University of Jijel - Mohammed Seddik Benyahia

Faculty of Letters and Foreign Languages

Department of English

Introduction to Human Aided Machine Translation

Semestrial lectures in Computer Assisted Translation for Master 2 students

Submitted by Benhabiles Bakir

2022-2023

Content

List of tables

List of figures

Introduction

Nowadays, machine translation (MT) has become an indispensable reality in the information society to which we belong. Compared to human translation, MT is much faster, offering the possibility to translate millions of words in seconds, and has the ability to improve as more data are stored. For very large projects, MT can not only handle the volume quickly, but it can also organize and tag the content using content management systems. Nevertheless, even though important progress has already been made in this field, the results of MT are still far from being perfect.

In the Algerian academic context, students are not sufficiently informed about the multiple uses of MT and the benefits that could result from integrating MT into the learning process, such as increasing learners' motivation, improving communication, enlarging access to information and knowledge and fostering learner autonomy among others. Hence, this booklet aims to introduce Master2 students to machine translation and computer-assisted technologies (CAT). It provides an overview of the research that has been going on since World War II in the field of computational linguistics and natural language processing. The content will be partly historical and the main approaches will be presented in a largely intuitive way, allowing entry-level students to easily understand the main principles, without any background knowledge.

The booklet comprises five chapter. Chapter 1 will introduce the reader to the most important terminology of MT. Different facets of MT will be discussed so that any potential misunderstandings or uncertainties will be eliminated. Chapter 2 will provide an overview of the historical and technical aspects of MT. Chapter 3 will address some linguistic and technical issues related to language processing in general and MT in particular. In Chapter 4, students will find comprehensive information about the most prevalent approaches in MT, along with a comparison between them. Finally, Chapter 5 initiate students to one of the most promising innovation in the domain of language technology, namely computer-assisted translation (CAT) tools. Consequently, two main core components of this technology will be discussed in order to offer Master2 students the opportunity to understand the developments and limitations of computerized translation.

1. Machine Translation Terminology

Introduction

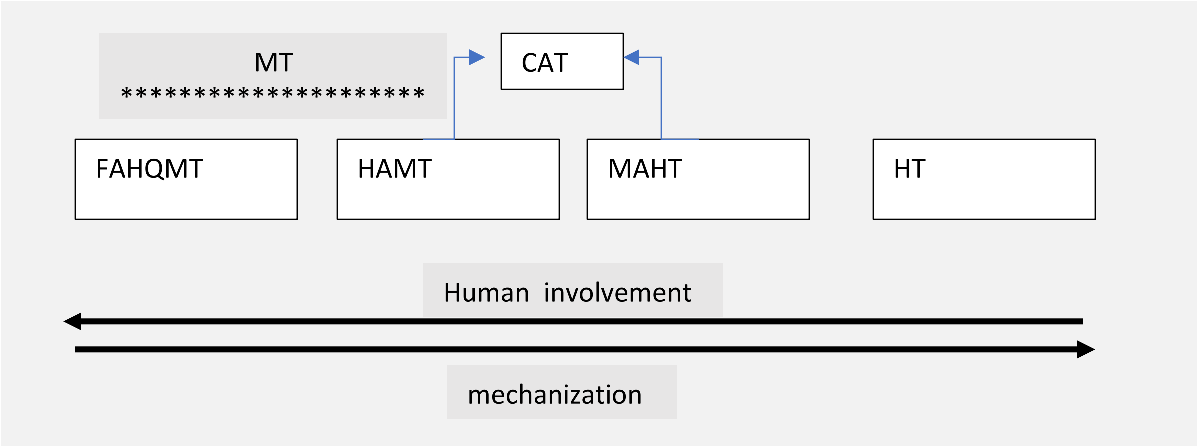
In computational linguistics, several terms are regularly employed to denote computerized tools that help translators manipulate and then generate a quality translation. These terms have several meanings and refers to significantly different products. Each product has its strengths and weaknesses, but all have the capacity to enhance the quality and efficiency of translations. In light of this, it is important to understand the different terminology used to describe machine translation in order to determine the best solution for the specific needs of each translation project.

Broadly speaking, five terms are used to designate the process of translation carried out by machines, namely:

• Machine translation (MT);
• Fully automatic high-quality (machine) translation (FAHQT/FAHQMT).
• Human-aided/assisted machine translation (HAMT);
• Machine-aided/assisted human translation (MAHT);
• Computer-aided/assisted translation (CAT);

Figure1

Human translation vs Machine translation types



Note. Adapted from Hutchins and Somers,1992, p.148

As originally introduced by Hutchins and Somers (1992) and categorized in Figure 1, there are four different ways in which humans interact with machines in translation: *fully automated high quality (machine) translation (FAHQT/FAHQMT), human-aided machine translation (HAMT), machine-aided human translation (MAHT*), and *human translation (HT)*. This linear continuum, though two decades old and may become less relevant as technology continues to evolve, is still beneficial as a means of classifying translations in relation to technology. Therefore, the aim of this chapter is to provide a basic introduction to the different systems used in translation technology.

1. Translation Tools Terminology

1.1 Machine translation

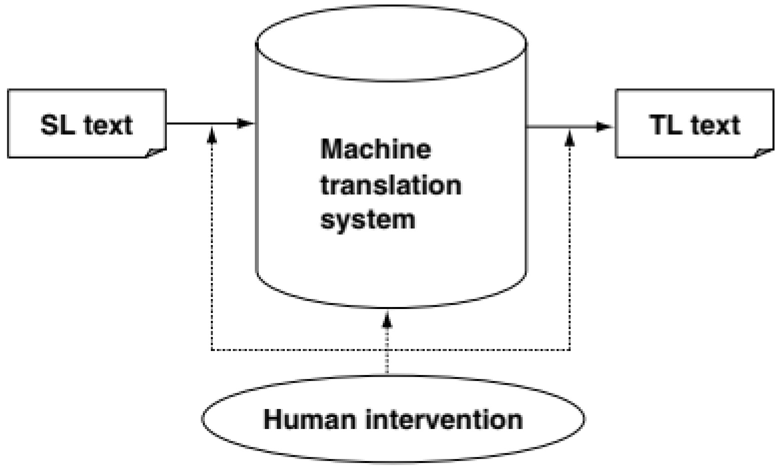
Machine translation (MT) is a contentious term that has been historically (since the 1950s) adopted to refer to three key features:

- The term originally referred only to automatic systems with no human involvement (Sager, 1994, p.326).
- ‘The attempt to automate all or part of the process of translating from one human language to another.(Arnold et al.,1994,p 1)
- Systems that are fully automated as well as those with human involvement (Somers 2003, p. 1)

As can be seen, the main area of disagreement between the three definitions lies in the human involvement during the translation process (Figure2). The first definition prioritizes the automatization of the translation process while excluding any human involvement. The second minimizes the degree of automatization and allows human translator to be part of the process when fully automatic translation is unfeasible. As for the third statements, Somers (2003) enlarges the scope of machine translation to encompasses both fully automated systems and those that involve some level of human involvement, acknowledging through it that machine translation can manifest in diverse forms.

Figue2

Machine translation architecture



Note. From Quah,2006, p.9

1.2 Fully automatic high-quality (machine) translation (FAHQT/FAHQMT)

This appellation refers to MT translation that appeared in the early 1950s. Scientists' initial goal, at the time, was to build high quality machines that are capable of proceeding texts with no human involvement and which generate high quality translations by relying on computational translation. Unfortunately, FAHQMT has failed to meet these high expectations because of the complexity of the human language and the difficulty of capturing all of its nuances and meanings.

Nowadays, the unfeasibility of FAHQT is generally accepted: a translation is either of high quality or generated in a fully automated way, but the combination of the two is presently not possible. Therefore, the FAHQT ideal has been replaced by MT.

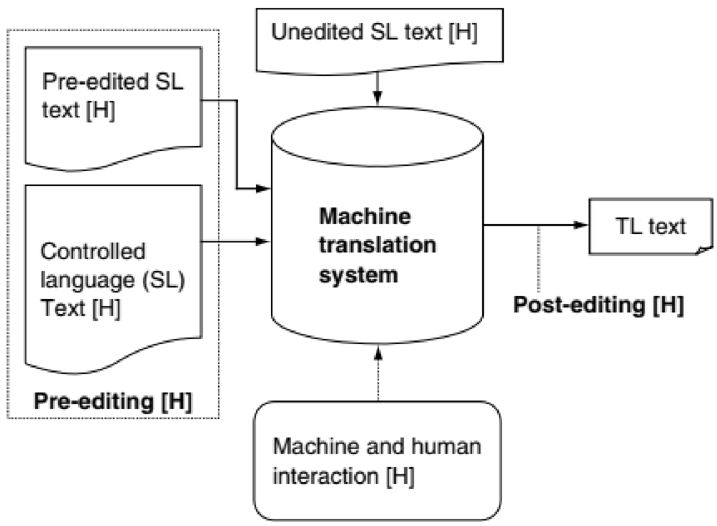
1.3 Human-Aided Machine Translation (HAMT)

Defined by Slocum (1988) as “a system wherein the computer is responsible for producing the translation per se, but may interact with a human monitor at many stages along the way” (p. 5).

In this case , the computer system is the main translator, without excluding a certain degree of human intervention. As shown in Figure 3, the system is in charge of the translation process but also might need human assistance either before (pre-editing), during (editing) or after (post-editing). Examples of human-aided machine translation systems include: MaTra Pro , Lite developed .

Figure3.

Human-aided machine translation model



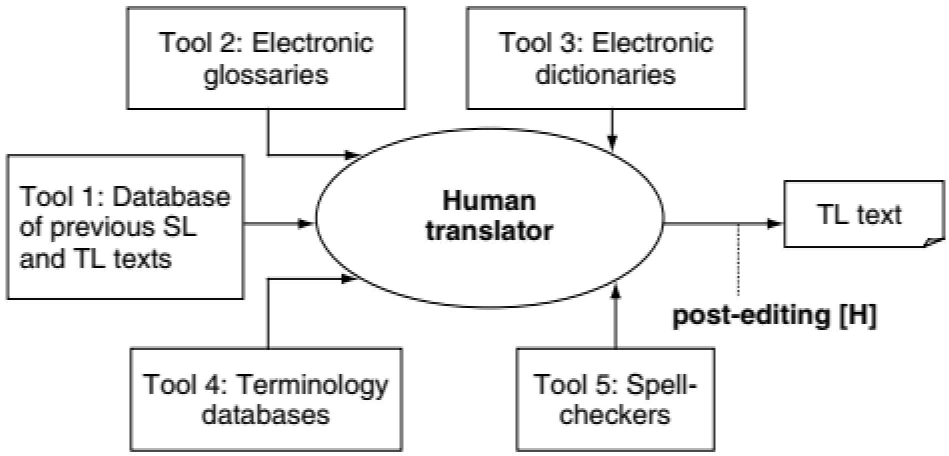
Note. From Quah,2006, p.12

1.4 Machine-Aided Human Translation (MAHT)

MAHT is also known as ‘workbenches’ or ‘workstations’, as they combine a number of features and functions in one software application. MAHT has been described as the use of computer software by translators “to perform part of the process of translation” (Sager, 1994, p. 326). Unlike HAMT, the human translator is the main responsible of all translation decisions (Figure 4). Using supporting electronic tools such as bilingual dictionaries, translation memories, corpora among others, the human involvement is central and irreplaceable in these models of systems. Some examples of commercial machine-aided human translation systems include Trados GmbH, Transit , SDLX Translation Suite and Déjà Vu.

