International Journal of Trend in Scientific Research and Development (IJTSRD)
Advancements in Multidisciplinary Research and Analysis – December 2023
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Large Language Models for Machine Translation in the Example of Uzbek-English

Khonkulova Nilufar Ravshanovna

English Applied Translation Department, Translation Faculty
Uzbekistan State World Languages University

ABSTRACT

The article consider about the language models for machine translation in the example of Uzbek-English. Because of the emergence of a large number of scientific and technical documents and the need for their prompt translation into other languages. There are described the principles of operation, the main types of machine translation systems, studied the features of linguistic means of scientific style, found out which of the linguistic means remain in the text after its translation and compared the text translated by a machine and a professional translator.

KEYWORDS: *information technology, professional translator, means of communication, automatic translation, translation tools, electronic dictionary, language models, operating principle.*

We live in a world of information technology, which has become a part of our lives. We use modern means of communication. The computer has become an integral element of our lives, not only in the workplace but also in everyday life. The rapid development of new information technologies testifies to the ever-increasing role of computer technology in the global information space.

The number of Internet users is increasing every day. Network technologies are increasingly influencing the development of science and technology itself. In recent years, the nature of education has begun to change significantly, moving to the distance level. This transition is taking place even in classical universities. The development of science and education, and indeed the formation of the global information space, is significantly hampered due to the so-called language barrier. This problem has not yet found its fundamental solution.

In recent years, the volume of information intended for translation has increased. The creation of a universal language such as Esperanto, the "Elvish languages" or any other language did not lead to a change in the situation. Using traditional means of intercultural communication may be a worthwhile solution. The present century dictates its own conditions: information changes twenty-four hours a day, electronic means of communication are widely used. In such a situation, the classical approach to translation does not always justify itself. It requires significant investment and time. In some cases, it may be more appropriate to use machine or automatic translation and machine translation systems (MTS).

The work is devoted to the study of the stylistics of the text, and the study of the features of the style of scientific and technical literature, in terms of the use of automated translation tools.

Currently, there is a fairly wide selection of software packages that make the work of a translator easier, which can be divided into two main groups:

- electronic dictionaries (electronic dictionary)
- machine translation systems.

Machine translation systems (MTS) of texts from one natural language to another simulate the work of a human translator.

Their usefulness depends on the extent to which they take into account the objective laws of language and thought. These laws have not yet been studied well. Therefore, when solving the problem of machine translation, it is necessary to take into account the experience of interethnic communication and the experience of translation activities accumulated by humanity. In the process of translation, the main units of meaning are not individual words, but phraseological phrases expressing concepts. It is concepts that are elementary mental images. Only by using them can you build more complex images that correspond to the text being translated.

In our opinion, it is not very true, and is partly biased. To make such a comparison, you need to know the type of MTS and the principle of its operation. Moreover, when they talk about translation "with the help of computer technology," the concepts are often confused. In general, in modern linguistics a number of areas of computer use can be distinguished.

Fully automatic machine translation systems are more of a pipe dream than a realistic idea. We will not consider them in this work. All machine translation systems (MT systems) operate with human participation to one degree or another. TM systems are sometimes also called "translator memory". They are simply a convenient tool rather than an element of automation.

Machine translation systems can use a translation method based on linguistic rules. The most suitable words from the source language are simply replaced with words from the target language.

It is often argued that to successfully solve the problem of machine translation, it is necessary to solve the problem of understanding text in natural language.

Typically, a rule-based translation method uses a symbolic representation (an intermediary) from which the text in the target language is created. And if we take into account the nature of the intermediary, we can talk about interlinguistic machine translation or transfer machine translation. These methods require very large dictionaries with morphological, syntactic and semantic information and a large set of rules.

If the machine translation system has enough data, it can produce a good quality translation. The main difficulty lies in

generating this data. For example, large text corpuses required for statistical translation methods turn out to be insufficient for grammar-based translation. Moreover, for the latter, an additional grammar task is required.

To translate related languages (Uzbek, Russian), a simple substitution of words may be sufficient.

Modern machine translation systems are divided into three large groups:
- rules-based;
- based on examples;
- statistical.

Next we will look at this classification in more detail.

Rule-based machine translation systems are a general term that refers to machine translation systems based on linguistic information about the source and target languages.

