AN OVERVIEW OF AUTOMATED TRANSLATION AND ITS LINGUISTIC PROBLEMS

Zahriddin Xaitqulov

Uzbekistan State World Languages University, Tashkent, Uzbekistan, haitkulovzakhiriddin@gmail.com

Follow this and additional works at: https://uzjournals.edu.uz/philolm

Part of the English Language and Literature Commons, Language Interpretation and Translation Commons, Linguistics Commons, Other Languages, Societies, and Cultures Commons, and the Reading and Language Commons

Recommended Citation

DOI: 10.36078/987654482
Available at: https://uzjournals.edu.uz/philolm/vol2021/iss1/6

This Article is brought to you for free and open access by 2030 Uzbekistan Research Online. It has been accepted for inclusion in Philology Matters by an authorized editor of 2030 Uzbekistan Research Online. For more information, please contact sh.erkinoz@edu.uz.
Zahriddin Xaitqulov
PhD student, Uzbekistan State World Languages University
Tashkent, Uzbekistan
E-mail: haikutlovzahiddin@gmail.com

AN OVERVIEW OF AUTOMATED TRANSLATION AND ITS LINGUISTIC PROBLEMS

ANNOTATION

Language, as the information carrier, has become the most significant means for humans to communicate. However, it has been considered as the barrier of communications between people from different countries. The problem of converting a language quickly and efficiently has become a problem of common concern for humanity. In fact, the demand for language translation has greatly increased in recent times due to effect of cross-regional communication and the need for information exchange. Most material needs to be translated, including scientific and technical documentation, instruction manuals, legal documents, textbooks, publicity leaflets, newspaper reports, etc. The issue is challenging and difficult but mostly it is tedious and repetitive and requires consistency and accuracy. It is becoming difficult for professional translators to meet the increasing demands of translation. In such a situation, the machine translation can be used as a substitute. Machine Translation is the process of converting a natural source language into another natural target language by computer. It is a branch of natural language processing and it has a close relationship with computational linguistics and natural language understanding. With the rapid development of the Internet and the integration of the world economy, how to overcome the barrier of language has become a common problem of the international community. This paper offers an overview of Machine Translation (MT) including the history of MT, linguistic problems of MT, the problem of multiple meanings in MT,
INTRODUCTION

Machine translation in a broad sense is a field of scientific research located at the junction of linguistics, mathematics, cybernetics, and aimed at building systems that implement machine translation in a narrow sense. The history of machine translation generally starts in the 1950s, although work can be found from earlier periods. The Georgetown experiment in 1954 involved a fully automatic translation of more than sixty Russian sentences into English. The experiment was a great success and ushered in an era of significant funding for machine translation research. The authors claimed that within three or five years, machine translation would be a solved problem. However, the real progress was much slower, and after the ALPAC report in 1966, which found that the ten years-long research had failed to fulfill the expectations, the funding was dramatically reduced. Starting in the late 1980s, as computational power increased and became less expensive, more interest began to be shown in statistical models for machine translation [Hutchins W.J., 1994; 166].

The term machine translation (MT) is understood in at least two ways. Machine translation in a narrow sense is the process of translating some text from one natural language to another, implemented by a computer completely or almost completely. In the course of this process, a text is supplied to the input of the machine, the verbal honor of which is not accompanied by any additional instructions, and the output is a text in another language, which is a translation of the input text into the output, and the conversion of the input text to the day off occurs without human intervention (sometimes post-editing is allowed). Today there is still no system that provides the holy-grail of "fully automatic high-quality translation" (FAHQT). However, there are many programs now available that are capable of providing useful output within strict constraints; several of them are available online, such as Google Translate and the SYSTRAN system which powers [Hutchins W.J., 1994; 21-34].

Machine translation of natural languages, commonly known as MT, has multiple personalities. First of all, it is a venerable scientific enterprise, a component of the larger area of studies concerned with the studies of human language understanding capacity. Indeed, computer modeling of thought processes, memory, and knowledge
is an important component of certain areas of linguistics, philosophy, psychology, neuroscience, and the field of artificial intelligence (AI) within computer science. MT promises the practitioners of these sciences empirical results that could be used for corroboration or refutation of a variety of hypotheses and theories. But MT is also a technological challenge of the first order. It offers an opportunity for software designers and engineers to dabble in constructing very complex and large-scale non-numerical systems and for field computational linguists, an opportunity to test their understanding of the syntax and semantics of a variety of languages by encoding this vast, though rarely comprehensive, knowledge into a form suitable for processing by computer programs [Quine W.V., 1987; 8–14].