Figure 4.

Machine-aided human translation model (Quah, 2006, p.13)



Note. From Quah,2006, p.13

1.5 Computer Aided Translation (CAT)

According to Sager (1994), a CAT is "a translation strategy whereby translators use computer programs to perform part of the process of translation" (p.326). Conceived in that way, a CAT tool is different from a MT in that the system` role is merely to support the human translator during the translation process. Similarly, Bowker (2002) describes CAT as a form of language translation in which a human translator uses computer hardware to support and facilitate the translation process. The computer hardware typically includes software tools such as translation memory, terminology management, and quality assurance checks, which can help the translator work more efficiently and effectively. However, the human translator remains the key decision-maker throughout the translation process.

1.6 Human translation (HT)

Human translation (HT) is, as the name implies, any type of translation performed solely by a human translator, a professional who has a perfect command of the language combination and is fully familiar with the translation process. When performed by a professional translator, HT does not pose quality problems. In fact, a professional translator has the meticulousness, the creative sense, and long experience to provide a smooth and quality final text. The reading experience will be natural and the reader will not perceive any trace of strangeness. Moreover, the professional translator usually translates into the mother tongue, and therefore has a perfect command of all linguistic and cultural subtleties.

As previously stated, the above classification of machine translation systems introduced by Somer and Hutchins (1992) is now relatively outdated and new issues have sparked since then. Scholars such as Quah (2006) has called for a new taxonomy, which takes into account the recent development in computational linguistics and the multifunctional character of new technologies.

2. The Origins of Natural Language Processing (NLP)

Introduction

In order to understand the ins and outs of translation in the current context of the internet and the information society, it is appropriate to briefly retrace the historical and methodological steps it has gone through. We present in the following paragraphs the major historical development that have introduced automation to translation and explore their theoretical underpinnings and practical applications.

2. The Origins

The origins of machine translation date back to the ninth century with the work of Al-Kindi who developed techniques for systemic language translation based on methods such as cryptanalysis, frequency analysis, probability and statistics. These methods are still used today in modern machine translation (DuPont, 2018). However, it is until the 17th century that the idea of machine translation was initiated, when René Descartes proposed a universal language to represent equivalent ideas in different languages by means of a single symbol. (Knowlson,1975)

In order to seize all the intricacies of MT, one should first understand the notion of "universal language" which was central to the emergence of studies in MT. In its broadest sense, a "Universal language"[1] means a language that allows communication between all humans. The idea was first promoted by philosophers of the 17th like Leibniz, Descartes and John Wilkins who proposed an intermediary abstract language in which all aspect of ambiguity is eliminated (Couturat,1914). If such a language existed, it would eliminate the question of translation or could at least provide the basis for an abstract language (to facilitate translation between the different existing languages.

Technically speaking, this language consists of the replacement of words and concepts with unambiguous identifiers i.e., "a single number". The symbolized concept can then be expressed in several different languages (Couturat,1914). The proposition inspired many scholars to build

[1] - From a historical perspective, the notion of "universal language" marked a turning point in machine translation research in that it has been explored by different scholars and yielded "Interlingua Approach". (see *Machine Translation Approaches)*

"digital dictionaries"[2] such as Cave Beck (1657), Johann Joachim Becher (1661) Athanasius Kircher (1663) and John Wilkins (1668).

However, one should be careful not to see these works as direct forerunners of machine translation systems. The aim of Leibniz, Descartes and others was essentially to solve philosophical, logical and moral problems. Though their research addressed the question of languages and translation, they did not specifically focus on the idea of automatic translation. Interestingly, their works on coding systems have been a source of inspiration for various researchers (and are often quoted in the literature of machine translation) but it seems that they have never been directly involved in the development of operational systems.

2.2 Precursors

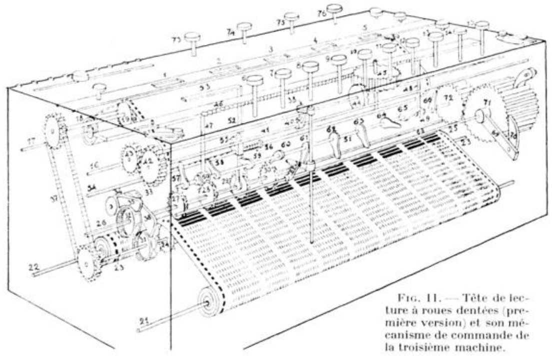
2.2.1 George Artsrouni (1933)

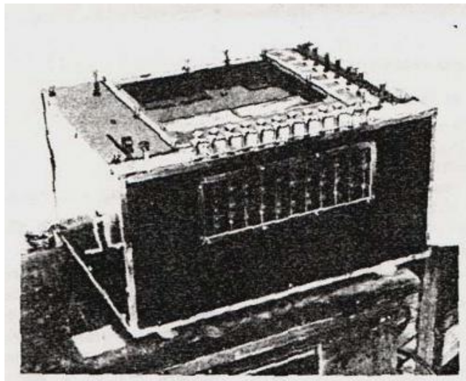
The first attempt was made by Georges Artsrouni, a French engineer of Armenian origin who had studied in Russia before immigrating to France in 1922. In July 1933, Artsrouni filed a patent application for a "mechanical brain" (cerveau mécanique). This general-purpose device it is not an ancestor of our modern computers but a machine allowing to store and find different types of information automatically. According to Hutchin (n.d) Two prototypes will be built (probably between 1932 and 1935), and will receive avid interest during public demonstrations. The machine will even obtain a Grand Prix at the time of the World Fair of 1937 in Paris (another prototype will be built but never completed, the two existing copies are preserved in the Museum of Arts and Techniques, in Paris).

[2] - we speak here of a dictionary where a numerical identifier is associated with each word or concept.

Figure 5.

Artsrouni's "mechanical brain"





Note. From Corbé 1960

At the end of the 1930's, the "mechanical brain" became the object of much attention from various organizations that had to handle large masses of information such as railway timetables, telephone directories, commercial telegraph codes, banking statements, and anthropometric records.(Hutchin, n.d) . Nevertheless, World War II prevented these actions from succeeding, after which the appearance of computers will make these purely mechanical machines obsolete. As reported by Hutchins(n.d), the system consisted of four main components (Fig.5):

- A ***memory*** (bande de réponse) of words in the four languages,
- an ***input device*** consisting of a *keyboard* activating a *reading head* (mécanisme de repérage),
- a ***search mechanism*** (sélecteur),
- and an **output mechanism** (mécanisme de sortie) activated in its turn also by the reading head. The four components were driven by a motor, and the whole apparatus was contained in a rectangular box measuring 25x40x21 cm.

Unfortunately, the system developed by Artsrouni did not allow him to go any further as he was not a linguist, and never considered the difficulties in conceiving a machine translation. Yet, posterity still regards him as the precursor of fully automatic translation system.

2.2.2 Petr P. Smirnov-Trojanskij (1933)

Unlike Astrouni, the proposals of Petr Petrovič Trojanskij (1894-1950) went far beyond the description of a mechanical dictionary. The Russian professor had described how a translation machine could be built using the electromechanical technology of the 1930s and 1940s.[3] Accordingly, three distinct stages were to distinguish: analysis, transfer, and synthesis (Hutchins, 2004). In the analysis stage an editor knowing only the source language would perform the "logical" analysis of the source words in their basic forms and syntactic functions. Then, a machine transforms sequences of these word forms and functions into their corresponding target sequences (transfer). Finally, a second editor knowing only the target language converted the output into the correct forms of the target language (synthesis). Further, he also explained how "translation processes" could work on the basis of "universal" symbols (Esperanto) for the encoding and interpretation of grammatical functions.

According to Hutchin (2004), Trojanskij's ambition was the development of a "translating machine," capable of translating words between several languages simultaneously by relying on Esperanto as an intermediary code language. In the end, Trojanskij's dream to see Esperanto becomes the new universal language of communication and technology remains a pipe dream. In 1937, Staline banned the use of Esperanto which he considered an international "bourgeois" language that facilitated espionage in the USSR) and 30,000 Esperantists were sent to the Gulag.

It should be noted that, despite their efforts, the two inventors will be largely ignored by the academic community. The system elaborated by Artsrouni will not have any continuation after the war, since it was clear that the future is favorable to the electronic calculators, much more powerful than mechanical calculators. The promising ideas of Trojanskij, having never given place to an operational system, will also be largely ignored in favor of fully automatic translation systems. (Poibeau, 2017)

2.3 Pioneers: Warren Weaver

The first experiments in machine translation really began after World War II, under the impulse of the American scientist Warren Weaver (1949). In his very influential memorandum

[3] - Trojanskij submitted the patent for his “translating machine” on 5 September 1933

entitled "Translation", Weaver was the first to discuss ideas about the feasibility of using computers to translate documents between different languages along with conceiving methods and setting objectives. His efforts provided the impetus for the creation of machine translation centers, mainly in universities, with the mission of producing serial translations of scientific texts from Russian into English. Ultimately, Weaver(1949) raised four interesting theoretical points:

1- The problem of multiple meaning (and its connection with context.)

Words should be contextualized before their exact meaning could be determined.

2- The logical basis of language,

Robots have the ability to compute written languages when provided with logical and recursive rules.

3- The application of communication theory and cryptographic techniques, (notably the work of Claude Shannon)
4- The possibilities of language universals.

Human languages possess universally common elements that may foster communication through abstract representation.

Weaver` memorandum elicited mixed reactions among the scientific community. Some scholars rejected the idea of computerizing a complex process such as translation. Others, such as Erwin Reifler and Abraham Kaplan were less dubious and maintained effort towards achieving weaver`s lifelong objectives.