They consist of bilingual dictionaries and grammars covering the basic semantic, morphological, syntactic patterns of each language. This approach to machine translation is also called classical.

Based on this data, the source text is sequentially, sentence by sentence, converted into the target text. Often, such systems are contrasted with machine translation systems that are based on examples.

The operating principle of such systems is the connection between the structure of the input and output sentences. The translation is not of particularly good quality. But it works in simple examples.

Translation from English to German would look like:
A girl eats an apple. □ Bir qiz olma yeydi.

These systems are divided into three groups:
- word-by-word translation systems;
- transfer systems;
- interlinguistic;

Word by word translation

Such systems are now used extremely rarely due to the low quality of translation. The words of the source text are converted (as is) into words of the target text. Often such a transformation occurs without lemmatization and morphological analysis. This is the simplest machine translation method. It is used to translate long lists of words (such as directories). It can also be used to compile a subscript for TM systems.

Transfer systems

Both transfer systems and interlinguistic systems have the same general idea. To translate, it is necessary to have an intermediary who carries the meaning of the expression being translated. In interlinguistic systems, the intermediary does not depend on the pair of languages, while in transfer systems it does.

Transfer systems operate on a very simple principle: rules are applied to the input text that match the structures of the source and target languages. The initial stage of work includes morphological, syntactic (and sometimes semantic) analysis of the text to create an internal representation. The translation is generated from this representation using bilingual dictionaries and grammatical rules. Sometimes, based on the primary representation that was obtained from the source text, a more "abstract" internal representation is built. This is done in order to emphasize places that are important for translation, and discard unimportant parts of the text. When constructing a translation text, the transformation of the levels of internal representations occurs in the reverse order.

When using this strategy, fairly high quality translations are obtained, with an accuracy of around 90% (although this greatly depends on the language pair). The operation of any transfer transfer system consists of at least five parts:
- morphological analysis;
- lexical categorization;
- lexical transfer;
- structural transfer;
- morphological generation.

Morphological analysis. Words in the source text are classified by parts of speech. Their morphological characteristics are revealed. Word lemmas are defined.

Lexical categorizations. In any text, some words may have more than one meaning, causing ambiguity in the analysis. Lexical categorization reveals the context of a word. Various kinds of notes and clarifications are possible.

Lexical transfer. Based on a bilingual dictionary, the lemmas of words are translated. The action is very similar to word-by-word translation.

Structural transfer. The words agree in a sentence.

Morphological generation. Based on the output data of the structural transfer, word forms of the translated text are created.

One of the main features of transphenial machine translation systems is the step during which an intermediate representation of the source language text is "transferred" into an intermediate representation of the target language text. This can work at one of two levels of linguistic analysis, or at both.

Levels:
- Surface (syntactic) transfer. This level is characterized by the transfer of "syntactic structures" between the source and target languages. Suitable for languages in the same family or type, for example in Romance languages, between Italian Spanish, Catalan, French, etc.
- Deep (semantic) transfer. The level is characterized by a semantic representation. It depends on the original language. This representation may consist of a number of structures that represent meaning. Translation also usually requires a structural transfer. This level is used for translation between more distant languages.

Interlinguistic machine translation

Interlinguistic machine translation is one of the classical approaches to machine translation. The source text is transformed into an abstract representation that is independent of language (unlike a transfer translation). The translated text is created based on this representation. The main advantage of this approach is that it allows you to add a new language to the system. It can be proven mathematically that within the framework of this approach, the creation of each new language interpreter for such a system will reduce its cost, compared, for example, with a transfer translation system. In addition, within this approach it is possible
- implement "text retelling", paraphrasing the source text within one language;

- relatively simple implementation of translation of very different languages, such as, for example, Uzbek and Arabic.

However, there are still no implementations of this approach that would work correctly for at least two languages. Many experts express doubts about the possibility of such implementation. The biggest challenge for creating such systems is designing an interlingual representation. It must be both abstract and independent of specific languages, but at the same time it must reflect the features of any existing language. On the other hand, within the framework of artificial intelligence, the task of identifying the meaning of a text has not yet been solved.

The interlinguistic approach was first proposed in the 17th century by Descartes and Leibniz, who proposed universal dictionaries using numerical codes. Others such as Cave Beck, Athanasius Kircher and Johann Joachim Becher worked to develop an unambiguous universal language based on the principles of logic and iconography. In 1668, John Wilkins, in his treatise "An Essay on Genuine Symbolism and Philosophical Language," spoke about his interlingua.

