UDC 81.322.4
DOI https://doi.org/10.32841/2409-1154.2022.53-2.8

Ogurtsova O. L.,
Ph.D. in Philology,
Associate Professor at the Department of English for Humanities 3
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

Shevchenko O. M.,
MBA,
Senior Lecturer at the Department of English for Humanities 3
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

SPECIAL ASPECTS OF MACHINE TRANSLATION TECHNOLOGIES

Summary. This paper describes some major machine translation methods designed to speed up the process of multilingual text translation. Machine translation is achieved by using a computer software transforming text from one language to another. At present several different machine translation approaches are used. Among them are: Rule-Based Machine Translation (RBMT), Statistical Machine Translation (SMT), Hybrid Machine Translation (HMT) and Neural Machine Translation (NMT). Most Rule-Based Machine Translation systems are word-based and create translations by parsing the source text. The focus is made on spelling and grammar of both source and target languages.. The text is split into grammatical constituents and the structure of a sentence is transferred into a target language. The translated words are fitted into the transferred structure. This method requires extensive lexicon with morphological, syntactic and semantic information and large sets of rules. Statistical Machine Translation is based on the analysis of the existing bilingual text corpora. Most modern SMT systems are phrase-based and generate translations using phrases found by statistical methods. The generated statistical models are analyzed by the system and the most likely translations are proposed. Neural Machine Translation uses artificial neural networks which predict the sequence of words and produce sentences. The main function of these trained neural networks is to encode and decode the source text. Each of these approaches has its advantages and disadvantages. However, current MT quality still remains imperfect as the natural languages are complex and work on different levels. Also, such neural machine translation systems as Google Translate, Microsoft Translator, Amazon Translate, Systran's Pure Neural Machine Translator and Yandex Translate were considered.

Key words: machine translation (MT), Rule-Based Machine Translation (RBMT), Statistical Machine Translation (SMT), Hybrid Machine Translation (HMT), Neural Machine Translation (NMT), source language, target language.

Introduction. Machine translation (MT) is an automatic translation from one language to another with the help of a computer software. This process is sometimes described as an automated translation performed by a computer.

The modern world offers enormous amounts of multilingual information and we are often faced with the problem of how to translate it in the shortest possible time. Also, today a large amount of information from all areas of life is available to the users of the Internet. However, the content of many interesting sites is presented only in a foreign language. To quickly overcome the language barrier a variety of machine translation systems are being widely used today.

In fact, automated translation may effectively solve the problem of growing number of translations and at the same time increase productivity of translation.

How does the program manage to coherently translate text from one language to another? What are the current approaches in machine translation (MT)?

At present, there exist several fundamentally different machine translation technologies. The first technology is based on language rules (Rule-Based Machine Translation or RBMT), the second – on the statistics (Statistical Machine Translation or SMT), the third one – Hybrid Machine Translation (HMT) is a combination of several types of machine translation systems, and the fourth type is the Neural Machine Translation which is based on neural networks.

The problem of machine translation was investigated by many researchers worldwide.Nowadays this branch of science finds a rapid development due to the rapid advancement of computer technologies. Researchers try to find new solutions to make the translation process more efficient, fast and accurate. Such scientists as J.M.Cohen, J.W.Hutchins, Andy Way, John Lehrberger and many others made their invaluable contribution to the study and the development of this branch of science.

All these technologies have their pros and cons, supporters and opponents, and the issue often discussed today is which of them allows to get the top quality result. This paper offers an analysis of the translation methods mentioned above:

1. Rule-Based Machine Translation

Rule-Based Machine Translation is based on the application of a great number of linguistic rules (algorithms) which are used in the process of translation in the following sequence: analysis, transfer and generation. The program analyzes the text and using the results of the analysis synthesizes translation. This method requires an extremely massive lexicon with information about the language morphological, syntactical and semantic structure. The translation is done with the help of built-in dictionaries for a given language pair. This translation process is also based on grammar rules which include morphological, syntactic and semantic analyses of words in both languages. On the basis of these complex sets of grammar rules, the grammatical structure of the source language is transferred into the grammatical structure of the target language [1, c. 4], The process performed by such a system is similar to

the process of human thinking: the system analyzes the text using a variety of algorithms.

This method, along with other methods, is used by some developers of machine translation systems, like (PROMT, SYSTRAN, Apertium, GramTrans, etc.)