2.4. The First-Generation Systems (Toy systems)

First-generation systems were early machine translation systems that were developed in the late 1940s and 1950s. These systems were based on simple linguistic rules and used a limited vocabulary to translate texts between two languages. The goal of these systems was to demonstrate the feasibility of machine translation, rather than to produce high-quality translations. While they were able to translate simple sentences and phrases, they were not capable of producing translations that were accurate or fluent. Despite their limitations, first-generation systems paved the way for the development of more sophisticated machine translation systems in the decades that followed. (Léon, 2015)

2.4.1 IBM and Georgetown University

The first public demonstration of machine translation was a collaboration between IBM and Georgetown University. The experiment took place in January 1954 in New York and was organized by the inaugural head of Georgetown's Institute of Languages and Linguistics Leon Dostert. it consisted of translating a few sentences from Russian into English with the help of a bilingual Russian-English lexicon of 250 terms and an "operational syntax" comprising six operations regulating the parsing. Paul Garvin of Georgetown university and Peter Sheridan of IBM were the main authors. (Léon, 2015)

Although the experiment was conducted on a small-scale, the IBM-Georgetown demonstration led to spur research in this area around the world: in the United States and Great Britain new research groups were created and financially supported by the National Science Foundation (NSF) and the CIA. In Russia, the first conference on automatic translation was held in Moscow (1958), bringing together some 340 participants from 79 different institutions. Many countries then became involved in this promising research: Japan in 1956, Czechoslovakia in 1957, China in 1958-59, Italy and France in 1959, Mexico in 1960, Belgium in 1961. The research remains little developed in the Federal Republic of Germany, in Sweden and in Finland (Hutchins, 2006).

2.5 The ALPAC report and its consequences (1966)

The ALPAC report, also known as the "Automatic Language Processing Advisory Committee Report," was an advisory committee created by the U.S. government in 1964 to evaluate the state of research in automatic language processing (NLP) and machine translation (MT) in particular. ALPAC was created in response to a 1963 government report that questioned federal investments in NLP and MT research, arguing that the results were insufficient. ALPAC published its own report in 1966, in which it concluded that MT was less accurate, slower and more expensive than human translation. The report stated that the technology was not sufficiently developed to replace human translators in a professional context, and that the research results up to that time were disappointing. In the words of Somers (1997), "language is too complex and the task of translation therefore requires human capabilities, which [...]cannot be easily simulated in a computer program."(p.194). Instead, the report made several recommendations for the future of machine translation research:

Continue Machine Translation Research: Despite the criticism of the state of machine translation research at the time, the report recommended that funding for machine translation research be continued and increased. The report argued that machine translation could be useful in providing approximate translations that could be improved by human translators.

Focus on statistical methods: The report recommended a shift away from rule-based methods in machine translation to statistical methods. The report argued that statistical methods had the potential to produce better results and could be more flexible in dealing with the complexities of language.

Collaboration between linguists and computer scientists: The report stressed the importance of collaboration between linguists and computer scientists in the development of machine translation. The report argued that effective machine translation systems should take into account the nuances of the language and the context in which it is used. (ALPAC,1967)

Overall, the report recommendations led to a reduction in funding for MT research in the 1970s and dampened the initial enthusiasm for MT. However, MT research has continued, and significant progress has been made since then through the use of massive data processing and machine learning methods.

2.6 The "Quiet Decade"(1967-1976)

The "Quiet Decade" is a term that has been used to describe the period from 1967 to 1976 in the field of machine translation (MT). This period is characterized by a reduction in funding and a lack of progress in the development of MT systems, following the release of the ALPAC report in 1966, which concluded that MT was not yet ready for practical use. (Somers,1997)

During the Quiet Decade, there was a reduced interest in MT and many researchers in the field shifted their focus to other areas of linguistics and computer science (Tong 1994). However, some researchers continued to work on MT, and significant advances were made in areas such as natural language processing, lexicography, and computational linguistics.

Despite the slow progress in the development of MT systems during the Quiet Decade, the period was not entirely unproductive. Researchers laid the foundation for future developments in the field, and a number of machine systems was developed such as TAUM-Météo, a machine translation system which translate weather forecasts between English and French , SUSY a German- English-Russian MT and CULT another MT (Chinese University Language Translator)(Quah,2006).

The Quiet Decade was a challenging time for the field of MT, but it also provided an opportunity for researchers to reflect on the challenges of MT and to explore new approaches and technologies. Today, MT is a rapidly growing field, and many of the advances made during the Quiet Decade have paved the way for the development of more advanced and effective MT systems.

2.7 Second Generation Systems (1976-1989)

The second generation of machine translation (MT) systems refers to the period between 1976 and 1989, which was marked by significant advances in the field of MT and the appearance of several new MT systems. The development of second-generation systems has been strongly influenced by advances in computer technology and the expansion of computing resources. The new advancements have permitted the development of more sophisticated and complex MT systems that are capable of processing larger volumes of data and analyzing them more quickly (Quah, 2006), namely:

SYSTRAN: Developed by Peter Toma in the 1968 and became one of the first direct machine translation systems to be used by various non/and government agencies such as IAEA, the European Communities (EC), General Motors, Dornier, and Aérospatiale.

METAL: a German-English rule-based MT system built by the US Air Force at the University of Texas in collaboration with Siemens.

EUROTRA: conceived by the European Commission between 1978 and 1992, EUROTRA was a rule-based MT system designed to translate between the official languages of the European Union. It was one of the largest MT projects of its time. unfortunately, the project failed to build a working MT system.

One of the key innovations of the Second-Generation systems was the use of *rule-based MT,* which involved “the application of morphological, syntactic and/or semantic rules to the analysis of a source-language text and synthesis of a target-language text.” (Carl and Way 2003b: xviii). This approach relied heavily on the creation of a set of rules and procedures for translating text from one language to another and on a deep understanding of the structure of the languages being translated.

Another important development during the Second-Generation period was the reintroduction of *statistical* MT systems.[4] These systems relied on large amounts of parallel texts in different languages and used statistical techniques to identify patterns and relationships between words and phrases in the source and target languages.

2.8 The Modern years

Modern approaches to machine translation can be divided into two broad categories: statistical approaches and neural approaches. Statistical approaches have dominated machine translation for most of the 1990s and 2000s. These approaches are based on probabilistic models that use linguistic data to estimate correspondences between words and sentences in different languages. Statistical approaches include rule-based translation models as well as statistical translation models based on parallel corpora.

Neural approaches became popular in the 2010s. These approaches are based on deep neural networks also called "Deep Learning" . These networks are complex mathematical models that are trained on large amounts of training data to learn to perform a specific task, such as text translation.

There are two types of neural machine translation approaches: "static" neural machine translation and "neural" machine translation (Sutskever et al., 2014). Static" neural machine translation is based on a neural network model that takes as input a sentence in the source language and produces as output a sentence in the target language (Bahdanau et al., 2014). This type of approach is often referred to as an "encoder-decoder" approach, as it encodes the source sentence in an abstract vector space and then decodes this representation to produce the target language translation. Neural" machine translation, also called "neural recurrent translation system", uses a network of recurrent neurons to translate a sentence (Luong et al., 2015). In this type of approach, the neural network uses a feedback method that allows it to take into account the global context of the source sentence, and generate a translation in the target language using an iterative process.

[4] - ideas of statistical machine translation were introduced by Warren Weaver in 1949

3. Challenges in Machine Translation

Introduction

Translation has always been a multifaceted operation that involves understanding the meaning of a text in one language and expressing it in another as closely as possible. While human translators can perform this task with relative ease, computers have long struggled to do so. This lecture will be dedicated to discuss some particular problems facing MT systems.

Roughly speaking there is at least two main reasons why translation is difficult for computers: *the nature of translation* itself and *the abilities of computers to deals with natural languages*.

3.1 The Nature of Translation

In his article *"Why translation is difficult for computers"*, Doug Arnold (2018) cited three main reasons that make translation a challenging task for computers. The first reason is related to the human languages in se, which are disproportionately complex and nuanced. Languages are very different from each other in terms of grammar, vocabulary, syntax, cultural connotations, figurative meaning, and tone. Indeed, a sentence in one language may have several possible interpretations in another language, depending on the context, register and culture. Secondly, the notion of equivalence is also a major challenge for translation, words and phrases cannot always be translated literally, as their meaning and usage vary according to context and culture. Therefore, translators must seek out appropriate equivalents that reflect the intent, tone and meaning of the original, while avoiding misunderstandings and false friends. Finally, creativity is an essential element of translation, language transfer often requires novel solutions to translate idioms, metaphors, wordplay and nuances of meaning. Machines must therefore be able to invent rules rather than simply apply them.

3.2 Computers' ability to deals with natural languages

The ability of computer programs to analyze and understand human language, also known as natural language processing (NLP) has permitted many translators to use their content and data to improve their products and services, but also to provide a more personalized customer experience. This technology has been applied to a variety of fields, including machine translation, speech recognition, speech processing and conversational systems. However, despite recent technological advances, NLP remains a complex and difficult field to master due to several challenges associated with achieving a complete understanding of natural language. At the root of them, Arnold (2003) identified four main limitations.

3.2.1 The inability of computers to perform vaguely specified tasks

The need for precisely formulated rules is essential since they have to be interpreted based on the standard operations of computer hardware. The challenges of NLP, especially MT, often stem from the struggle to come up with accurate and precise formulations for seemingly simple ideas. Additionally, formulating precise rules is not always sufficient to come up with soultions since there are some problems that can be expressed in a precise way but still cannot be solved computationally.

3.2.2 The inability of computers to learn things (as opposed to being told them)

Learning is another major challenge for computers, because it relies on classification based on the vague notion of similarity, and the need to invent rules rather than simply follow them. Although learning algorithms exist for some tasks, there is no general, reliable procedure for learning the kinds of knowledge needed for machine translation. Therefore, for the computer to understand what it needs to translate, explicit rules must be provided in written form by humans.

3.2.3 The inability of computers to perform common-sense reasoning.

It is difficult for computers demonstrate common sense reasoning because of the large number of facts about the world that are involved in this type of reasoning (real world knowledge). The task of encoding all this knowledge is enormous, making it difficult for computers to have a thorough understanding of what we call "common sense." In practice, most aspects of this form of reasoning are far beyond the capabilities of modern computers.