During the 18th and 19th centuries, many universal languages were developed, including Esperanto. It is known that the idea of a universal language for machine translation did not manifest itself in any way at the initial stages of the development of this technology. Instead, only pairs of languages were considered. However, during the 1950s and 60s, researchers in Cambridge led by Margaret Masterman, in Leningrad led by Nikolai Andreev and in Milan by Silvio Ceccato began work in this area. In the 1970s and 1980s, some progress was made in this area and a number of machine translation systems were built.

In this translation method, interlingual representation can be seen as a way of describing the analysis of a text in the original language. At the same time, the morphological and syntactic characteristics of the text are preserved in the representation. It is assumed that in this way the "meaning" can be conveyed when creating a translated text. In this case, two interlingual representations are sometimes used. One of them more reflects the characteristics of the source language. The other is the target language. The translation in this case is carried out in two stages.

In some cases, two or more representations of the same level are used (equally close to both languages), but differing in topic. This is necessary to improve the quality of translation of specific texts. This approach is not new to linguistics. It is based on the idea of the proximity of languages. To use the interlinguistic machine translation system you need:

- dictionaries for analysis and generation of texts;
- description of language grammars;
- concept knowledge base (to create an interlingual representation);
- concept projection rules for languages and representation.

The hardest part about creating this type is the inability to build a base for broad areas of knowledge. And those databases that are created for very specific topics have high computational complexity.

Example-based translation is one of the approaches to machine translation that uses a bilingual text corpus. This corpus of text is used as a knowledge base during translation. Roughly speaking, this is a translation by analogy.

If we think about how a person translates, we are unlikely to come to the conclusion that the translator carries out a deep linguistic analysis. It is assumed that people decompose the source text into phrases, then translate these phrases, and then compose the translated text from the phrases. Moreover, the translation of phrases usually occurs by analogy with previous translations.

To build a machine translation system based on examples, you will need a language corpus made up of pairs of sentences.

Language pairs - texts containing sentences in one language and corresponding sentences in the second, can be either variants of writing two sentences by a person who is a native speaker of two languages, or a set of sentences and their translations made by a person.

Translation based on examples is best suited for things like phrasal verbs. The meanings of phrasal verbs depend greatly on context. Phrasal verbs are very common in spoken English. They consist of a verb with a preposition or adverb. The meaning of such an expression cannot be derived from the meanings of its constituent parts. Classical translation methods are not applicable in this case.

This translation method can be used to determine the context of sentences.

Bilingual text corpora

The expected question arises, where to get such pairs. Examples of bilingual text corpora include parliamentary records in Canada, Hong Kong and other countries. The texts are minutes of debates in parliament. Also, the official documents of the European Economic Community are a good example. They are published in 11 languages. The United Nations publishes documents in several languages. These materials have proven to be very useful for machine translation.

Statistical machine translation is a machine translation technique. It uses comparison of large volumes of language pairs, as well as example-based machine translation.

Statistical machine translation has the property of "self-learning". The more language pairs there are and the more closely they match each other, the better the statistical machine translation result.

Statistical machine translation is based on finding the most likely translation of a sentence using data from bilingual text corpora. As a result, when performing translation, the computer does not operate with linguistic algorithms, but calculates the probability of using a particular word or expression. The word or sequence of words that has the optimal probability is considered to be the most consistent with the translation of the source text and is substituted by the computer into the resulting text.

In statistical machine translation, the task is not to translate the text, but to decipher it. We assume that an article written in English is actually an article written in English, but the text is encrypted (or corrupted by noise). With this approach, it becomes clear why the further the languages are, the better the statistical method works, compared to classical approaches.

More details about the mathematical model of statistical machine translation (Shannon Model) are described below.

Shannon model

The model consists of five elements: information source, transmitter, transmission channel, receiver and final target, arranged linearly.

The transmitter encodes the information received from the source and transmits it to the channel. Through the transmission channel, which is affected by noise - interference of any kind that distorts information, the data enters the receiver, where it is decoded and transmitted to the final target.

Due to noise, the information received by the receiver generally does not match the information sent by the transmitter. However, according to Shannon, by creating redundant information, the original data can be restored with arbitrarily high probability. Checksums are used to detect errors, and special correction codes are used to correct them (provided that the degree of noise does not exceed a certain limit).

