https://www.sworldjournal.com/index.php/swj/article/view/swj19-03-063
DOI: 10.30888/2663-5712.2023-19-03-063

UDC 81.322

MACHINE TRANSLATION METHODS AND SOME CURRENT TRENDS

Shevchenko O.M.
MBA, Senior lecturer
ORCID: 0000-0001-6726-7269
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Prosp.Peremohy 37, 03056

Abstract. *This paper provides an overview of some major machine translation methods designed to speed up the rate of multilingual text translation. Machine translation is achieved by computer software transforming text from one language to another. At present, different approaches in machine translation (MT) are used: rule-based machine translation (RBMT), statistical machine translation (SMT), neural machine translation (NMT) and others. RBMT relies on linguistic rules and dictionaries to translate text and can provide accurate translations for specific domains. At the same time, this method requires much manual effort to develop and maintain the rules. SMT uses statistical models to translate text. These models are trained on large parallel corpora which consist of parallel sentences in the source and target languages. SMT can handle a wide range of languages though its translations can often be inconsistent or ungrammatical. Neural machine translation (NMT) is a recent approach that uses artificial neural networks to perform translations. This approach has shown a significant improvement in translation quality, fluency and consistency.*

Key words: *Rule-based machine translation (RBMT), Statistical machine translation (SMT), Neural machine translation (NMT), parallel corpora, encoder, decoder*

Introduction.

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

The modern world offers a huge volume of multilingual information and we are often faced with the problem of how to translate it as quickly as possible. Also today, large amounts 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 different machine translation systems are being widely used today.

At present, automated translation may effectively solve the problem of the 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)?

Until recently there existed two fundamentally different machine translation technologies. One is based on the rules (rule-based machine translation or RBMT), and another - on the statistics (statistical machine translation or SMT).

Both technologies have their pros and cons, supporters and opponents, and the issue in question is which of them allows the user to get the best results.

Rule-based MT technology

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, syntactic 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. 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 is used by such machine translation services and platforms as SYSTRAN in France, USA and South Korea, Apertum in Spain, GramTrans in Scandinavian countries, etc.

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

Stage 1: 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 programmer writes a code’ (Source language – English, target language – Ukrainian). In this sentence ‘a’ is an indefinite article; ‘programmer’ is a noun; ‘writes’ 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 record = to write smth. down) or as a noun (a record = 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.

Stage 2: 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, of 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 ("Switch off the computer"). In our example, the system provides syntactic information about the verb: writes = Present Simple, 3rd person singular, Active Voice.

Stage 3. 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 in view of 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.'The programmer' (subject) + 'writes' (predicate) + 'a code' (direct object) ⇒ Ukr.. 'Програміст' (subject) + 'пише' (predicate) + 'код' (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, and 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.

Statistical MT translation technology

Statistical machine translation is based on statistical translation of language models obtained from the analysis of bilingual texts. It does not use linguistic translation algorithms, and relies on a statistical calculation of the probability of a match. A bilingual corpus containing a 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 bigger and more versatile the dictionary, the better the results of statistical machine translation. 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 easily add 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 system 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 illustrates how the SMT system splits the sentence to create pairs of phrases:

'The space station was launched'.

Pairs of phrases:

The space station	космічна станція
The space station was	космічна станція була
Station was	станція була
Station was launched	станція була запущена
Was launched	була (був) запущена (запущений)

The weak point of the statistical system is the lack of a mechanism for grammatical analysis of sentences of both source and target languages. It is hard to imagine that a system which does not analyze text in terms of grammar, is able to provide any adequate translation.

Neural Machine Translation (NMT) technology

Machine translation has undergone significant evolution over the last few years One of the most innovative MT methods is Neural Machine Translation (NMT) introduced in 2014. It uses artificial neural networks to translate text from one language to another. NMT is based on 'sequence-to-sequence' model which consists of two main components: an encoder and a decoder. The encoder processes the input sentence in the source language and encodes it into a 'context vector', The decoder takes the 'context vector' as an input and generates the translated sentence in the target language. NMT models are usually trained on parallel corpora. Parallel corpora ensures that each sentence or text in one language has a corresponding translation in the other language. Parallel corpora also provides a great number of translation examples with a wide range of vocabulary and sentence structures thus enabling the NMT model to produce accurate and adequate translations,

Summary and conclusions.

In this paper we have analyzed the performance of Rule-Based Machine Translation, Statistical Machine and Neural Machine |Translation systems, 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.

References:

1. Bahdanau D,, et al. (2014). Neural Machine Translation by Jointly Learning to Align and Translate [Computer Science CoRR]. https://www.semanticscholar.org/paper/Neural-Machine-Translation-by-Jointly-Learning-to-Bahdanau-Cho/fa72afa9b2cbc8f0d7b05d52548906610ffbb9c5

2. Baniz B.(2020). Machine Translation: A Critical Look At The Performance Of Rule-Based And Statistical Machine Translation [Cadernos de Traducao], issue 40,vol 1 pp. 54-71. https://www.scielo.br/j/ct/a/6yh6JpvZDG4Rq6nFG6KQXmy/?lang=en

3. Koehn P., Och F.J., Marcu D. (2003). Statistical phrase-based machine translation [Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics], pp.127-133. https://aclanthology.org/N03-1017/

4. Shen G.R. (2010). Corpus-based Approach to Translation Studies [Cross Cultural Communication] issue 6, volume 4, pp.181-187. http://www.cscanada.net/index.php/ccc/article/view/j.ccc.1923670020100604.010

Article sent: 25.05.2023

© Shevchenko O.M.