LITERATURE REVIEW

Machine Translation is a sub-field of computational linguistics that aims to automatically translate text from one language to another using a computing device. To the best of our knowledge, Petr Petrovich Troyanskii was the first person to formally introduce machine translation. In 1939, Petr approached the Academy of Sciences with proposals for mechanical translation, but barring preliminary discussions these proposals were never worked upon. Thereafter, in 1949, Warren Weaver proposed using computers to solve the task of machine translation. Since then, machine translation has been studied extensively under different paradigms over the years. Earlier research focused on rule-based systems, which gave way to example-based systems in the 1980s. Statistical machine translation gained prominence starting late 1980s, and different word-based and phrase-based techniques requiring little to no linguistic information were introduced. With the advent of deep neural networks in 2012, the application of these neural networks in machine translation systems became a major area of research. Recently, researchers announced achieving human parity on automatic Chinese to English news translation using neural machine translation. While early machine translation systems were primarily used to translate scientific and technical documents, contemporary applications are varied. These include various online translation systems for the exchange of bilingual information, teaching systems, and many others [John H., 2000; 187–221].

RESEARCH METHODOLOGY

According to Sebastian Pado, Michel Galley, Dan Jurafsky, and Chris Manning, the machine translation process can be simplified into three stages: a) the analysis of the original-language text, b) the transformation from original-language text to target-language text c) the target-language generation. Seemingly, the core issue of machine translation is the accuracy of the translation. In fact, from the aspect of technology, the core issue is the methodology adopted by the machine translation system [Sebastian P., 1894–1950; 187–221].
MT are lexical problems, word conjunction, and polysemy, syntactic problems, etc. The problem of resolving lexical polysemy is one of the most complex applied problems related to lexical meaning. The problem of automatic (less often semi-automatic) resolution of lexical polysemy was first formulated within the framework of the direction of science and technology related to the creation of machine translation systems. In the future, the problem of resolving lexical ambiguity became one of the key problems not only in the creation of machine translation systems but also in the development of natural language text processing systems for other purposes (search, classification) [Marchuk Y.N., 1983; 40].

**Polysemy** (from the Greek polysemous – polysemous) (polysemic) – the presence of more than one meaning in a language unit – two or more. Often, when talking about polysemy, they mean the polysemy of words as units of vocabulary. Lexical polysemy – the ability of a single word to serve to denote different objects and phenomena of reality. There are two main reasons for the existence of polysemy. First is the principle of economy, which stands in contradiction between the infinite nature of human experience and the limited resources of language. Reality, therefore, often forces the person to use the same word for different classes of objects [Alexey N., 2011; 1-3].

**Grammatical polysemy** – the coincidence of different grammatical forms of the same lexeme. The implementation of a particular meaning of a word is carried out by the context or situation, the general topic of speech. Just as the context determines the specific meaning of a polysemantic word, under certain conditions it can create semantic diffuseness, that is, the compatibility of individual lexical meanings, when their differentiation is not carried out (and does not seem necessary). Some meanings appear only in combination with the defining word; in some combinations, the meaning of a polysemantic word is represented as phraseologically related. Not only lexical compatibility and word-forming features characterize different meanings of words, but also, in some cases, features of grammatical compatibility [Voronovich V.V., 2013; 22].

Applications of computational linguistics do not distinguish homonymous and the only word meanings. This is because, in the vast majority of applied tasks, it is not so much the etymology of the word that is important, as its semantics. Recognizing and separating groups of homonymous meanings is also part of the task of resolving lexical polysemy, as it can sometimes be useful from a practical point of view. The problem of polysemy is considered solved if a regular meaning is chosen for a word or if a synonymous equivalent is found in the form of a regular meaning for metaphorical use.

It is known that there are several independent problems in solving polysemy. In particular, we can distinguish the largest, "classic" tasks:

1) The task of ascribing a known meaning to a known lexeme. 2) The task of attributing a known meaning to a new lexeme. 3) The task of discovering a new meaning for a known lexeme. 4) The task of discovering a new meaning for a new lexeme.