3.2.4 The inability of computers to deal with some problems where there is a large number of potential solutions.

Combinatorial explosion is frequent in languages with significantly different grammatical structures and word orders. This can lead to a number of possible translations that is quickly unmanageable for a computer. To deal with this problem, machine translation systems often use statistical models and algorithms that can identify the most likely translation based on probability distributions derived from large amounts of training data. However, even with these techniques, combinatorial explosion can still be a challenge in machine translation, especially when dealing with complex sentences or rare word combinations. As a result, human translators are often needed to ensure accurate and natural translations.

3.3 Systematic Problems

3.3.1 The analysis Problems

The first problem is related to language analysis, which involves transferring "content" from the source language to the target language. Machine translation systems often have difficulty analyzing sentence structure and producing abstract representations of the content for the following reasons. (Arnold,2003)

- ***Robustness***:

  Robustness is the ability of a computer system to cope with errors at runtime and to deal with erroneous inputs. Grammatical errors, spelling, typos, homonymy, colloquialism, slang, among others, can then pose a significant challenge to NLP systems, which rely on accurate and well-formed input in order to produce reliable results.

- ***Ambiguity***:

  ambiguity is a quality of language that makes speech or written text open to multiple interpretations. That quality makes the meaning difficult or impossible for a person or artificial intelligence (AI) program to reliably decode without some additional information.

Lexical ambiguity

Lexical ambiguity is a considerable barrier for MT systems, since it can lead to incorrect or inaccurate translations. A word is said to be lexically ambiguous when it has more than one meaning depending on the context in which it is used.

Examples

a. The company designs a new <u>*pen*</u>. ('writing implement' or 'animal enclosure'?)

b. Ahmed has grown another <u>*foot*</u>. ('limb' or 'unit of measurement'?)

c. "I saw her duck." ('to bend suddenly , or the bird.?)

d. There was not a <u>single</u> man at the party. (No men at the party or all the men there were married?)

Structural ambiguity

Structural ambiguity, also known as *syntactic ambiguity*, is a type of ambiguity that occurs when a sentence or phrase has two or more distinct meanings. In literature, structural ambiguity is used to create a panoply of stylistic effects, such as adding depth and complexity, forging wordplays, metaphors, irony, humor and many other images.

Examples

a. I saw the man with the telescope. ("I saw the man, who had a telescope" or "I saw the man, while I was holding a telescope"?)
b. Leila writes to her friends in Algiers. (Is Leila writing in Algiers, or are the friends in Algiers?)
c. "Visiting relatives can be a nuisance." ("The act of visiting relatives can be a nuisance" or "Relatives who are visiting can be a nuisance.?")

Pragmatic ambiguity

Pragmatic ambiguity occurs when the meaning of a sentence (the utterance) depends on the context in which it is said. Hence, in order to grasp the meaning, computers need to understand the surrounding context, or the real-world knowledge, which can frequently be unattainable as they are still not on par with humans in this respect.

Examples

a. Do you know what time is it? (Asking for time or expressing anger to someone who missed the due time)`
b. I'm sorry. (Apology or expressing empathy?)
c. some think he has a big mouth (large mouth size, or revealing confidential information)

3.3.2 Transfer Problem

Transferring the abstract representation produced by the source-language analysis component and producing the equivalent output is another challenging problem. Languages differ formally and semantically and thus require appropriate adaptation to achieve an accurate translation.
Examples

- **Translating the conditional**

1-If you finish work early, I'll visit you. اذا تنهي العمل باكرا فسوف أزورك (x)

2-If you finished work early, I would visit you. إذا انهيت العمل باكرا فسوف زرتك (x)

3-If you'd finished work early, I would have visited you. إذا انهيت العمل باكرا فسوف كنت زرتك (x)

There are three types of conditional sentences in English, but only two in Arabic. The main problem here is the translation of the future past ''would have '' into Arabic. The three translations are unacceptable in Arabic because (اذا) cannot be followed by the present (1), and (سوف) doesn't precede the past (2,3).

A better translation to the previous examples would be.

1-If you finish work early, I'll visit you. إذا انهيت العمل باكرا فسوف أزورك

2-If you finished work early, I would visit you. لو أنهيت العمل باكرا لزرتك

3-If you'd finished work early, I would have visited you. لو(أنك) أنهيت العمل باكرا (لكنت زرتك) لزرتك

3.3.3 Synthesis Problem

Synthesis problems are generated from the combination of both analysis and transfer problems. In this case, when the content is transferred into the target language, there is more than one way to express it. Arnold (2003) made a distinction between two main aspects to this problem. The first occurs when there are several valid ways to express the same content, but only one of them is the right option. The second is opposite to the first and may arise when there is no evident means of identifying the correct manner of expressing the content. Therefore, there are multiple possible ways to communicate the same meaning in the target language.

Examples (source: Arnold,2003, p.133)

Several to one option

a. What time is it? (✓)

b. How late is it? (x)

c. What is the hour? (x)

Only (a) is idiomatic in English

No evident option

Sam saw a black cat.

a. Sam saw a cat. It was black.

b. Sam saw something black. It was a cat.

c. Sam saw a cat which was black.

d. Sam saw a black thing which was a cat.

e. A black cat was seen by Sam.

f. Something happened in the past. Sam saw a cat.

g. There was a black cat. Sam saw it.

etc.

According to Arnold (2003), the representation of the content in *Sam saw a black cat* can be expressed in English in many other ways (as listed above), which may be problematic to computer systems because they have to make a choice from all these alternatives.

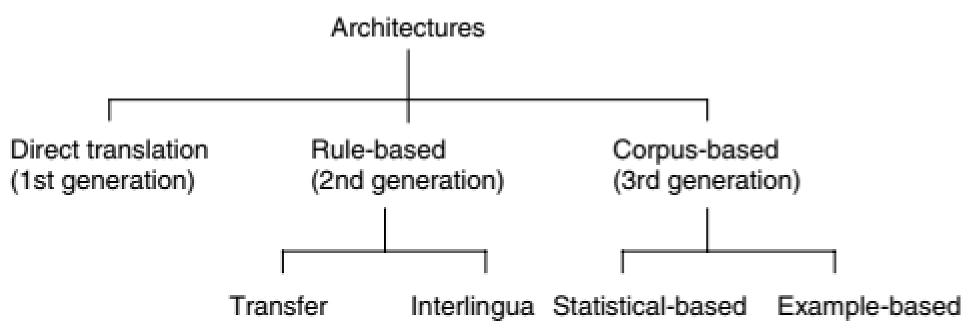
4. Machine Translation Approaches

Introduction

Over time, several theoretical conceptions about automatic translation have emerged. Some have been successfully implemented giving rise to commercial systems; others remained at the research stage or the demonstration prototype. Nevertheless, every conception calls on specific resources and develop a distinct linguistic approach. In the following section we briefly make a presentation of the currently available approaches from historical and theoretical perspectives.

Figure 6.

The architectures of different machine translation systems



Note. From Quah, 2006, p.88

4. Machine translation approaches

There are several approaches to MT and each one of them is generally characterized by its own terminology, technicality, methodology, strengths and weaknesses.

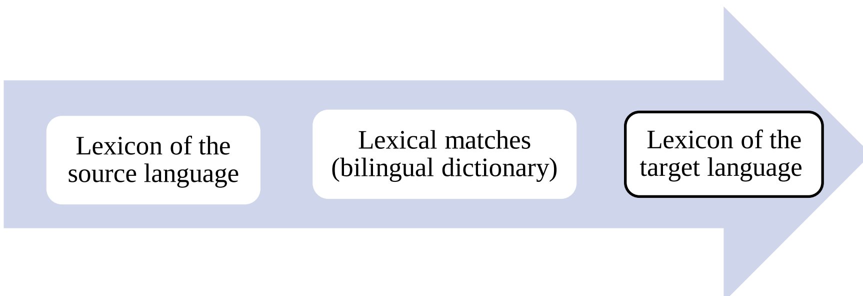
4.1 Direct Approach (Dictionary/lexical)

Lexical approach characterizes *direct machine translation systems*. These systems are based on the search for lexical equivalents in a more or less elaborate bilingual dictionary. In this approach, the systems translate the source language text directly from the lexicon, i.e., without going through an analysis phase of the two languages involved in the translation. The system only uses the resources of the dictionary (lexical correspondences) to carry out the translation automatically. Figure7 illustrates this process.

Because of its apparent simplicity, this approach has been used to develop the *"first generation" systems.*

Figure7.

Direct approach



In the most advanced versions of these systems, the bilingual dictionary is associated with a morphological analyzer that processes the basic forms of the language and with a rudimentary grammar that processes the syntactic rules. However, despite this evolution, the final result remains disappointing: agreements not respected, literal translation, structure layers, erroneous equivalences, unresolved ambiguities, etc.

Examples of early direct MT systems include the English -French weather forecast system Météo (1976), the Chinese–English MT system CULT (1968) and the old Systran system (1968).

4.2 Rule-based Approach

Ruled-based machine translation (RBMT) is defined by Carl and Way (2003) as “the application of morphological, syntactic and/or semantic rules to the analysis of a source-language text and synthesis of a target-language text “(p. xviii).

Stated differently, RBMT is a syntactic method of translation which relies on a multitude of built-in linguistic rules and bilingual dictionaries to accurately translate specific content. In order to do so, the software translates the text indirectly through a parsing phase, converted into a transitional representation, from which the target translation is generated. RBMT does not translate isolated words by looking up their equivalents in a dictionary, but proceeds to the transfer of a certain structure of the source sentence to an equivalent structure in the target language. Two main subtypes are to be distinguished: *Transfer* and *Interlingua*.

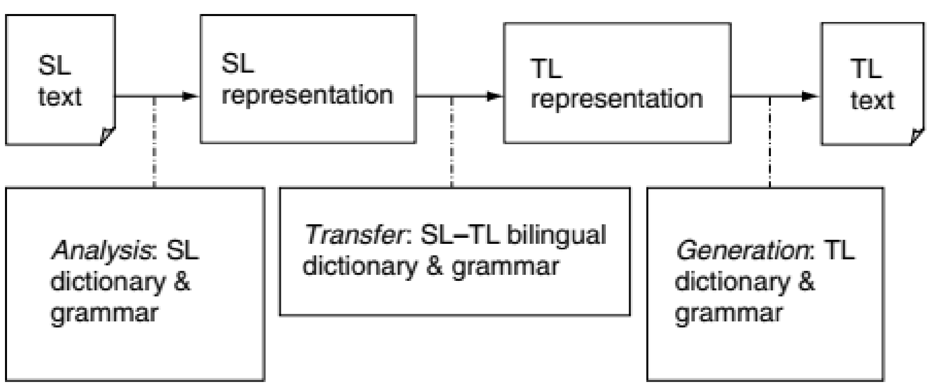
4.2.1 Transfer Approach

Transfer systems are more complex than direct translation systems because they first perform a detailed syntactic analysis of the sentence to be translated. The translation is performed at the level of syntactic structures, which eludes the simple reordering of words, (the main weakness of the direct approach), since a simple reordering is often not sufficient to obtain correct sentences in the target language. In this approach, the system translates the text indirectly through a formal representation phase of the syntactic structures of the two languages in question. Thus, the system does not translate isolated words by searching their equivalents in a dictionary, but proceeds by "transferring” the structure of the source sentence to an equivalent structure in the target language.