It is worth noting that all information is in some way redundant (Shannon, 1948: 380). Human speech is redundant - to understand the meaning of a sentence, it is often not necessary to hear it in full. Similarly, written language is also redundant, and this can be used in translation. If a sentence as a whole is clear, but there are a few unfamiliar words, it is usually not difficult to guess their meaning.

Thus, to translate text it is necessary to find a decoding method that uses natural redundancy, and therefore decoding must be probabilistic.

The task of such decoding is to, given a message, find the original message that has the highest probability. To do this, it is necessary for any two messages to be able to find the conditional probability that the translated message, having passed through a channel with noise, will be converted into the original message.

In this case, we need a source model (language model) and a channel model (translation model). The language model estimates the probability of phrases in the target language, and the translation model estimates the probability of the original phrase given the phrase in the target language.

If we need to translate a phrase from Uzbek to English, then we must know what exactly is commonly spoken in English and how English phrases are distorted into the Uzbek language. Translation itself turns into a process of searching for an English phrase that would maximize the products of the unconditional probability of the English phrase and the probability of the Uzbek phrase (original) given the given English phrase.

$$\max_E P(E|R) = \max_E P(E) \cdot P(R|E)$$

- E - translation phrase (English);
- R - original phrase (Uzbek).

In statistical translation systems, variants of the n-gram model are used as a language model (for example, in Google translator, a 5-gram model is used). According to this model, the correct choice of a particular word depends only on the preceding (n-1) words.

The simplest statistical translation model is the literal translation model. In this model, known as IBM Model No. 1, it is assumed that to translate a sentence from one language to another, it is enough to translate all the words (create a "bag of words"), and the language model will provide them in the correct order. The only data array that Model No. 1 operates on is the table of probabilities of paired translation correspondences of words in two languages (Rakhimberdiev, 2003: 101). More complex translation models are usually used. Many of them are trade secrets of the development companies.

The operation of statistical systems, as well as systems based on examples, occurs in two modes: training and operation.

In the learning mode, parallel corpora of text are viewed and the probabilities of translation matches are calculated. A model of the target language is built. The probabilities of each n-gram are immediately determined.

In operation mode, for a phrase from the source text, a phrase in the target text is searched, so as to maximize the product of probabilities.

After the operation of the MTS (transfer type, Example-Based), unidentified text fragments are translated into a foreign language manually. In this case, you can use the procedure of approximate search for these fragments in the database, and use the search results as a hint. The results of manual translation of new text fragments can be entered back into the database. Then, as more and more documents are translated, the "translator's memory" will gradually be enriched, and its effectiveness will increase. The indisputable advantage of the "translator memory" technology is the high quality of translation of the class of texts for which it was created.

But translation correspondence databases built for homogeneous texts of one enterprise are suitable only for homogeneous texts of enterprises with similar profiles, since sentences and large fragments of sentences extracted from the texts of some documents, as a rule, are not found or are very rarely found in the texts of other documents. Practical implementation is associated with large labor costs for creating a "translator's memory" or replenishing arrays of bilingual texts (bilingual). Scientific, technical and mathematical texts are most often translated using this system. The authors of this work, in particular, know that a similar approach is often used by the Kurchatov Institute.

Let us briefly consider the advantages and disadvantages of existing systems.

Word-by-word translation systems are currently used only for interlinear translation, as noted earlier.

Advantages:
- simplicity;
- high speed of operation;
- not demanding on resources.

Disadvantages: low quality of translation

There are no prominent representatives on the market; in this case, it is more convenient to create a new system for a specific task.

Transfer systems are very widespread.

The most famous representatives are:
- ImTranslator;
- PROMPT.

All such systems have similar advantages and disadvantages.

Advantages:
- high quality translation (if the necessary dictionaries and rules are available);
- there is usually a choice of text topics, which improves the quality of the translation;
- it is possible to refine the translation by making changes to the translator's database (thus, the user receives a potentially infinite set of terms that can be freely manipulated, and "infinite" translation quality can be achieved).

Flaws:
- high cost and development time;
- to add a new language, you have to redo the system again;
- a team of qualified linguists is needed to describe each source and each target language.
- demands on resources at the stage of compiling the database.

Interlinguistic translation systems have never been brought to the level of industrial systems.