Separately, we can consider problems that are still relatively rarely considered
in the theory of polysemy but are relevant in applied problems

1) The task of identifying a proper name and assigning it to an ontological class.
2) The task of identifying the use of a word in a figurative meaning (metaphor, metonymy, synecdoche).

For new words, including proper names, we must first formulate a list of possible values and then proceed to the solution of the classical problem of resolving polysemy.

There are two main classes of mechanisms for resolving polysemy:
1. These mechanisms are automatic, assuming a completely computer-based solution to this problem.
2. These mechanisms are interactive (interactive, semi-automatic), involving a joint solution of the problem by a person and a computer, and are reduced to the fact that the computer provides the user with a set of alternatives from which he must choose one option [Voronovich V.V., 2013; 22-24].

One of the automatic methods for resolving polysemy is filters, that is, methods that do not reveal the exact value, but explicitly impose restrictions on their spectrum. Examples of such filters are the rules for combining lexemes, the rules for entering octants into syntaxes, and predicative structures.

The mechanisms for resolving polysemy also include those that do not use a lexical meaning that is postulated explicitly, as is done, for example, in an explanatory dictionary. Such methods are usually statistical in nature. To translate polysemantic words, contextual dictionaries are also used, the dictionary entries of which are algorithms for querying the context for the presence or absence of contextual determinants of meaning. For each polysemantic word, its priority translation equivalent is specified, which is specific to the subject area under consideration. At present, there is no need to combine a contextual dictionary and sets of contexts with a special algorithmic procedure, since modern programming languages make it possible to implement a dictionary system on a computer in a variety of ways, depending on the general conditions of its functioning. In the interactive method, the author (editor) of the text is reference dictionary native language semantic additions and translations of words, phrases with amends use special dictionaries, the source, and target languages, as agreed with the dictionary [Michele E., 1991; 2-3].

The reference interpretive dictionary of the source language combines functions of an interpretive dictionary and a translation dictionary. This dictionary reflects those elements of the source language that are of particular importance when translating into at least one of the target languages included in the system of agreed (translated) dictionaries of this source language. The meanings are presented in a separate section, following the descriptions of the semantic meanings of the word for which they are common. This allows us to take into account the diversity of not only lexical but also grammatical meanings in the coding process.

The process of semantic coding of the source text is in the computer is the author of the source text using the utilities containing the mentioned reference dictionary source language and implementing the instructions of the operation of the formation of meaning add-ons. In the process of coding, the author analyzes the source text
sequentially, word by word, and highlights the next word in a special font if, in the opinion of the author (in some cases on the initiative of the utility program), this word has at least one of the following features:

a) this word is polysemantic, and its combination with neighboring words may not contain enough information to select the semantic meaning closest to the source text;

b) the grammatical form of this word and related words does not reflect a particular shade of the actual meaning of the text, although in translation into the target language, this word and (or) related words may have specific grammatical forms, the choice of which strictly depends on the context;

c) there are verbs, participles, adverbs, the form of which is in source language clearly does not reflect a particular character described in the text of validity and (or) state reached in the result of the action, while in a particular target language for the expression of these shades of actions and (or) States are used, depending on the actual meaning of the text, verbs, participles, adverbial participles, having specific grammatical forms;

d) this word, together with some neighboring words, represents a phrase, for the translation of which it may be necessary to search among the known phrases related to this word, and in some cases, there may be differences in the lexical composition or structure that do not affect the allegorical meaning of the phrase, for example, there are "wedged" words, in particular, definitions or other circumstances to certain words that clarify the meaning of the phrase as a whole, or introductory words, variable components are added to the beginning or end, certain proper words or their sequence are changed; In this regard, there is a problem to determine that some words belong to a combination, find the boundaries within the phrase, determine the leading (key) word, and finally choose the meaning that corresponds to the context [Borisevich A.D., 1972 28-36].

Next, the utility program calls from the reference explanatory dictionary, a dictionary article corresponding to the word marked by the author, then the author explains the meaning of this word, comparing the source text with certain elements of the article. The utility has some of the initiative functions, for example, the author points to the mismatch of use words or phrases in the source text and the marked element dictionary entry as well as the author indicates words missed in the process of analysis of the source text, but perhaps has a particular ambiguity.

Syntactic transformations in machine translation: Verb-nominal transformations - are the central issue of forming the structure of a translated utterance. New content is given to the problem of language transformations by modern realities: the need to design and develop training components of machine translation and text knowledge processing systems based on existing and newly created parallel text corpora [Warren W., 1955; 15-23].

