MT has been created, raising the quality standard for MT to new heights.

The problem of MT post-editing for the Kazakh language found its place in works . Automatic post-editing allows improving the quality of translation efficiently. In (Abeustanova and et., 2017), incorrect words in the translated sentences from English to Kazakh are found using the maximum entropy model. (Shormakova and et., 2019) proposes a method for determining incorrect words in the translated texts. These words are found by comparing a right target sentence and a translated sentence from the source English language to the target Kazakh language. When the incorrect words are identified, they are replaced with the most suitable ones.

The technology for solving the problem of unknown words in neural machine translation (NMT) is described in (Turganbayeva and et., 2020). The unknown words are replaced with their synonyms in the dictionary. The quality of NMT is increased by applying the segmentation method based on the complete set of endings (CSE) proposed by (Tukeyev and et., 2020).

This work aims to study the problem of the quality of MT of the Kazakh language. The quality of the translation depends on the subject matter and style of the source text and the grammatical, syntactic, and lexical affinity of the languages between which the translation is done. Having identified the principal errors of translation into the Kazakh language and vice versa, it may be possible to determine the quality problems of MT.

3 Methodology of machine translation

Machine translation (MT) is an automated translation, where software is used to translate a text or phrases from one natural language into a second language. The following approaches in machine translation are rule-based (RBMT), statistical (SMT), and neural machine translation. The RBMT uses knowledge of the language, like structural, morphological, lexical rules of determining languages. So, if you do not know all the grammar and syntax of the language, you can not cover all the rules to create MT. Herein, mostly used the Hidden

Markov model (HMM) to predict part of speech of a word (1):

$$\arg\max_{t_1,\dots,t_n} \prod_{t=1}^{n} p(w_i|t_i)p(t_i|t_{i-1}) \tag{1}$$

where:

t – tag (noun, pronoun, adjective, etc.); w – words in the text; $p(w|t)$ – the probability of w correspond to tag t; $p(t_1|t_2)$ – the probability of t_1 goes after t_2.

At the heart of the statistical machine translation system is the comparison of the text of large language pairs. This type of system is based on probability and uses statistical translation models. The Bayesian theorem applied to the probability approach is as follows (2):

$$P(T|S) = P(S|T)P(T), \tag{2}$$

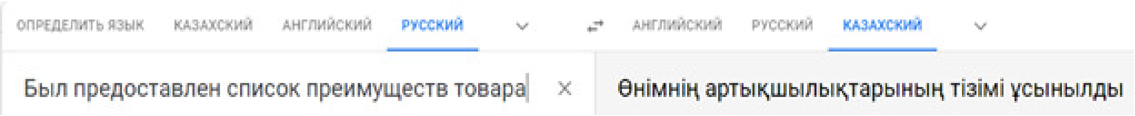
where $P(T|S)$ is the probability that the string in the source language is a translation of the string in the target language, and $P(T)$ is the probability that the string in the target language is obtained. The process of creating a statistical translation model is a bit faster, but the technology here largely depends on the volume of parallel corpora.

Today, NMT is often used in machine translation as a trend. NMP is based on the construction of large neural networks and computations. Ehe process in the NMT is divided into two phases. In the first, each word of the original sentence is passed through an "encoder" which generates what we call the "original context" based on the current word and the previous context. The translation is completed when the decoder reaches the stage of generating the actual last word in the sentence.

4 The translation problems and quality for Kazakh language in MT

The study of the problems of the quality of translation into the Kazakh language is very relevant since the development of our country at the global level and the growth of external relations. There is a need for translation into the Kazakh language or from the Kazakh language into other languages for various segments of the population and industry. The difference between the Kazakh language and the other languages is that it has special characteristics: proximity of the lexical structure, the harmony law, agglutination (a series of affixes), the lack of a category, the lack of auxiliary words (prepositions), and special word order.

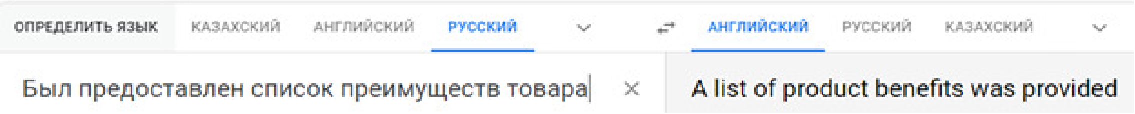
Therefore, a pilot study, consisting of several stages, was carried out to determine the quality of MT. First, several machine translation systems (MTS) were taken, and MT was done. They are presented in Fig. 1 – 4



Был предоставлен список преимуществ товара

Өнімнің артықшылықтарының тізімі ұсынылды

Figure 1: An example of translation from the Russian language to the Kazakh language

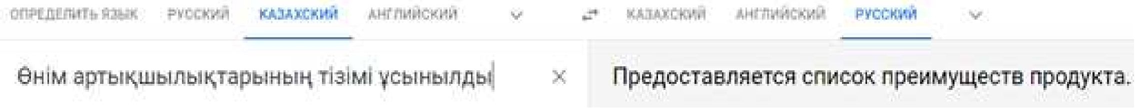


Был предоставлен список преимуществ товара

A list of product benefits was provided

Figure 2: An example of translation from the Russian language to the English language