Transfer MT breaks translation into three steps (figure8): ***analysis*** of the source language text to determine its grammatical structure, ***transfer*** of the resulting structure to a structure suitable for generating text in the target language, and finally ***generation*** of this text. (Jurafsky & Martin, 2009, pp. 906-908)

Figure8

Transfer approach



Note. From Quah,2006, p.74

Because of the qualitative leap it represents compared to the "direct" approach", this conception of machine translation has been endorsed for the development of the so-*called "second generation" systems "*. It is also the most dominant approach and continues to be the backbone of many commercialized software.

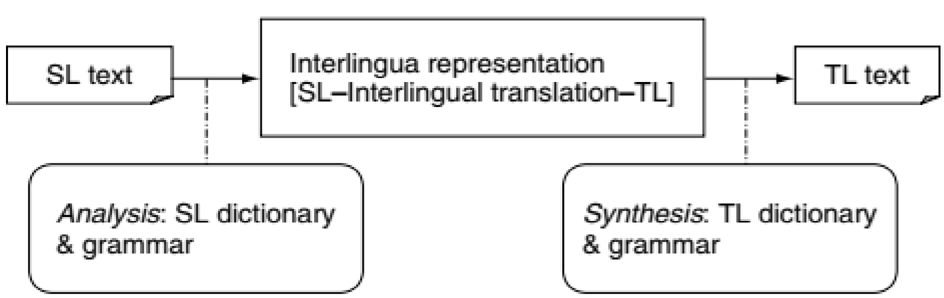
The focal point of transfer approach lies in the application of different formal rules for each pair of languages: during the analysis phase, the syntactic rules of the source language are implemented; whereas in the generation phase, it is the syntactic rules of the target language that are applied. Thus, the approach avoids many of the pitfalls of the direct approach, in particular, the non-respect of agreements and verbatim translation. The quality of the translations resulting from this approach remains dependent on the degree of precision and abstraction of the syntactic rules used for cross-linguistic transfer.

4.2.2 Interlingua Approach

Translation in this approach is conceived from a "universalist", philosophical perspective. By exploring the possibility to put into practice an abstract representation ("interlanguage"), which is valid for several language pairs, interlingua is intended to function as an intermediary "universal language" between natural languages. "During the analysis stage, a source-language text is analyzed and transformed into its interlingua representation. Target language sentences are produced from this interlingua representation with the help of target-language dictionaries and grammar rules during the synthesis stage (Lewis,1992, p.78). Figure 9 illustrates the whole process.

Figure 9.

Interlingua approach



Note. From Quah,2006, p.72

In practical terms, interlingua systems use highly abstract language-independent representation in order to translate texts between different languages. The text is first converted into an interlingua (intermediary non-natural language) and then a target translation is generated from this abstraction.

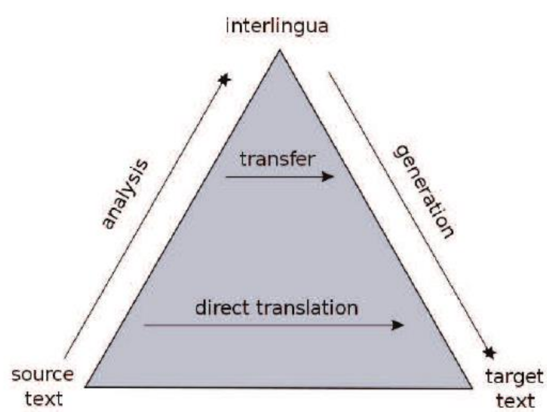
Despite the development of a few experimental prototypes, this approach has not yet resulted in an operational system, mainly due to its very high level of abstraction.

The main interest of this approach lies in the possibility to reuse the same representation for several linguistic combinations, thus saving - in theory – the effort of elaborating a set of rules specific to each language pair. However, this is only possible if a high degree of abstraction and systematization is adopted.

In short, this type of approach is potentially exploitable for certain texts presenting a great lexical and syntactic homogeneity (medical, legal, etc.), but it remains utopian for the treatment of the others.

Figure10.

Vauquois' triangle of MT approaches



The three types of approaches considered above are well visualized in Vauquois` triangle which describes the classic strategies to machine translation. As shown in (Figure 10), The depth of analysis required for each approach increases as one moves from direct to interlingual, while the amount of knowledge transfer required decreases as one moves from direct to interlingual.

4.3 Corpus-Based Approach

Corpus-Based Approach characterizes machine translation systems based either on the statistical calculation of linguistic occurrences or examples matching. These systems explore the resources of large bilingual parallel corpora (bi-text) to generate relevant equivalences. This approach favors effective and authentic uses of specific speech and does not seek to model the potential achievements of the system. The combination of statistical and linguistic rules should result in translations that are more accurate. As Carl (2000) clarified:

> All corpus-based machine translation systems use a set of so-called 'reference translations' containing source language texts and their translations. Source and target-language texts are aligned and the equivalent translation is extracted using a specific statistical method or by matching a number of examples extracted from the corpus (p. 997).

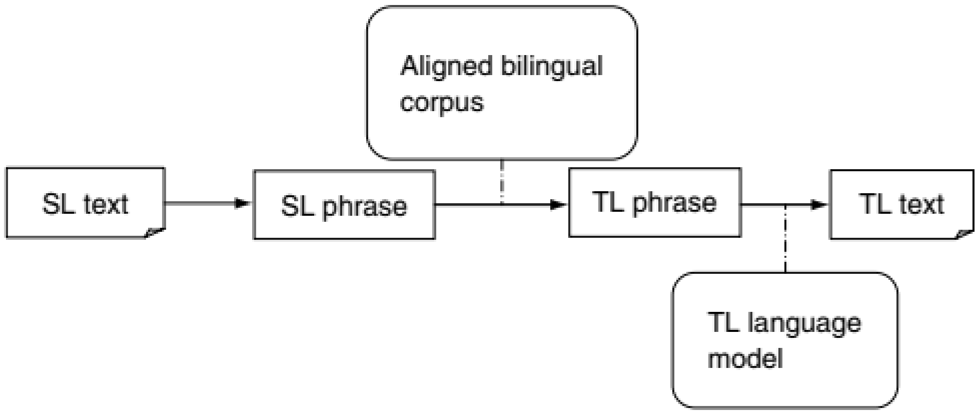
4.3.1 Example-based approach (Translation by Analogy)

Example-based machine translation (EBMT) was first introduced by Makoto Nagao in 1984. Nagao proposed the Example-based approach in order to replace second-generation systems, which were, at the time, costly and requires extensive syntactic analysis. By observing the process of translation among professional translators, Nagao concluded that human translators tended to exclude word for word translation and generally proceeded by fragment without prior analysis of the sentences. (Poibeau 2017). Instead of building large and complex bilingual dictionaries and wasting precious time on tedious preliminary analysis of sentence chunks, Nagao proposed to benefit from the valuable information contained in parallel corpora. Thus, he recommended the use of previously translated fragment stored in bi-texts. Quah (2006) judiciously explained the modus operandi of the system.

> An example-based machine translation requires a bilingual corpus of translation pairs and employs an algorithm to match the closest example of a source-language segment to its target-language segment as the basis for translating the new source text. A matched pair of segments is called an 'example'. A segment can be of any length or operate at any linguistic level […] the basic idea of an example based translation is to 'translate a source sentence by imitating the translation of a similar sentence already in the database'. (p.81)

Figure 11.

Example based approach



Note. From Quah, 2006, p.81

As depicted in Figure 11, EBT privileges the authentic uses of specific discourses derived from corpora rather than attempting to model the language system in general. However, the stumbling block of this system remains the corpus from which translations are retrieved. This corpus is not always validated or immediately available and its creation is often costly in terms of time and energy. Nonetheless, the corpus remains essential in order to ensure seamless functioning of the system and must be developed with precision and care.

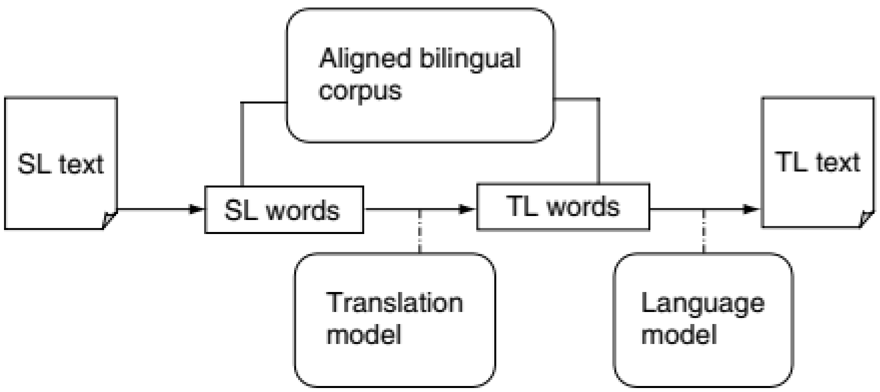
4.3.2 Statistical -based approach

The availability of large body of texts on the Internet and the continuous development in computing capacities as well as in translation memory databases have all contributed to the appearance of the very promising *Statistical -based approach*. Most of the current translation systems, especially the most popular ones such as *"Google Translate"*, *"Bing Microsoft Translator"*, are based on statistical calculation. These systems are no longer relying on dictionaries and grammar rules that have been designed by linguists, but on "the alignment of texts [...] that are structurally matched often at sentence level. Statistical calculations are then performed on the aligned bilingual texts to establish the probabilities of various translation equivalents." (Quah, 2008, p.64)

The first statistical translation systems rested on the automatic identification of word-level equivalences between source and target languages, thanks to large bilingual corpora "aligned" at the sentence level (i.e., each sentence in the source language has its corresponding translation in the target language. Large international institutions such as the UN, the EU along with the Internet can provide this type of data in large quantities, especially if English is the source or the target language.

Figure 12.

