**Machine Translation: Using tools of technology for translation**

The term machine translation (MT) refers to computerized systems responsible for the production of translations with or without human assistance. A distinction is commonly made between **human-aided MT (HAMT)** and **machine-aided human translation (MAHT)**. The latter comprises computer-based translation tools which support translators by providing access to on-line dictionaries, remote terminology databanks, transmission and reception of texts, stores of previously translated texts (‘translation memories’), and integrated resources, commonly referred to as translator workstations or translator workbenches. The term **computer-aided translation (CAT)** is sometimes used to cover all these computer-based translation systems.

**Approaches to Translation:**

The human translation process may be described as:

**Decoding the meaning of the source text; and Re-encoding this meaning in the target language.**

Behind this ostensibly simple procedure lies a complex cognitive operation. To decode the meaning of the source text in its entirety, the translator must interpret and analyse all the features of the text, a process that requires in-depth knowledge of the grammar, semantics, syntax, idioms, etc., of the source language, as well as the culture of its speakers. The translator needs the same in-depth knowledge to re-encode the meaning in the target language.

Therein lies the challenge in machine translation: how to program a computer that will “understand” a text as a person does, and that will “create” a new text in the target language that sounds as if it has been written by a person. There are various types of Machine Translation:

➢ **Rule-based machine translation** (**RBMT**; “Classical Approach” of MT) is machine translation systems based on linguistic information about source and target languages basically retrieved from (unilingual, bilingual or multilingual) dictionaries and grammars covering the main semantic, morphological, and syntactic regularities of each language respectively. Having input sentences (in some source language), an RBMT system generates them to output sentences (in some target language) on the basis of morphological, syntactic, and semantic analysis of both the source and the target languages involved in a concrete translation task.

➢ **Transfer-based machine translation** is a type of machine translation (MT). It is

currently one of the most widely used methods of machine translation. In contrast to the simpler direct model of MT, transfer MT breaks translation into three steps: 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. Transfer-based MT systems are thus capable of using knowledge of the source and target languages. The methods which are chosen and the emphasis depends largely on the design of the system, however, most systems include at least the following stages:

• **Morphological analysis**. Surface forms of the input text are classified as to part- of-speech (e.g. noun, verb, etc.) and sub-category (number, gender, tense, etc.). All of the possible “analyses” for each surface form are typically made output at this stage, along with the lemma of the word.

• **Lexical categorisation**. In any given text some of the words may have more than one meaning, causing ambiguity in analysis. Lexical categorisation looks at the context of a word to try to determine the correct meaning in the context of the input. This can involve part-of-speech tagging and word sense disambiguation.

• **Lexical transfer**. This is basically dictionary translation; the source language lemma (perhaps with sense information) is looked up in a bilingual dictionary and the translation is chosen.

• **Structural transfer**. While the previous stages deal with words, this stage deals with larger constituents, for example phrases and chunks. Typical features of this stage include concordance of gender and number, and re-ordering of words or phrases.

• **Morphological generation**. From the output of the structural transfer stage, the target language surface forms are generated.

➢ **Interlingual machine translation** is one of the classic approaches to [machine translation.](https://en.wikipedia.org/wiki/Machine_translation)

In this approach, the source language, i.e. the text to be translated is transformed into an

interlingua, i.e., an abstract language-independent representation. The target language is then generated from the interlingua. Within the rule-based machine translation paradigm,

the interlingual approach is an alternative to the [direct approach a](https://en.wikipedia.org/wiki/Dictionary-based_machine_translation)nd the [transfer](https://en.wikipedia.org/wiki/Transfer-based_machine_translation) [approach.](https://en.wikipedia.org/wiki/Transfer-based_machine_translation)

In the direct approach, words are translated directly without passing through an additional representation. In the transfer approach the source language is transformed into an abstract, less language-specific representation. Linguistic rules which are specific to the language pair then transform the source language representation into an abstract target language representation and from this the target sentence is generated.