At the present stage of linguistic research and development, a synergistic combination of functional and level approaches is necessary. The functional approach integrates language tools (syntactic, lexical, word-forming, and inflectional), belonging to different levels of the language, based on their functional and semantic characteristics. Each lexical form is related to grammatical forms in two ways. On the
one hand, the lexical form, even when taken by itself, abstractly, reveals a meaningful grammatical structure. On the other hand, the lexical form in any particular utterance, being a special language form, is always accompanied by one or another grammatical form. It acts in a certain function, and the cases in which this lexical form mainly occurs, together constitute its grammatical function. Lexical forms that perform any general functions belong to the same formal class. Partially overlapping formal classes may arise based on different functions. Thus, the performance of the function of the actor is typical for substantive expressions and typically infinitive phrases [Nelyubin L.L., 1991; 2-4].

Transformations are understood primarily as transformations of predictors into names and names into predictors: \textit{run} – \textit{running}, \textit{teacher–teaching}, while maintaining a partial identity of the form-the root or base of the word and a certain identity of semantics. Transformations constantly act as one of the two main means-along with periphrases-of creating utterances.

The lack of complete coincidence between language constructs in different languages can be found when studying the comparative frequency of use of individual parts of speech in them, which is important for building machine translation systems.

The functional-semantic approach, which explores the relations of "functional synonymy" of heterogeneous and multi–level language units is extremely relevant at the moment when experiments are being conducted to identify functional and isomeric language structures from parallel text corpora. This approach allows finding correspondences in texts in different languages. In fact, it is not possible to determine in advance with full certainty, exactly how a particular language structure in a text body was translated. Therefore, it is necessary to build and investigate various hypotheses when designing a linguistic processor. Functions are implemented when language objects and their contexts interact [Krasnova T.I., 2015; 124-126].

\textbf{a) Transformations "verb - name"}. The scientific presentation as a whole is characterized by a sign of nominative, i.e. greater use of nouns than in other functional styles. At the same time, a comparative analysis of translations shows that, for example, in Russian, this trend is expressed more clearly than in English, and when translating English verbs are often replaced by nouns. The most productive types of verb-name transformations in English-Russian translation correlates with the following functional values.

1. The circumstances of the goal and the consequence expressed by the infinitive.
2. Compound predicate with an infinitive.
3. Adjective transformations of the infinitive.
4. The infinitive in the second complement function.
5. The infinitive that stands at the beginning of the sentence and performs the function of the subject.

\textbf{b) transformation of the "name/verb"}. The translation of the English gerund into Russian causes difficulties due to its dual nature - it is a non-personal form of the verb that performs in the sentence the functions inherent in the noun: the subject, the complement; as well as the functions of definition and circumstances inherent,
respectively, in the adjective and adverb. Morphologically, the gerund coincides with the actual participle of the English language, which can also play the role of definition and circumstance but can be neither the subject nor the complement in the sentence. This situation is an unerring source of translation errors, even for a human translator (this is one of the central topics in the course of translation theory and practice), and in existing machine translation systems, the distinction between participle and gerund forms does not occur at all, some types of constructions are implemented only in fragments.

The system of transfer rules for machine translation is initially built on the principle of invariant rules, when the translated correspondence it was chosen as the broadest way to translate a certain construction, although not always completely grammatical, but still ensuring the "intelligibility" of the translation in the greatest number of cases. In this approach, preference was always given to the version that was closest in the form to the original English construction: to avoid transformations during translation, which always lead to the appearance of "noise" and a sharp increase in computational costs and, consequently, programming efforts [Kulagina O.S., 1979; 30].

The relevance of the problem of modeling the transformations of verbal and The reason for the lack of nominal constructions for machine translation systems and the extraction of knowledge from texts is that so far these phenomena have been little studied in terms of the possibilities of their computer implementations and, accordingly, are not sufficiently taken into account in existing machine translation systems. The urgent need to create functional-semantic representations of verb-noun transformations is also caused by the fact that the further development of machine translation systems is carried out using machine learning on parallel corpora and the rules that set the functional synonymy of language constructions to allow you to extract the necessary information and avoid the formation of redundant rules and "noise" [Kozerenko E.B., 2007; 18-38].