Statistical-based approach



Note. from Quah, 2006, p.81

This type of analysis has since been considerably refined. Today, it is possible to identify groups of frequent words (compound words, common phrases, collocations) that are best translated as a whole and for which an equivalent in the target language can easily be found (see Figure 12). Subsequently, translation systems are presently dealing with fragments of translation from which coherent sentences are expected to be formed. However, this last step has long been the Achilles heel of statistical translation: The point here was to build a coherent sentence from disparate fragments, as if the system were trying to recreate a clear picture from several jumbled puzzles.

4.4 Hybrid Approach

Motivated by the inability of the previous systems to achieve the optimum level of accuracy when used separately, Hybrid machine translation (HMT) takes advantage of the strengths of both statistical and rule-based translation approaches. The translations are performed using rules based engine and statistics are used in an attempt to adjust and correct the output from the rules engine.

Several machine translation organizations claim a hybrid approach that uses both rules and statistics but also other combinations such as rule-based MT CAT2 (Constructors, Atoms and Translators) and examples-based (EDGAR) systems (Carl et al. 2000), or an example-based engine and an interlingua approach in one single system called Pangloss Mark II (Bel et al. 2001).

More recently, with the advent of Neural MT[5], a new version of hybrid machine translation is emerging, combining the benefits of rules, statistical machine translation and neural machine translation . The approach allows for pre- and post-processing in a rule-driven workflow as well as taking advantage of NMT and SMT. The drawback is the inherent complexity that makes the approach suitable only for specific cases.

Summary of key points

Table 1

Strength(s) and Weakness (es) of Mt Approaches

APPROACH	STRENGTH	WEAKNESS
Corpus-Based	• It can mimic the style of the training data to generate output based on the frequency of patterns allowing them to produce more fluent output	• Creating High quality bilingual corpora is expensive and time consuming.
Rule-Based	• It is built with much less data instead uses language rules and dictionaries. • It achieves good quality results.	• Language is constantly changing, which means rules must be managed and updated where necessary. • It sometimes lacks fluency
Hybrid	• It is more effective, flexible and still the leading approach. It is combination of rule- based and statistical-based approach strengths.	• It is costly.

[5] - According to Bahdanau et al. (2015), "NMT models are end-to-end models (they need no intermediate steps to perform translation) that jointly learn to align and translate, requiring no explicit segmentation between the encoder and the decoder" (p. 1).

5. Computer Assisted Translation (CAT)

Introduction

As we have explained in the previous chapters, MT has many limitations. In certain situations, the amount of time spent by the translator reading and correcting automated translation is greater than the time invested in doing the translation itself. CAT, on the other hand, is very relevant for the translator who still has the main role of translating or validating the translation proposed by the translation memory. While this software ensures the harmonization of translations, its use also increases the quality and volume of the translation process and helps to reduce costs and turnaround times in the long term.

5.1 CAT: A brief historical overview

Computer-assisted translation (CAT) harks back to the 1950s. Early machine translation used simple grammatical rules to translate sentences from one language to another; however, their use was very limited due to the high cost and the low quality of the translation. It is until the 1990s that CAT tools began to develop, especially with the introduction of translation memories (TMs), which store previously completed translations for reuse in new translation projects (Foster, 2010). CAT software also became more affordable and user-friendly, leading to greater adoption by professional translators (Carl and Way, 2015).

During the 2000s, CAT continued to grow with the emergence of new tools, such as speech recognition and online translation (Specia et al., 2018). Advances in AI technology have also improved the quality of machine translation (Koehn, 2017).

5.2 CAT basics

CAT tools are based on the use of technology to help the translator produce high-quality translations more quickly and efficiently. According to Caignon (2012) a CAT tool relies on three basic principles:

The use of translation memory: A CAT software stores all previous translations performed by the translator in a translation memory. This memory is used to suggest translations for similar sentences in the text being translated saving the translator time by not having to translate the same sentences multiple times.

The use of terminology: A CAT tool also allows translators to add glossaries and terminology databases to ensure that domain-specific terms are used consistently throughout the translation. This ensures that translations are accurate and professional.

The use of filters: a CAT tool uses filters to extract the text to be translated from various file formats, such as PDF, HTML and XML files. This allows translators to easily translate the text without having to copy and paste the content into a separate document.

5.3 Core Components

A computer-aided translation (CAT) system can vary depending on the method and technology used for its conception. Below is a compilation of components that are frequently utilized in a CAT systems:

- Morpho-syntactic analyzer: analyzes grammatical sentence structures and morphological marks to identify parts of speech and word relationships.
- Bilingual dictionary: stores translations of words, phrases and idioms in both source and target languages.
- Corpus Aligner: aligns parallel texts, i.e., texts in both source and target languages, to allow comparison of structures and identification of possible translations.
- Translation engine: uses grammatical rules, statistical models, machine learning techniques and/or neural translation to generate translations
- Target Generator: produces the final translation by assembling the segments translated by the translation engine and applying the rules of the target language.
- Post-Editor: Allows human translators to proofread, edit and improve the quality of translations produced by the CAT system
- Terminology tool: manages specialized terms and acronyms to ensure terminology consistency in translations

These components can be integrated into a complete CAT system or used independently for specific assisted translation tasks.

Two main translation tools in workstations will be described here, namely translation memory systems and terminology management systems.

5.4 Translation-Memory Systems

5.4.1. The idea behind using TM

Translation Memory systems (TMS) were developed in the late 1980s and early 1990s to address the challenge of locating previously translated content in an organized and accessible format. Prior to TM systems, many translators did not keep archives of previous translations, and those who did collect them unsystematically, often on paper or in file cabinets. One of the key advantages of TM systems is that they can significantly reduce the time and effort required to translate new content. By leveraging previously translated content, translators can focus on the unique and challenging aspects of the content, rather than re-translating content that has already been translated. Additionally, TM systems can improve the consistency and accuracy of translations by ensuring that common terminology and phrasing are used consistently throughout the translated content. As the demand for translation services continues to grow, TM systems will play an increasingly important role in streamlining the translation process and improving the quality of translations.

5.4.2. What is a Translation Memory?

A TM (Translation Memory) has been defined by Bowker & Fisher (2010) as

> A tool that allows users to store previously translated texts and then easily consult them for potential reuse. To permit this, the source and target texts are stored in a TM database as bitexts. An aligned bitext is created by first dividing the texts into segments –which are usually sentences – and then linking each segment from the source text to its corresponding segment in the translation. (p.61)

Stated differently, a TM is a memory that stores previously translated segments in a database and assists translators in translating new contents by suggesting matches based on the similarity between the source text and the target sentence *(translation unit)*. Hence, by storing previously translated content in a structured and standardized format (also known as *leveraging)*, TM allows translators to easily locate and reuse previously translated documents, saving time and increasing consistency. Some examples of TMS are Trados Workbench, DéjàVuX, SDLX, Star Transit, MultiTrans, Similis, MetaTexis…

Figure 13

Display of translation units in SDL Trados 2019

5.4.3 How does a TM work?

TM proceeds by comparing the new segment with previously translated segments in its database. Every time a new segment is inserted, the TM system searches its database to highlight similarities between new and old segments. If any is detected, the TM system presents the previous translation to the translator so that s/he determine whether to integrate it into the new translation or not.

5.4.3.1 Segmentation

Segmentation is the process of dividing the source text into units such as words or strings of words in order to be stored in the TM database. These "translation units" serve as the basis for finding matches when a new text is to be translated. (Quah ,2006).

Segmentation is an important step in the TM creation process, as it determines the size of the segments that will be stored in the database. On the whole, segmentation is performed so that each segment of text represents a coherent semantic unit. This means that full pair of sentences are often chosen as translation units, but shorter segments such as headings, lists and bullet point can also be used, depending on the desired level of granularity. (Quah, 2006)

In the same vein, Bowker (2002) pointed out the difficulties the system encounters in identifying complete sentences. Punctuation marks such as periods, exclamation points, and question marks are generally used to indicate the end of a sentence, but problems arise in other cases such as abbreviations, periods in decimal numbers or in numbered section headings. The ellipsis, colon, and semicolon are additional punctuation marks that can be adopted as end-of-segment markers. Finally, non-Latin-script languages may present further obstacles to the segmentation process.

In order to address those issues, Bowker (2002) proposed the incorporation of *stop lists*[6] into the TM system. Other segmentation decisions can be left to the user appreciation.

Table 2

Types of segmentation

Heading	EN:	Warning:
	FR:	Avertissement :
One sentence translated by one sentence	EN:	This computer program is protected by copyright law and international treaties.
	FR:	Ce logiciel est protégé par les lois et les traités internationaux sur le droit d'auteur.
One sentence translated by two sentences	EN:	Unauthorized reproduction or distribution of this program, or any portion of it, may result in severe civil and criminal penalties, and will be prosecuted to the maximum extent possible under the law.
	FR:	Toute reproduction ou distribution partielle ou totale, par quelque moyen que ce soit, est strictement interdite. Toute personne ne respectant pas ces dispositions se rendra coupable du délit de contrefaçon et sera passible des sanctions pénales prévues par la loi.
Two sentences translated by one sentence	EN:	The "0" button and the "1" option affect the current application. The other options affect all applications.
	FR:	Les modifications apportées par le bouton « 0 » et l'option « 1 » n'affecteront que l'application en cours alors que les autres options seront répercutées dans toutes les applications.

Note. From Bowker,2002, p.96

[6] - Stop-lists are lists of words or phrases that the system should ignore when segmenting the text into translation units. They can include common abbreviations such as "Mr." or "Dr.", as well as acronyms and frequently used expressions such as "etc.", "e.g." or "in other words".

5.4.3.2 Matches

The term "matching" generally refers to an approach used in TM and Cat tools which consists in detecting linguistic similarities between different TUs in the database. The process is automatically executed via algorithms that calculate the *sensitivity threshold*[7] (degree of similarity) for each pair of segments. Several types of matching are to be distinguished (Table3), but the most recognized ones are the exact match and the fuzzy match.

Table 3

Types of matches

Exact match	A segment from the new text is identical in every way to one in the TM database.
Full match	A segment from the new text is identical to one in the TM database save for proper nouns, dates, figures, etc.
Fuzzy match	A segment from the new text has some degree of similarity to a segment stored in the TM database. Fuzzy matches can range from 1% to 99%, and the threshold can be set by the user. Typically, the higher the match percentage, the more useful the match; many systems have default thresholds between 60% and 70%.
Sub-segment match	A contiguous chunk of text within a segment of the new text is identical to a chunk stored in the TM database.
Term match	A term found in the new text corresponds to a termbase entry in the TM system's integrated TMS.
No match	No part of a segment from the new text matches the contents of the TM database or termbase. The translator must start from scratch; however, the new translation can itself go into the TM for future reuse.