In the process of translation by using Rule-Based MT method, the sentence from a source language normally goes through the following stages:

Morphological analysis

Before starting a translation of a sentence, the program first analyzes the words in each sentence in terms of morphology, i.e. indicating their gender, number, person, and other morphological characteristics. At this stage, the program does not solve the question of grammatical ambiguity, but only keeps this information. The following example is a good illustration of the general frame of this method: 'A computer executes a program' (Source language – English, target language – Ukrainian). In this sentence 'a' is an indefinite article; 'computer' is a noun; 'executes' is a verb; 'a' is an indefinite article; 'program' is a noun.

After morphological analysis the system performs the following actions:

It solves the problem of grammatical ambiguity (determines the meaning of words, which may belong to different parts of speech) on the contextual level.

For example, if the word belongs to different parts of speech, like the English word 'record' which can be used as a verb (to rec**o**rd = to write smth. down) or as a noun (a r**e**cord = a written account of smth.), the system determines that 'to record' is a form of a verb and provides it with the appropriate morphological characteristics.

Syntactic Analysis

The next stage in the translation process is the process of determination of parts of the sentence and their place in the sentence, the boundaries of simple sentences and their relationships with each other in complex sentences. First, the program searches for a predicate, then for a subject which precedes the predicate (it is assumed that the word order is direct). If, however, there is no subject before the predicate, the system searches it in the postposition, or it assumes that there is no subject at all like, for example, in impersonal sentences ("It is cold") or in imperative sentences ("Turn on the light"). In our example, the system provides syntactic information about the verb: 'executes' = Present Simple, 3rd person singular, Active Voice.

Sentence Synthesis

This is the final stage of the translation process when the elements within groups are coordinated, e.g. predicate and words that depend on it (subject, direct and / or indirect object) are arranged according to the rules of the target language and the correct word-order is used. In the process of translation, the program uses a set of algorithms that help make translation according to the grammatical and other features of a particular target language. In our example the elements of a sentence are coordinated and arranged according to the rules of the target language: En.'A computer' (subject) + 'executes' (predicate) + 'a program' (direct object) → Ukr.. 'комп'ютер' (subject) + 'виконує' (predicate) + 'програму' (direct object).

As a result, in spite of certain inaccuracies found in the translation, the user will understand the gist of the text translated with the help of the Rule-based MT system.

The advantages of systems based on grammar rules are: fairly good grammatical and syntactic accuracy, stable results, the ability to customize text.

However, the creation of such systems requires much time and huge linguistic resources, like thousands of specialized bilingual dictionaries, and good knowledge of grammar, syntax, semantics, etc. both in the source and target languages. This makes the process of the RBMT system development very time-consuming and expensive.

2. Statistical Machine Translation

Statistical machine translation is based on statistical translation of language models obtained from the analysis of bilingual text corpora. It does not use linguistic translation algorithms, and relies on a statistical calculation of the probability of a match [2, c. 182], A bilingual corpus containing huge amount of text in the source language together with its human translation into the target language is downloaded into the system. Then the system analyzes the statistical data about interlingual matches, syntactic structures, etc. In fact, it is a self-learning system which is based on previously obtained statistical results. The larger and more versatile the dictionary, the better the results of Statistical MT. If you work with large databases of parallel texts, you can expect higher quality of the translation. Every newly translated text improves the quality of subsequent translations.

The systems of statistical machine translation are characterized by quick setting and by the ability to add easily new language pairs. Thus the Statistical MT can be described as the process of finding and matching identical pairs from source and target languages.

In the process of translation the Statistical MT systems breaks up source sentences into phrases. This method of finding relevant pairs of phrases yields fewer errors in target language sentences as they include the word combinations and keep the word order of the target language.

The following example is a good demonstration of how a parallel corpus works. It consists of two collections of documents : a source language collection and an identical target language collection. Every sentence in the target language is a translation of the source language sentence with a certain degree of probability (from high to low). One of the main tasks in the process of SMT is to estimate the probability of translation and to find a sentence with the highest probability. Thus, in the sentence pair *He welcomes his friends/ Він вітає своїх друзів* we can see that *He* produces *Він*, *welcomes* produces *вітає, his* produces *своїх*, *friends* produces *друзів.* In this case we can say that each word is aligned with the word it produces. However, not all pairs of sentences are as simple as in this example. In the pair *(He never saw an elephant/Він жодного разу не бачив слона)* we can align *He* with *Він, saw* with *бачив, elephant* with *слон-а*. But *never* aligns with three words: *жодного/разу/ не,* thus creating difficulties in the process of generating a word by word translation. Sometimes words in the source language (SL) sentence align with nothing in the target language (TL) sentence which often leads to inappropriate translations.