➢ **Statistical machine translation (SMT)** is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. In linguistics, a **corpus** (plural corpora) or text corpus is a large and structured set of texts (nowadays usually electronically stored and processed). In corpus linguistics, they are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory. The statistical approach contrasts with the rule-based approaches to machine translation as well as with example-based machine translation. The first ideas of statistical machine translation were introduced by Warren Weaver in 1949, including the ideas of applying Claude Shannon’s information theory. Statistical machine translation was re- introduced in the late 1980s and early 1990s by researchers at IBM’s Thomas J. Watson Research Center and has contributed to the significant resurgence in interest in machine translation in recent years. Nowadays it is by far the most widely studied machine translation method.

Statistical machine translation tries to generate translations using statistical methods based on bilingual text corpora, such as the Canadian Hansard corpus, the English-French record of the Canadian parliament and EUROPARL, the record of the European Parliament. Where such corpora are available, good results can be achieved translating similar texts, but such corpora are still rare for many language pairs. The first statistical machine translation software was CANDIDE from IBM. Google used SYSTRAN for several years, but switched to a statistical translation method in October 2007. In 2005, Google improved its internal translation capabilities by using approximately 200 billion words from United Nations materials to train their system; translation accuracy improved. Google Translate and similar statistical translation programs work by detecting patterns in hundreds of millions of documents that have previously been translated by humans and making intelligent guesses based on the findings. Generally, the more human-translated documents available in a given language, the more likely it is that the translation will be of good quality. Newer approaches into Statistical Machine translation such as METIS II and PRESEMT use minimal corpus size and instead focus on derivation of syntactic structure through pattern recognition. With further development, this may allow statistical machine translation to operate off of a monolingual text corpus. SMT’s biggest downfall includes it being dependent upon huge amounts of parallel texts, its problems with morphology-rich languages (especially with translating into such languages), and its inability to correct singleton errors.

➢ **Example-based machine translation (EBMT)** approach was proposed by Makoto Nagao in 1984. Example-based machine translation is based on the idea of analogy. In this approach, the corpus that is used is one that contains texts that have already been translated. Given a sentence that is to be translated, sentences from this corpus are selected that contain similar sub-sentential components. The similar sentences are then used to translate the sub-sentential components of the original sentence into the target language, and these phrases are put together to form a complete translation.

➢ **Hybrid machine translation (HMT)** leverages the strengths of statistical and rule-based translation methodologies. Several MT organizations (such as Omniscien Technologies (formerly Asia Online), LinguaSys, Systran, and Polytechnic University of Valencia) claim a hybrid approach that uses both rules and statistics. The approaches differ in a number of ways:

• **Rules post-processed by statistics:** Translations are performed using a rules based engine. Statistics are then used in an attempt to adjust/correct the output from the rules engine.

• **Statistics guided by rules:** Rules are used to pre-process data in an attempt to better guide the statistical engine. Rules are also used to post-process the statistical output to perform functions such as normalization. This approach has a lot more power, flexibility and control when translating. It also provides extensive control over the way in which the content is processed during both pre-translation (e.g. markup of content and non-translatable terms) and post-translation (e.g. post translation corrections and adjustments).

More recently, with the advent of Neural MT, a new version of hybrid machine translation is emerging that combines the benefits of rules, statistical and neural machine translation. The approach allows benefitting from pre- and post-processing in a rule guided workflow as well as benefitting from NMT and SMT. The downside is the inherent complexity which makes the approach suitable only for specific use cases. One of the proponents of this approach for complex use cases is Omniscien Technologies.

➢ **Neural machine translation (NMT)** is an approach to machine translation that uses a large artificial neural network to predict the likelihood of a sequence of words, typically modeling entire sentences in a single integrated model. Deep neural machine translation is an extension of neural machine translation. Both use a large neural network with the difference that deep neural machine translation processes multiple neural network layers instead of just one.