Note. Bowker, L., & Fisher, D., 2010, p.61

Several different approaches are employed to extract translation segments from the previously stored texts. However, the most common types are exact and fuzzy matches.

[7] - According to Bowker (2002), S*ensitivity threshold* is the percentage of a match that must be achieved in order for a segment stored in the translation memory to qualify as a fuzzy match for a segment in the new source text.

5.4.3.2.1 Exact match

A perfect or exact match occurs when a new source language segment is completely identical including spelling, punctuation and inflections, to the old segment found in the database, that is in the translation 'memory' (Austermühl, 2001, p. 136). i.e., two segments are said to be exactly matching each other solely when the new segment is 100% identical to the old one in terms of spelling, inflection, numbering, punctuation and formatting, as shown in Table3.

Table 4

Example of Exact match

Sentence	English (source)	Arabic (target)
Old	ES6.1: delete the document.	AS6.1 قم بحذف المستند.
New	ES6.2: delete the document.	AS6.2 قم بحذف المستند.

ES= English sentence; AS= Arabic sentence

Regarding the rest of the segments which are not perfectly identical to the original segments, the system will treat them as *fuzzy matches.*

5.4.3.2.2 Fuzzy match

When a perfect match cannot be found in the database of a TM, it is possible to search for a less accurate match. To do so, the translator simply lowers the fuzzy match percentage to a value below 100% and the database will retrieve the segments that are similar but not identical according to the sensitivity threshold.(Bowker,2002)

The main objective of this method is to assist the translator by helping him avoid repetitive, daunting tasks and save precious time. On the other hand, it should be noted that fuzzy matching is not intended to replace the human translator. This approach is mainly used as a support to accelerate the translation process through the manipulation of the sensitivity threshold option which ranges between 1% to 99%.

Table 5

Example of fuzzy match

Sentence	English (source)	Arabic (target)
Old	ES6.1: delete the document.	AS6.1 .قم بحذف المستند
Old	ES6.2: New files are added	AS6.2 تم إضافة ملفات جديدة
New	ES6.3: Print the document	AS6.3 .اطبع الملف
New	ES6.4: Old files are added	AS6.4 .تم إضافة ملفات قديمة

ES= English sentence; AS= Arabic sentence

5.4.4 Creating a TM

The two main methods that are generally used in order to create TMs are Interactive translation and post-translation alignment.

5.4.4.1 Interactive translation

The term *Interactive translation* is used when a human translator carries out translations manually and stores the resulting translation in the TM so that every time a new segment is translated, the paired unit is preserved in the database and becomes part of it. (Bowker,2002)

Throughout the translation cycle, the human translator may interact with the system by taking back the suggested translations or by adjusting them when deemed required. This active mechanism generally result in a consistent, more accurate TM and significantly higher-quality database as the whole process is under the control of the translator. Probably, the main issue of this approach is related to time and the effort that requires to build a sizeable or TM.

5.4.4.2 Post-translation alignment

As its name implies, *post-translation alignment* involves aligning[8] existing translation (On the condition that these texts exist in an electronic version). In order to identify similar segments *post-translation alignment* relies heavily on statistical algorithms which use several techniques

[8] - Alignment is the process of comparing a source text and its translation, matching the corresponding segments, and binding them together as translation units in a TM. (bowker,2002,p.109)

for automatic alignment including *natural correlation between the lengths of translated segments* (analyzing the lengths and the structures of segments in both language in order to build correspondences), *cognate word pairs* (words sharing the same spellings, pronunciations, and meanings in two different languages) and *known translations* (using existing TUs as a reference in automatic alignment).

It should be noted, however, that creating a strong TM using the two aforementioned approaches is highly dependent on the quality of the input (the more accurate the pairs of sentences, the more robust the TM). Consequently, it is then necessary to assure that the source translations are of optimum quality and that the TM is incrementally updated.

5.4.5 Texts that are suitable for use with a TM

TMs expedite translation process and provide greater consistency in communication by eliminating the risk of repetitive translations for the same segments. However, the effectiveness of a TM is particularly noticeable with large volumes of text or projects containing *internal repetitions* such as technical, scientific and legal text. Another example of texts that is suitable for TMs is *Revision* which " is an amended version of a previous text" such as manuals and webpages. (Bowker,2002, p.113). *Recycled texts* can also benefit from TMs as they may treat the same subject matter or contain similar passages from other texts (external repetition). Finally, *updates* can usefully take advantage from TMs qualities when the project is still in progress. Unlike human translation in which it is very hard to notice small changes in the original text during the translation process , a TM is very useful in detecting any update and making comparison between the old-new segments .

5.4.6 Benefits and Drawbacks of TM

A translation memory is a practical tool not only for the translator, but also for the client who may benefits from it and wants to obtain quality translations at a lower cost. However, like any product, it has also some potential limitations that are worth mentioning.

5.4.6.1 Benefits

There is a plethora of advantages that can be derived from adopting a TM. In the following paragraphs, we present some of them.

Saving time

One of the most commonly cited advantage of a TM is gain of time. In fact, using a TM can save precious time for the translator and faster translation delivery. It is above all a matter of mastery: the more acquainted the translator with the use of the TM, the quicker would be the time delivery. A TM is therefore very interesting to set up for clients with recurring translation needs.

Translation quality improvement

A TM compiles all the work done by the translator through the alignment process, providing him/her with the possibility to consult the translations at any time, making changes, but most importantly maintaining the same terminology throughout the whole project. This allows the translator to ensure consistency in terminology and style. Nonetheless, it is essential to emphasize that in order to obtain a high-quality translation, the human input should be correct and of good quality in the first place.

Saving/ earning money

TMs help both client and translator saving/earning money in the long run. Thanks to the matching process (*exact vs fuzzy*), TMs allow clients to avoid being billed multiple times for recurring or even redundant passages in their translation projects. Similarly, TMs help translators earn decent money particularly when in possession of larger databases.

5.4.6.2 Drawbacks

Although TMs are powerful tools of memorization and organization, they also have some drawbacks.

Time consuming

learning how to use a CAT tool in general is time consuming. It takes weeks and month for a novice translator to become familiar with the interface and functionality of the tool and years of practice to become proficient.

Cohesive issues

Translation Scholars such as Baker (1992) and Hatim and Mason (1990) underlined the importance of cohesion for effective translation. They both emphasized that a text is viewed as a block and cannot be treated in terms of separate sentence processed in isolation. When applied in TM, it seems that this aspect is not given enough attention since TMs align texts segments by segment without consider their inner relationships or the wider context. Hence, this aspect should be taken into consideration for the future elaboration of the TUs.

Tools transition

Transition from one MT to another is another problem facing this technology. Translation market abound with a wide range of products that are not compatible with each other and may use with different formats to store their translations (e.g., SDL Trados vs Omega).

Languages with peculiar

Bowker (2002) pointed out that some language are more difficult to process than others such as Arabic and Hebrew , Chinese, Japanese, and Korean(also known as *double-byte languages*). All these languages have unique segmentation challenges that may affect the quality of the translation.

Other constraints

Other constraints may be relevant to the issue of TM` legal ownership and costs[9] between the client and the translator, the high cost of CAT tools software (Good systems are fairly expensive. e.g., SDL Trados ($408 to $1,500) as well as the technical challenges of converting hard copy texts to electronic format (scanner, OCR, Voice recognition, alignment programs...)

5.5 Terminology Management Systems (TMS)

The use of electronic media to manage terminology and associated data dates back to the 1960s. This recourse is explained by the mass of data to be processed in certain organizations, but also by the nature of the terms that seem to lend themselves well to rigorous and systematic encoding. The

[9] - Some Clients, for instance, refuse to pay translators the full rate for reusing previous translations(fuzzy match), which is regarded as unacceptable by many translators who view that both fuzzy and exact matches need proofreading and analysis in context before being incorporated into the new translation.

electronic medium has even strongly influenced the techniques for describing terminology data in specialized repositories, since encoding standards and the computer medium almost always went hand in hand.

5.5.1 Important distinctions

Terminology management system (TMS)

TMS is a computer-based tool designed to manage terminology in a systematic and structured way, providing a repository for storing, organizing and sharing terminology resources (Hajlaoui & Khennouf, 2017). Examples of TMS are SDL MultiTerm, LogiTerm and Termex.

Term

A word, phrase associated with a particular company, brand, vertical market or field of activity**.**

Terminology

The study of terms and their use.

Termbase

database which includes the term, translations and associated metadata, such as a description or definition, the rules for usage or formatting.

Dictionary, TM and Term base

A dictionary provides definition for words in one language, while a termbase stores words with translations, descriptions and usage rules in different language. Finally, a TM stores translations of segments for potential use.

5.5.2 Main functions

5.5.2.1 Storage

The Terminology Management System (TMS) is an essential tool for consolidating and storing terminology information used in translation projects. Older TMSs stored this information in structured text files with one-way mapping, which caused problems for translations in different language directions. Newer TMSs use a relational model to store information in a more conceptual manner, allowing for mapping across multiple languages.

Modern TMSs also offer more flexibility in terms of the information fields that can be stored on each terminology record, unlike previous TMSs where fields were often fixed. This allows users to store more relevant information for each record, rather than being limited to a predefined number of fields and characters.

5.5.2.2 Retrieval

in order to search and retrieve the terminology needed by the translator several methods and techniques are used by TMSs, mainly

5.5.2.2.1 Exact match

Considered by many scholars as the basic and the most direct technique to retrieves terms that exactly match the query character by character. (Cabré, 1999, p. 114. This approach is often used in TMS to retrieve precise terms, allowing the highest level of precision in retrieval results and ensuring that users can quickly and easily find the terms they are looking for with minimal ambiguity.

5.5.2.2.2 Stemming

“ Stemming is a technique used in TMS to retrieve words that share the same root or stem as the query word, even if they are not exact matches" (ISO 12620:2018, p. 8). In concrete terms, this approach allows morphological variants during retrieval process (e.g. speak, spoke, spoken, speaker, speaking). Stemming is then an efficient way to improve the search results that have been overlooked by exact match. By taking into account subtle variations in languages and declensions at the formal level, *stemming* expands the scope of search and enables the retrieval of relevant information

5.5.2.2.3 Wildcard search

In a wildcard search "TMS allows users to search for words using partial matches or patterns" (Mellado & Cobo, 2015, p. 30). In other words, and as explained by Bowker (2002) a translator can use any character to represent one or more characters in search string, such as using an asterisk to refer to different character patterns. (e,g "translat*" would retrieve "translate," "translated," "translates," "translating," "translation," "translationese," "translations," "translator," "translators," and so on.