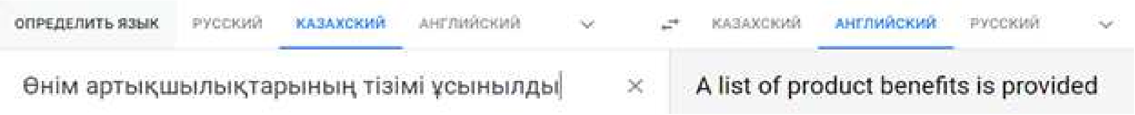
After that, the translation of the sentence "Өнім артықшылықтарыны1 тізімі ұсынылды" from the Kazakh language into the Russian and the English languages was also carried out. Although the translation was done correctly, there were translation errors associated with the ambiguity of words. They were displayed in Fig. 3 – 4



Өнім артықшылықтарының тізімі ұсынылды

Предоставляется список преимуществ продукта.

Figure 3: An example of translation from the Kazakh language to the Russian language



Өнім артықшылықтарының тізімі ұсынылды

A list of product benefits is provided

Figure 4: An example of translation from the Kazakh language to the English language

AAs a result of the translation, it was possible to identify various translation errors that were associated with the classification of errors (Rakhimova and et., 2020).

The second stage of the study consisted of determining the development of the selected SMT. In order to study the evolution of the SMT, the same texts were translated with three online translators in June 2020 (Table 1) and March 2021 (Table 2). Also, one of the translators was a human. The less the editor had to correct the text, the better the system was. If all the translations had to be rewritten, the MT was ineffective. Thus, for each studied fragment, there were several MT options for verification and quality assessment. The results of a comparative analysis of transfers with a difference of nine months allow us to draw the following conclusions:

1. The Yandex SMT is developing more actively than the other studied SMTs (32% of changes). Now, it has the ability to translate from the Kazakh language written in Latin. However, when comparing the translation results, it was noted that the quality of the output text changed during translation, and previously absent spelling errors appeared.

2. The translation of the Google Translate SMT has undergone the smallest changes (2%), but the important thing is that the quality of the translation is improving.
3. The changes in the translation of the Tridentsoftware SMT were not made. The text remained unchanged.

Table 1. Translation results in 2020

Желіде кемшіліктер бар болған Артықшылықтардың тізімі көрсетілген еді Биіктен аққан ақ сәуле Арналарымыздың түрілері шектеулі Қанаттарымыздың ұзындығы анықталмаған	Были проблемы с сетью Был предоставлен список преимуществ Белый свет течет сверху Типы наших каналов ограничены Длина наших крыльев неизвестна	Есть недостатки в сети Был показан список преимуществ Ао утечки излучения с высоты Ограниченные типы каналов Длина крыльев не определена	Недостатки есть в сети Список преимуществ было указана в Ао утечки излучения с высоты С Арналарымыздың түрілері Длина Қанаттарымыздың не установлена

Table 2. Translation results in 2021

Желіде кемшіліктер бар болған Артықшылықтардың тізімі көрсетілген еді Биіктен аққан ақ сәуле Арналарымыздың түрілері шектеулі Қанаттарымыздың ұзындығы анықталмаған	Были проблемы с сетью Был предоставлен список льгот Белый свет льется сверху Типы наших каналов ограничены Длина наших крыльев неизвестна	В Сети появились недочеты Список излишеств был указан Белый луч с высоты С Арналарымыздың түрілері Длина крыльев не определена	Желіде кемшіліктер бар болған Артықшылықтардың тізімі көрсетілген еді Биіктен аққан ақ сәуле Арналарымыздың түрілері шектеулі Қанаттарымыздың ұзындығы анықталмаған

The comparative analysis of translations showed that all SMTs are lexically developing, and the quality of translation is improving. Nevertheless, the translation problems still remain. In terms of the best translation, Google Translate leads the way.

The third stage of the research was to find ways to improve the quality of the translation. For this purpose, we use MT in translation projects, compare different SMTs, evaluate the translation in order to choose the best system.

The translation company Pairaphrase has identified how the quality of MT can be improved. They distinguish two approaches. The first is based on changing the way the input text is written. It is implemented by using short sentences, the structure of which should be simple, rare adverbs and not utilizing slang, complex and ambiguous words. The second way is related to improving the MT engine. It implies the use of software that includes a translation memory. Translation memory is a key component of any translation training tool (Pairaphrase, 2020). It has been around for over 30 years and represents a way for the translation industry to reuse previous translations to improve the quality of user translations over time.