Expected benefits:
- high quality of translation, regardless of the choice of language.
- meaning is extracted from the source text once and then written down in any language, including the source one (we get a "retelling of the text");
- low cost of labor for adding a new language to the system.

Flaws:
- controversial potential;
- high complexity of development;

The amount of work required to improve the quality of a translation by a certain percentage increases with the order of magnitude of that percentage.

- systems are not scalable.

The interlinguistic model led to the development of the ETAP machine translation system in Russia, on which a huge amount of time and effort was spent, without any visible result.

The Abbyy company, for ten years, under the leadership of V. P. Selegey, has also been trying to create an interlinguistic system.

You can also mention the DIALING project. The result of the project was only a library of machine morphology of the Uzbek language (website aot.ru).

MTS, based on examples, also does not have prominent representatives. Existing prototypes are used in academic settings to illustrate the method itself. Often they are supplied not as a finished product, but as a set of libraries:
- Marclator - MTS of the University of Dublin;
- Cunei is a hybrid SMT based on analogical translation and statistical translation.

Let's consider the advantages and disadvantages of such systems:

Advantages:
- high quality translation (if the system has been trained for a long enough time);
- copes well with many contextual tasks (phrasal verbs);
- qualified linguists are not needed directly to build the system, only engineers are needed;
- logical simplicity of the device;
- it is possible to train the system during its operation.

Flaws:
- to train the system, large parallel corpuses of text, marked in a certain way, are needed.
- translation is highly dependent on the corpora that were used in training;
- to create such systems, specialized programming languages are required;
- long training time;
- demands on resources at the learning stage.

Statistical machine translation systems have been actively developed (and are being developed) by IBM. Thanks to her developments, translation models IBM Model 1-5 were created. But this method became most famous thanks to Google. In addition to Google translator, there are a number of systems and libraries that use a statistical approach:
- Giza++ ;
- Moses;
- Pharaoh;
- Rewrite;
- BLEU scoring tool.

Not very long ago, a statistical translator for Yandex appeared, although so far it only speaks Uzbek, English and Russian.

Advantages:
- high translation quality (for phrases that fit entirely into the n-gram model):
- if the system has been trained for a sufficiently long time.
- in the presence of high-quality text corpora;
- qualified linguists are not needed directly to build the system, only engineers are needed;
- human labor is minimized to create such systems;
- there is no need to rebuild the system when adding a new language;
- it is possible to train the system during its operation.

Flaws:
- training requires large parallel corpora of text;
- complex mathematical apparatus;
- high-quality translation is possible only for phrases that fit entirely into the n-gram model;
- translation is highly dependent on the corpora used in training.
- when adding a new language, you have to analyze a large number of parallel corpora;
- long training time;
- demands on resources at the learning stage.

The advantages and disadvantages of example-based MTSs and statistical MTSs overlap in many ways. However, the huge advantage of the latter is that such systems are trained without human intervention. Statistical translation does not require additional marking of text corpora; this greatly simplifies their construction. On the other hand, for high-quality training of both, significant volumes of parallel texts are needed. Therefore, translators are often additional services of search engines (Google, Yandex). At the moment, statistical systems are leaders in terms of price/quality ratio for all MTSs.

Reference

[1] Brown R. Adding Linguistic Knowledge to a Lexical Example-Based Translation System, in Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99), Chester, UK, 1999.

[2] Brown R., Automated Generalization of Translation Examples, Pittsburg, PA, USA, 2000.

[3] Hutchins W.J., Somers H.L. An Introduction to Machine Translation. London: Academic Press, 1992.

[4] Samigova, H., Guo, T., & Zhao, Y. (2022). Dialogic rhetoric of English and Uzbek. Translation Studies: Problems, Solutions and Prospects, (1), 304–307. retrieved from https://inlibrary.uz/index.php/translation_studies/article/view/6101

[5] Kurganov, A., & Samigova, H. (2022). Dialogical rhetoric: tadcits and conversations. in Library, 22(2), 1–266. retrieved from https://inlibrary.uz/index.php/archive/article/view/12349

[6] Botirovna, S. Kh., & M. B, A. (2022). Expressiveness in English and Uzbek Languages. Central Asian Journal of Literature, Philosophy and Culture, 3(3), 16-21. Retrieved from https://www.cajlpc.centralasianstudies.org/index.php/CAJLPC/ article/view/299