3. Hybrid Machine Translation

This method uses several MT approaches (rules and statistics) within a single MT system. Translations are performed using a rule-based approach followed by a statistical approach. Also, rules can be used to pre-process the input data and then to post-process the statistical output and then to enhance the quality of translation

Combination of both rule-based MT, statistical MT and neural MT demonstrates marked improvements in machine translation quality.

Table 1

MT technologies used by some major MT companies and services

Company/Service	MT technology
Google Translate	Statistical and Neural
Microsoft Translator	Statistical and Neural
PROMT	Hybrid, Rule-based, Statistical, Neural
SYSTRAN	Hybrid, Rule-based, Statistical, Neural
Yandex Translate	Statistical and Neural

4. Neural Machine Translation

Neural Machine Translation is a relatively new approach to machine translation. This method allows to make use of artificial neural networks for predicting possible sequence of words and is based on numerical substitutions.

In 2016 Google introduced a neural machine translation system (GNMTS) with the aim to improve the performance of Google translate service. Google started to use NMT which replaced Phrase-Based Machine Translation used previously by Google Translate service. This novel approach which is distinct from the existing methods is very similar to the work of a human brain. It employs neural network technology, a kind of artificial intelligence (AI), and a machine learning system. [3, c. 18], First, the system is 'trained' by a very huge number of human-translated texts. During this process every word is analyzed within the context and then is turned into a digital representation. After that the system finds the same word representation in the target language using the 'knowledge' obtained previously.

The indisputable advantage of the neural machine translation system is that it can be 'trained' directly on both source and target texts without using any specialized systems which are required in the case of statistical machine translation [4, c. 90]. The NMTS consists basically of three key components: encoder, attention mechanism and decoder. The linguistic pair from source and target languages are 'trained' together to ensure a higher degree of probability of the correct translation of a source sentence. In other words, during the translation process the NMT system makes an attempt to build a single neural network that reads a sentence and produces a correct translation. NMT works similarly to human brain. That allows to translate a whole sentence at a time without breaking it.

However, despite the clear benefits the system can have certain problems when translating long sentences. A technical difficulty is that the internal representation has a fixed length to decode each translated word. [5, c. 127]. To solve this problem, the attention mechanism is used to 'teach' the model where to put the focus while the output text is decoded.

Another limitation of the NMT system is that it still has low performance, requires a lot of resources and remains relatively slow.

Neural machine translation method has already demonstrated in practice its obvious advantages over Phrase-Based Machine Translation (PBMT) method.

Google's NMT system is now available through the Standard Edition of the Google Cloud Translation API.

In 2016 Microsoft also started to use NMT system to improve translations provided by its free translation service known as Microsoft Translator. It uses a translation cloud service provided by Microsoft and offers text and speech translation.

Microsoft translator provides immediate translations from one language to another and uses neural networks which mimic the work of a human brain. Microsoft NMT system is claimed to have achieved parity with human translations especially when performing tasks from Chinese to English and vice versa.

In 2017 Facebook Artificial Intelligence (AI) Research introduced its convolutional neural networks (CNNs) which are very efficient for translation and recurrent neural networks (RNNs) which provide an automatic content translation for Facebook users. It is important to Facebook to enable its multilingual users to read posts or to watch videos in their preferred language with the highest accuracy and speed. However, one of the problems remains still unsolved – the neural translation system cannot efficiently cope with lots of informal language, slang and acronyms.

Amazon Translate is a NMT service which was introduced in 2017. Due to its deep learning models it helps users to identify content, such as websites or applications, and to translate large volumes of text very quickly and efficiently. At present Amazon Translate supports translations between English and 14 other languages – Chinese, Spanish, French, German, Arabic, Portuguese, etc.. In near future Amazon Translate plans to provide support for other languages which are highly in demand.

Systran (the pioneer in machine translation) takes an active part in the development of NMT system. Systran's Pure Neural Machine Translator, which is used in an open source community (open NMT), is currently capable of translating between more than 100 different languages. Its neural network engine provides more reliable and accurate translations than ever before. The new translation technology allows to consider the entire input sentence as a unit taking into account distinct features of speech and meaning.

In 2017 Yandex Translate introduced a hybrid machine translation system which uses both statistical and neural machine translation methods. This approach is based on the so-called double translation when a sentence is translated twice, i.e. by using statistical and neural machine translation model. Then both outputs are ranked and the best quality translation is selected. Each of these methods has its own advantages and drawbacks. For example, Statistical Machine Translation (SMT) models are more efficient at memorizing examples and less frequent words and phrases. On the other hand, SMT breaks sentences into words or phrases which sometimes creates difficulties when constructing sentences in a target language. Neural Machine Translation (NMT) models can translate the whole sentences at once. They rely on the context and thus provide more human-like translations. However, as the neural network uses context to find out the meaning of each word, it may fail to translate correctly words which it does not encounter very often.