Translation of phraseological combinations in machine translation systems. The problem with a machine translation of idioms is that it is not always possible to give an accurate translation, following the usual rules. However, it should be taken into account that idioms should be identified at the initial stage in order to avoid their loss, and they should be treated as a single word. 

Idiomatic resolution is one of the formal operations that provide analysis and synthesis in machine translation systems and is performed either by using standard grammatical and lexical text analysis programs working in conjunction with an automatic dictionary or by directly correlating the input and output segments. In the second case, both the input and output segments are considered indivisible revolutions. One or more output turns or word forms assigned to each input turn make up an automatic dictionary of turns. As a result of the analysis, idiomatic expressions are assigned a certain digital equivalent, and they are excluded from the further grammatical analysis. 

The compilation of the algorithm for searching and translating turns in the
text is preceded by a linguistic study of their distribution. It is necessary to find out through syntactic analysis whether the turnover is complete and whether it includes changeable forms. A turnover is considered whole if it has an unchanged composition and no other units can be inserted between its elements. If the turnover is not complete, you should take this into account when composing the algorithm. Thus, parsing data is needed to process turns broken by other members of the sentence. When creating machine translation systems for texts containing idiomatic expressions, the following principles should be followed:

The main units of language and speech that should be included in the machine dictionary should be phraseological units (in particular, idiomatic expressions). Individual words can also be included in the dictionary, but they should only be used in cases where it is not possible to translate, relying only on phraseological units. Along with idiomatic expressions consisting of continuous sequences of words, machine translation systems should also use so-called "speech models" - phraseological units with "empty spaces" that can be filled with various words and phrases, generating meaningful segments of speech [Kurdi M.Z., 2017; 15-30].

Real texts, regardless of their belonging to a particular thematic area, are usually polythematic, if they have a sufficiently large volume. And they differ from each other not so much in their vocabulary as in the probability distributions of the occurrence of various words from the national vocabulary. Therefore, a machine dictionary designed to translate texts even from only one subject area should be polythematic, and even more so for translating texts from different subject areas.

Large machine dictionaries are required. Such dictionaries should be created on the basis of automated processing of bilingual texts that are translations of each other, and in the process of functioning of translation systems. Along with the main polythematic dictionary of a large volume, in the system of phraseological machine translation, it is also advisable to use a set of small-volume additional thematic dictionaries. Additional dictionaries should only contain information that is not available in the main dictionary (for example, information about the priority translated equivalents of phrases and 33 words for different subject areas, if these equivalents do not coincide with the priority translated equivalents of the main dictionary). Along with the translation of texts in automatic mode, it is advisable to provide an interactive mode of their operation in phraseological machine translation systems. In this mode, the user should be able to interfere with the translation process and configure additional machine dictionaries on the subject of the translated texts.

CONCLUSION

Machine translation (MT) plays an important role in benefiting linguists, sociologists, computer scientists, etc. by processing natural language to translate it into some other natural language. And this demand has grown exponentially over the past couple of years, considering the enormous exchange of information between different regions with different regional languages. Machine Translation poses numerous challenges, some of which are: a) Not all words in one language have an equivalent word
in another language b) Two given languages may have completely different structures c) Words can have more than one meaning. Owing to these challenges, along with many others, MT has been an active area of research for more than five decades. Numerous methods have been proposed in the past which either aim at improving the quality of the translations generated by them or study the robustness of these systems by measuring their performance on many different languages [Kulagina O.S., 1979; 30].

Machine translation has a long history but is still a relatively immature technology. For the past decade researchers and developers have been trying to determine the efficacy of existing MT systems and to find solutions for optimizing these MT systems. The progress in the field of MT depends on systematic evaluation and quality control. Every new system works better than the previous one. There are still certain limitations in applications but MT accuracy increases every year. In this article, we discussed and analyzed an overview of machine translation, literature review, and its linguistics problem. Finally, the use of machine translation to help people to obtain information has become an inevitable trend. Machine translation has a long history but is still a relatively immature technology. For the past decade researchers and developers have been trying to determine the efficacy of existing MT systems and to find solutions for optimizing these MT systems. Moreover, progress in the field of MT depends on systematic evaluation and quality control. Every new system works better than the previous one. There are still certain limitations in applications but MT accuracy increases every year.

REFERENCES

26. https://doi.org/10.1017/S0033291700008977