Today, the best SMTs are based on Neural machine translation (NMT). The research and practical application of NMT in a professional environment provide reviews and comparative characteristics of NMT and Statistical machine translation (SMT) in terms of quality. In a study (Koehn and Knowles, 2017; Koehn, 2017), the NMT system was found to outperform the SMT system in a study involving training data of 15 million words. These authors noted problems with NMT, including domain mismatch and the use of rare words. The conducted review of the results of modern MT systems can be used as a reference material when choosing a system for use in a professional translation service or for personal use.

5 Results and discussion

In this paper, MT is considered to be the process of translating some text from one natural language into another, namely in the case of Kazakh language. The main advantages of MT are its speed and low cost. Currently, there is a large number of MT systems. The research identified that the most popular MT systems among users are Tridentsoftware, Yandex, and Google. Tridentsoftware uses the rule-based MT, unlike Google, which until recently used a statistical translation method. In March 2017, Google completely switched to neural networks to improve the quality of the output text. Yandex has implemented a hybrid system that can choose between neural and statistical MT models.

The consideration of the methods for assessing the effectiveness of the MT systems led to the conclusion that the variety of approaches and methods for assessing the quality of MT indicates ongoing research in this area and the absence of a single standard for determining the effectiveness of existing systems.

The results of the research can be used to improve the quality of machine translation systems from Kazakh into another language, and into Kazakh from another source language; they will also be useful in the preparation of text for MT, as well as in translation training. The received results will be used in further research, namely in training neural MT based on linguistic features of Kazakh language and in the task of post-editing. Post-editing – human text processing after receiving MT texts or sentences.

References

[1] Zhumanov Zh.M., Tukeyev U.A., "Development of machine translation software logical model (translation from Kazakh into English language)", *Reports of the Third Congress of the World Mathematical Society of Turkic Countries. Edited by Academician Bakhytzhan T. Zhumagulov* 1 (2009): 356-363.

[2] Tukeyev U., Zhumanov Zh., Rakhimova D., "Features of development for natural language processing", *In the book "ICT - from theory to practice" edited by M. Milosz. Polish Information Processing Society, Lublin* (2010): 149-174.

[3] Tukeyev U., Rakhimova D., "Augmented attribute grammar in meaning of natural languages sentences", *The 6th International Conference on Soft Computing and Intelligent Systems, and the 13th International Symposium on Advanced Intelligent Systems, SCIS-ISIS2012/Kobe, Japan* (2012): 1080-1085.

[4] Tukeyev U.A., Rakhimova D.R., Baisylbayeva K., Umirbekov N., Orazov B., Abakhan M., Kyzyrkanova S., "Kopmagynalyk beineleu keste tasili negizinde orys tilinen kazakh tiline mashinalyk audarmasynyng morfologiyalyk analizben sintezin kuru [Synthesis of morphological analysis of machine translation from Russian to Kazakh on the basis of the method of ambiguous mapping]", *In proceedings of the I International Conference on Computer processing of Turkic Languages* (2013): 182-191.

[5] Tukeyev U.A., "Razrabotka tehnologii mashinnogo perevoda na osnove metoda mnogoznachnyh otobrazhenii dlya morfologicheski slojnyh yazykov [Development of machine translation technology based on the multivalued mapping method for morphologically complex languages]", *Proceedings of the 4th International Scientific and Practical Conference "Informatization of Society"* (2014): 130-132.

[6] Abeustanova A., Tukeyev U., "Automatic Post-editing of Kazakh Sentences Machine Translated from English", *In Advanced Topics in Intelligent Information and Database Systems. ACIIDS 2017. Studies in Computational Intelligence, Springer* 710 (2017): 283-295.

[7] Rakhimova D.R., Turarbek A.T., Karyukin V., Karibayeva A., Turganbayeva A., "Kazakh tiline arnalgan zamanaui mashinalyk audarma tehnologiyalaryna sholu [Review of modern machine translation technologies for the Kazakh language]", *Bulletin of KazNRTU named after K. Satpayev. Technical science* 5 (141) (2020): 103-109.

[8] Shormakova A., Zhumanov Zh., Rakhimova D., "Post-editing of words in Kazakh sentences for information retrieval", *Journal of Theoretical and Applied Information Technology* 97 (6) (2019): 1896-1908.

[9] Turganbayeva A., Tukeyev U., "The solution of the problem of unknown words under neural machine translation of the Kazakh language" , *Journal of Information and Telecommunication* (2020): 214-225.

[10] Tukeyev U., Karibayeva A., Zhumanov Z., "Morphological segmentation method for Turkic language neural machine translation" , *Cogent Engineering* 7(1) (2020): 1-16.

[11] Pairaphrase, "Two Approaches for How to Improve Machine Translation Quality" , *https://www.pairaphrase.com/how-to-improve-machine-translation-quality* (2020).

[12] Philipp Koehn, Rebecca Knowles, "Six Challenges for Neural Machine Translation" , *Proceedings of the First Workshop on Neural Machine Translation* (2017): 28-39.

[13] Philipp Koehn, *Statistical Machine Translation. Draft of Chapter 13. Neural Machine Translation* (arXiv:1709.07809v1 [cs.CL], 2017): 117.