NMT requires a large parallel corpus to be effective, and is known to fail when the training data is insufficient. At present, a great majority of language pairs lack the required parallel corpora for training the NMT system.

Despite the fact that billions of words are being translated daily by multilingual machine translation services like Google Translate, Microsoft Translator, Systran's Pure Neural Machine Translator, and others, machines have a long way to go before they can function as fluently as humans do when translating into different languages.

Conclusions. In this paper we have analyzed the performance of Rule-Based Machine Translation, Statistical Machine Translation Hybrid Machine Translation and Neural Machine |Translation, Our main observations are:

1. Machine translation is a method that provides more efficient, fast and accurate translation from one language to another.

2. Different MT technologies have their advantages and limitations.

3. A Rule-Based system requires deep knowledge about the source and the target languages to develop morphological, syntactic and semantic rules to generate the translation.

4. Statistical Machine translation is a three-step process: 1)finding the correct word in the given context; 2) finding the best translation of a given word; 3)finding the correct word-order.

5. Neural Machine Translation is a system which uses trained neural networks that read sentences and output their translation. These are efficient end-to-end systems which require only one model for the translation.

Література:

1. Josep Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Young, Jean Senellart, Egor Akhanov, Patrice Brunelle et al. Systran's Pure Natural Machine Translation Systems – arxiv.org/pdf/1610.05540.pdf, 2016. – 4 c.
2. Shen G.R. Corpus-based Approach to Translation Studies. Cross Cultural Communication, 6(4), 2011. C. 181–187.
3. Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, and Mohammad Norouzi et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation – arXiv:1609.08144, 2016. C. 18–19.
4. Ha Nguyen Tien, Huyen Nguyen Thi Minh. Long Sentence Preprocessing in Neural Machine Translation. International Conference on Computing and Communication Technologies IEEE-RIVF – Danang, Vietnam, 20-22 March, 2019. 90 c.
5. Junczys-Dowmunt, Marcin, Tomasz Dwojak, and Hieu Hoang. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions – Proc. of the IWSLT – Tokyo, Japan, 2016. 127 c.

Огурцова О.Л., Шевченко О.М. Особливості технологій машинного перекладу

Анотація. Ця стаття дає короткий аналіз деяких основних методів машинного перекладу, призначеного для прискорення темпів перекладу багатомовного тексту. Машинний переклад досягається шляхом комп'ютерного програмного забезпечення, що трансформує текст з однієї мови на іншу. В даний час в машинному перекладі (MT) використовується кілька різних підходів: машинний переклад на основі правил (RBMT), машинний переклад на основі статистичних моделей (SMT), гібридний машинний переклад (HMT), нейронний машинний переклад (NMT).

Більшість систем машинного перекладу на основі правил базується на аналізі слів і забезпечує переклад шляхом синтаксичного аналізу вихідного тексту. При цьому ситема аналізує орфографію та граматику речень вихідної мови та мови перекладу. Текст розбивається на граматичні складові, а існуюча структура вихідного речення застосовується у мові перекладу. Перекладені слова інтегруються в запропоновану структуру. Цей метод вимагає великої лексичної інформації та великого набору правил для кожної лінгвістичної пари. Статистичний машинний переклад базується на аналізі існуючих двомовних текстових корпусів. Більшість сучасних систем SMT здійснюють переклади за допомогою фраз, знайдених через застосування статистичних методів. Сформовані статистичні моделі аналізуються системою, яка пропонує найбільш вірогідні переклади. Нейронний машинний переклад використовує штучні нейронні мережі, які можуть передбачати послідовність слів і формувати речення. Основною функцією цих навчених нейронних мереж є кодування та декодування вихідного тексту. Кожен з них підходів має свої переваги і недоліки. Тим не менш, на даному етапі розвитку якість машинного перекладу (MT) залишається поки недосконалою, оскільки природні мови є складними і працюють на різних рівнях. Також розглядаються різні системи нейронного машинного перекладу, включаючи Google Translate, Microsoft Translator, Amazon Translate, Systran's Pure Neural Machine Translator, Yandex Translate.

Ключові слова: машинний переклад (MT), нейронний машинний переклад (NMT), машинный переклад на основі правил (RBMT), статистичний машинный переклад (SMT), гібридний машинный переклад (HMT), мова перекладу, вихідна мова.