5.5.2.2.4 Fuzzy search

Fuzzy search retrieves words that are similar to the query word, taking into account factors such as misspellings, typographical errors, and variations in word order" (Biel & Marquès, 2014, p. 110).

5.5.2.2.5 Semantic retrieval

"Semantic retrieval is a technique used in TMS to retrieve words based on their meaning rather than their form, using techniques such as natural language processing, machine learning, and ontologies" (Wüster, 1979, p. 32).

5.5.3 Other features

5.5.3.1 Active terminology recognition

Active terminology recognition is described by Bowker (2002) in the following terms

> A type of automatic dictionary lookup. The terminology-recognition component compares items in the source text against the contents of the term base, and if a match is found, the term record in question is displayed for the translator to consult. (p.141)

To put it differently, *active terminology recognition* is the process that automatically identifies and extracts important terms and phrases from a given text while the translator is typing (AutoSuggest).[10] It involves recognizing key concepts and vocabulary used in a particular domain and understanding their contextual meaning. Active terminology recognition is often used in areas such as information retrieval, natural language processing and machine learning to improve text analysis and classification tasks. This process can also help communication and understanding within a specific domain by promoting consistent use of terminology.

5.5.3.2 Term extraction

Term extraction, also known as *term-recognition* or *term identification tool* is a functionality in a TMS that identifies key terms in a source text and adds them to a terminology database. This

[10] - In the case of SDL Trados 2019 , as soon as the user start typing the first letter of a term stored in the term base (terminology suggestions are displayed on the top right) the system automatically offers the full term, which can be inserted manually in the target segment by pressing the ENTER key. (SDL Trados Studio 2019 user manual, n.d., Chapter 13, para. 3)

functionality can be monolingual or bilingual and uses algorithms to identify candidate terms from a corpus of text. The identified terms can then be added to the terminology database for reuse in future translation projects. However, the list of candidate terms must be verified and validated by a human translator to ensure the accuracy of the terms added to the database. This feature can help speed up the process of creating a terminology database by automating the task of identifying key terms.(Bowker,2002)

5.5.3.3 Additional features

- create and manage concept systems or thesauri,
- Merge multiple term bases,
- import from or export to other formats,
- print out the contents of a term base in a user-specified glossary format.

5.5.4 Benefits and drawbacks

5.5.4.1 Benefits

Speed & flexibility

TMSs offer several advantages in terms of speed and flexibility in the translation process. First, the use of a terminology database allows translators to quickly and easily access pre-defined key terms and phrases, saving time when translating similar documents. In addition, *term extraction* feature allows key terms in a source text to be quickly identified and added to the terminology database for future use, making it easier to translate similar documents in the future. Finally, TMSs allow for flexibility in terminology management. Terms can be added, modified or deleted from the terminology database according to the needs of the user, allowing the terminology used to be adapted to changes in the market or customer preferences.

Shareability of information

TMS can be networked between different users in a translation team or between different teams working on similar projects, allowing for consistent use of terminology throughout the organization and rapid updating of the terminology database. (Bowker,2002)

With respect to CAT tools, Bowker (2002) explained that both TMS and CAT tools can be interconnected with each other, allowing CAT tools to access and utilize the terminology resources stored in the TMS while improving the quality and consistency of translations. The integration can be achieved through different standard formats such as Term Base eXchange (TBX), MARTIF - ISO12200 and Open Standards for Container/Content Allowing Reuse (OSCAR). Those softwares enable a seamless exchange of data between the TMS and CAT tools, resulting in a streamlined workflow for translation teams.

5.5.4.2 Drawbacks

In addition to character set limitations of languages with double-byte character sets (DBCS) like Chinese and Japanese, TMS can be costly, especially if they are integrated with a top-tier CAT tool. Costs can vary depending on several factors, such as the complexity of the system, the size of the organization, the number of users and the functionality required. (Bowker,2002)

References

Arnold, D., Balkan, L., Meijer, S., Humphreys, R. L., & Sadler, L. (1994). Machine translation: An introductory guide. Oxford: NCC Blackwell.

Arnold, D.J. (2003). Why translation is difficult for computers. In H. Somers (Ed.), Computers and Translation: A translator's guide (pp. 119-142). John Benjamins.

Automatic Language Processing Advisory Committee (ALPAC). (1966). Language and Machines: Computers in Translation and Linguistics (Tech. Rep.). Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences and National Research Council.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Baker, M. (1992). In other words: A coursebook on translation. Routledge.

Bel, N., Dorr, B., Hovy, E., Knight, K., Iida, H., Boitet, C., Maegaard, B., & Wilks, Y. (2001). Chapter 4: Machine Translation. In E. Hovy, N. Ide, R. Frederking, J. Mariani, & A. Zampolli (Eds.), Multilingual Information Management: Current Levels and Future Abilities. Retrieved February 2003, from http://www-2.cs.cmu.edu/~ref/mlim/chapter4.html.

Biel, L., & Marquès, L. (2014). Terminology management in the cloud: the TermNet experience. In M. Temmerman & K. Kerremans (Eds.), Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity (pp. 109-116). EURALEX.

Bowker, L., & Fisher, D. (2010). Computer-aided translation. In Y. Gambier & L. van Doorslaer (Eds.), Handbook of translation studies (Vol. 1, pp. 60-65). John Benjamins Publishing Company.

Cabré, M. T. (1999). Terminology: Theory, methods, and applications (Vol. 2). John Benjamins Publishing.

Caignon, P. (2012). Introduction à la traduction assistée par ordinateur. Presses universitaires de Louvain.

Carl, M. (2000). A Model of Competence for Corpus-based Machine Translation. In Proceedings of COLING2000, Vol. 2, Germany. Retrieved from http://acl.upenn.edu/C/C00/C00-2145.pdf

Carl, M. and Way, A. (2015). Recent advances in example-based machine translation. IEEE Transactions on Neural Networks and Learning Systems, 26(9), pp. 2008-2021.

Carl, M., & Way, A. (2003). Introduction. In M. Carl & A. Way (Eds.), Recent Advances in Example-based Machine Translation: Text, Speech and Language Technology (Vol. 21, pp. xvii–xxxi). Kluwer Academic. (Original work published 2003)

Carl, M., C. Pease, L.L. Iomdin and O. Streiter (2000) ‘Towards a Dynamic Linkage of Example-based and Rule-based Machine Translation’, Machine Translation, 16: 223–557.

Corbé, M. (1960): “La machine à traduire française aura bientôt trente ans”, Automatisme 5(3): 87-91.

Couturat, L. (1914). The Logic of Leibniz. Open Court Publishing Company.

DuPont, Q. (2018). Les origines cryptologiques de la traduction automatique: D'al-Kindi à Weaver. Amodern, (8).

Foster, G. (2010). A brief history of machine translation. Machine Translation, 24, pp. 3-8.

Hajlaoui, N., & Khennouf, D. (2017). A systematic review of terminology management systems: A bibliometric analysis. Terminology, 23(1), 29-59. doi: 10.1075/term.23.1.02haj

Hatim, B., & Mason, I. (1990). Discourse and the translator. Longman.

Hutchins, J. (2004). Two precursors of machine translation: Artsrouni and Trojanskij. International Journal of Translation, 16(1), 11-31.

Hutchins, J. (2006). The first public demonstration of machine translation: The Georgetown-IBM system, 7th January 1954. Machine Translation, 20(1), 1-13. https://doi.org/10.1007/s10590-006-9009-1

Hutchins, W. J. (n.d.). W. John Hutchins. Retrieved from http://ourworld.compuserve.com/homepages/WJHutchins

Hutchins, W. J., & Somers, H. L. (1992). An introduction to machine translation. Academic Press.

ISO 12620:2018. (2018). Language resource management — Lexical markup framework (LMF). International Organization for Standardization.

Jurafsky, D., & Martin, J. H. (2009). Speech and Language Processing. Pearson.

Knowlson, J. (1975). Schémas linguistiques universels en Angleterre et en France, 1600-1800. Presses de l'Université de Toronto. ISBN 978-0-8020-5296-4.

Koehn, P. (2017). Neural machine translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1-10.

Léon, J. (2015). Histoire de l'automatisation des sciences du langage. ENS Éditions. Retrieved from http://books.openedition.org/enseditions/3733

Lewis, D. (1992). Computers and Translation. In C.S. Butler (Ed.), Computers and Written Texts (pp. 75-113). Oxford: Blackwell.

Luong, M. T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.

Mellado, I., & Cobo, C. (2015). A methodology for the development of terminological databases in an academic context. Terminology, 21(1), 25-49.

Poibeau, T. (2017). Machine Translation. MIT Press.

Quah, C. K. (2006). Translation and technology. Palgrave Macmillan.

Sager, J. C. (1994). Language engineering and translation: consequences of automation. John Benjamins Publishing Company.

SDL Trados Studio 2019 user manual. (n.d.). Chapter 13, para. 3.

Slocum, J. (1988) ‘A Survey of Machine Translation: Its History, Current Status, and Future Prospects’, in Machine Translation Systems, J. Slocum (ed.). Cambridge: Cambridge University Press: 1–40.

Somers, H. L. (1997). A practical approach using machine translation software: "Post-editing" the source text. The Translator, 3(2), 193-212.

Somers, H. L. (2003). Introduction. In H.L. Somers (Ed.), Computers and Translation: A Translator's Guide (pp. 1-11). Amsterdam: John Benjamins.

Specia, L., Turchi, M., and Federico, M. (2018). Machine translation and quality estimation: The OMTrEX project. Berlin: Springer.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (Eds.), Advances in neural information processing systems (pp. 3104-3112). Curran Associates.

Tong, L. C. (1994). Translation: Machine-aided. In R. E. Asher & J. M. Y. Simpson (Eds.), The Encyclopedia of Language and Linguistics (Vol. 9, pp. 4730-4737). Oxford: Pergamon Press.

Weaver, W. (1949). Translation. In W. N. Locke & A. D. Booth (Eds.), Machine translation of languages: Fourteen essays (pp. 15-23). Cambridge, MA: Technology Press of the Massachusetts Institute of Technology. (Original work published 1949)

Wüster, E. (1979). Einführung in die allgemeine Terminologielehre und terminologische Lexikographie. Walter de Gruyter.