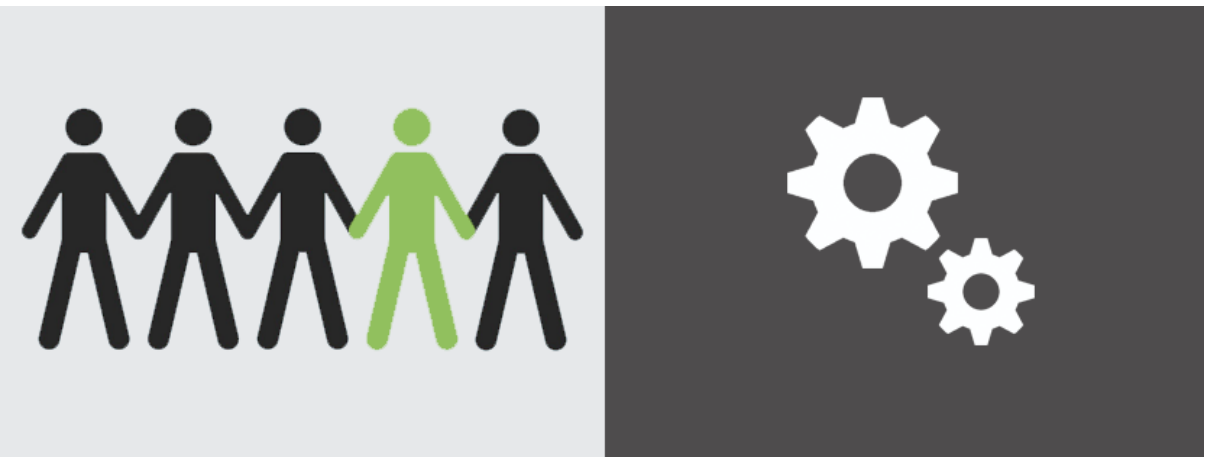
What is Machine Translation?

Machine translation (MT) is automated translation or “translation carried out by a computer”, as defined in the Oxford English dictionary. It is a process, sometimes referred to as Natural Language Processing which uses a bilingual data set and other language assets to build language and phrase models used to translate text. As computational activities become more mainstream and the internet opens up the wider multilingual and global community, research and development in Machine Translation continues to grow at a rapid rate.

A few different types of Machine Translation are available in the market today, the most widely use being Statistical Machine Translation (SMT), Rule-Based Machine Translation (RBMT), and Hybrid Systems, which combine RBMT and SMT.

Human vs. Machine Translation

In any translation, whether human or automated, the meaning of a text in the source (original) language must be fully transferred to its equivalent meaning in the target language’s translation. While on the surface this seems straightforward, it is often far more complex. Translation is never a mere word-for-word substitution.



A human translator must interpret and analyse all of the elements within the text and understand how each word may influence the context of the text. This requires extensive expertise in grammar, syntax (sentence structure), semantics (meanings), etc., in the source and target languages, as well as expertise in the domain.

Human and Machine Translation each have their share of challenges. For example, no two individual translators will produce identical translations of the same text in the same language pair, and it may take several rounds of revisions to meet the client's requirements. Automated translations find difficulties in interpreting contextual and cultural elements of a text and quality is dependent on the type of system and how it is trained, however it is extremely effective for particular content types and use cases, e.g. automotive, mechanical, User Generated Content (USG), repetitive texts, structured language and many more.

While Machine Translation faces some challenges, if implemented correctly MT users can achieve benefits from economies of scale when translating in domains suited to Machine Translation.

Rule-Based Machine Translation (RBMT) Technology

RBMT relies on countless built-in linguistic rules and millions of bilingual dictionaries for each language pair. The RBMT system parses text and creates a transitional representation from which the text in the target language is generated. This process requires extensive lexicons with morphological, syntactic, and semantic information, and large sets of rules. The software uses these complex rule sets and then transfers the grammatical structure of the source language into the target language.

Rule-based Machine Translation systems are built on gigantic dictionaries and sophisticated linguistic rules. Users can improve translation quality by adding terminology into the translation process by creating user-defined dictionaries, which override the system's default settings.

In most cases, there are two steps: an initial investment that significantly increases the quality at a limited cost, and an ongoing investment to increase quality incrementally. While rule-based MT may bring a company to a reasonable quality threshold, the quality improvement process is generally long, expensive and needs to be carried out by trained experts. This has been a contributing factor to the slow adoption and usage of MT in the localization industry.

Statistical Machine Translation (SMT) Technology

Statistical Machine Translation utilizes statistical translation models generated from the analysis of monolingual and bilingual training data. Essentially, this approach uses computing power to build sophisticated data models to translate one source language into another. The translation is selected from the training data using algorithms to select the most frequently occurring words or phrases.

Building SMT models is a relatively quick and simple process which involves uploading files to train the engine for a specific language pair and domain. A minimum of two million words is required to train an engine for a specific domain, however it is possible to reach an acceptable quality threshold with much less. SMT technology relies on bilingual corpora such as translation memories and glossaries to train it to learn language pattern, and is uses monolingual data to improve its fluency. SMT engines will prove to have a higher output quality if trained using domain specific training data such as; medical, financial or technical domains.

SMT technology is CPU intensive and requires an extensive hardware configuration to run translation models at acceptable performance levels. Because of this, cloud-base systems are preferred, whereby they can scale to meet the demands of its users without the users having to invest heavily in hardware and software costs.

RBMT vs. SMT

- RBMT can achieve good results but the training and development costs are very high for a good quality system. In terms of investment, the customization cycle needed to reach the quality threshold can be long and costly.

- RBMT systems are built with much less data than SMT systems, instead using dictionaries and language rules to translate. This sometimes results in a lack of fluency.

- Language is constantly changing, which means rules must be managed and updated where necessary in RBMT systems.

- SMT systems can be built in much less time and do not require linguistic experts to apply language rules to the system.

- SMT models require state-of the-art computer processing power and storage capacity to build and manage large translation models.

- SMT systems can mimic the style of the training data to generate output based on the frequency of patterns allowing them to produce more fluent output.

The Verdict

Statistical Machine Translation technology is growing in acceptance and is by far, the clear leader between both technologies. The increasing availability of cloud-based computing is providing a solution to the high computer processing power and storage capacity required to run SMT technology effectively, making SMT a game changer for the localization industry.

Training data for SMT engines is becoming more widely available, thanks to the internet and the increasing volumes of multilingual content being created by both companies and private internet users. High quality aligned bilingual corpora is still expensive and time consuming to create but, once created becomes a valuable asset to any organization implementing SMT technology, with translations benefiting from economies of scale over time.