



Machine Translation for Enterprises

Fabrice Chabot
Computational Linguist
Lexcelera
2015

What is machine translation?

Machine translation refers to a computer-generated translation of a text from one language to another so that it can be understood by speakers of the target language.

Why use machine translation?

There are many reasons for an enterprise to use machine translation, including:

- an explosion of new content that needs to be translated,
- international expansion, with new stakeholders located around the globe,
- a large amount of content that needs to be understood by staff, at least at a superficial level ("gisting"),
- a corporate intranet that is regularly updated,
- highly confidential content that cannot be outsourced for translation.

The purpose of machine translation tends to fall into two broad categories: assimilation (understanding a text in a foreign language) and dissemination (communicating with speakers of other languages).



Main types of machine translation

There are several types of machine translation technology:

Rule-based machine translation (RBMT)

Rule-based machine translation was the first type to appear on the market. It uses syntactic analysis and transformational grammar rules combined with bilingual and multilingual dictionaries to build a translation engine. The applications using RBMT usually include tools such as bilingual and normalization dictionaries for customizing translations.

With its grammar-based design, RBMT is relatively easy to implement, and the key to improving accuracy lies in building customized dictionaries specific to the company's business domain. RBMT tools also increase accuracy by developing syntactic rules and normalizing the text. While rule-based machine translation generally produces rather literal translations, it strictly adheres to the terminology and grammatical structure of the translated sentences.

Among the leaders in this sector are Systran, ProMT and Linguatrec.



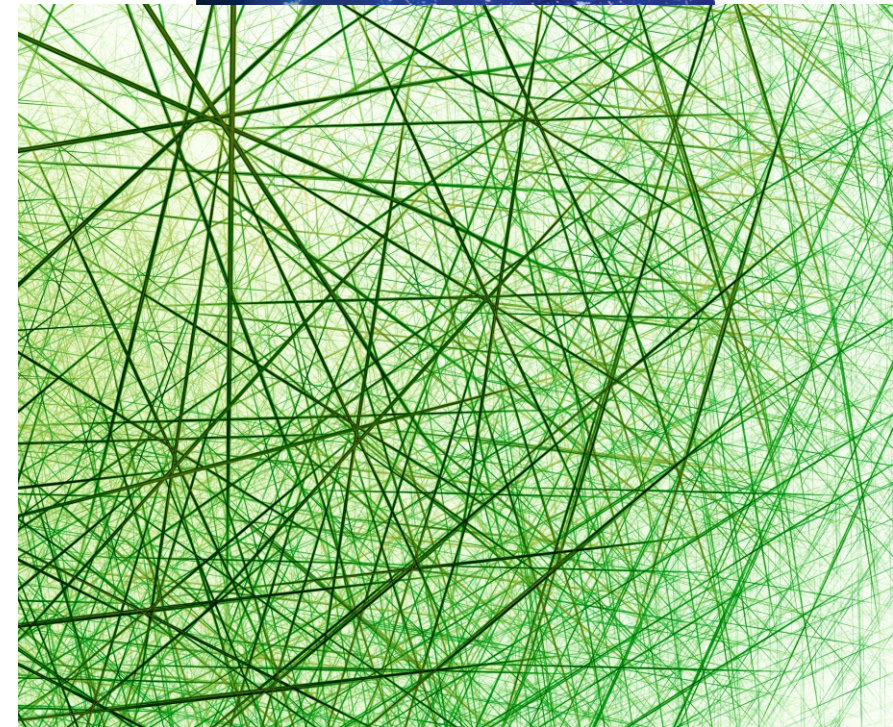
Statistical machine translation (SMT)

Statistical machine translation uses algorithms and large volumes of bilingual text – or corpora – to develop a translation engine. The translation rules are generated by pattern matching during a process in which the bilingual texts are analyzed. This is known as “training”. The quality of the results is directly related to the quality of the bilingual texts on which the engine is trained, including accurate grammar and spelling, matching of the corpora's content with the source texts, style and terminological consistency.

The systems' accuracy can be increased by adding and improving new specialized corpora (“in-domain”), which may initially seem easier to implement than a rule-based system that requires adding words to a dictionary. But it can be more complex to tune the engines, with the corpora needing to be prepared specifically for this task. Its operating method naturally produces a more fluid writing style, but the grammatical accuracy is more uncertain and the terminology is not always predictable.

Engines are available that have already been trained on out-of-domain or multi-domain corpora. When publicly available, they can be accessed online using various sales or security methods. There are also systems you can develop on your own for proprietary purposes using open-source software such as Moses.

The main players in the statistical machine translation market are Google Translate, Morphologic and IBM WebSphere.



Hybrid machine translation: technology convergence

Recent years have seen the emergence of so-called "hybrid" technologies, which draw on the best features of the two systems described above. Depending on the system, these hybrid methods may use rule-based technology supplemented by a statistical-type autocorrect capability based on large bilingual corpora. Or they may first consist of statistical technology with the ability to analyze grammar. Systran, Microsoft Translator (Bing) and AppTek all take this approach.

	Rule-based machine translation	Statistical machine translation
Basic principles	syntactic analysis rules transformational grammar rules multilingual rules	corpus pattern matching
Updates	adding terms to dictionaries adding rules	adding new prepared and approved corpora training
Translation	literal style consistency with dictionary terminology	fluid style consistency and variable syntax



Products

Open-source products

Open-source products that allow you to develop your own proprietary translation system are available for the two main types of technology.

Two examples are Open Logos and Apertium, used for rule-based machine translation. Statistical technology, for its part, has greatly benefited from the Moses development kit. Many enterprise products are now developed using Moses, including those offered by Asia Online and Kantan MT. The open-source version, however, offers very general tools and raw data, requiring extensive development of both the engine and interface to achieve effective results.

Products for LSPs (language service providers)

In most cases, translation companies with the technical skills necessary for implementation have become a logical market for the automated services provided by machine translation products. These services enable the companies to develop customized engines for their end-clients. The engines are used for the machine pre-translation of texts, which are then post-edited by professional translators. This process increases the companies' productivity and shortens completion times for large projects. It also helps them meet their end-clients' growing needs for distributing content with a short lifespan at a competitive translation cost.



On-line translation products

The first thing that often comes to mind when talking about machine translation is free online services that are relatively well developed, such as Google Translate and Microsoft Bing. These tools generally appear in the form of type-in windows but they may also be available via plugins on the browsers' toolbar or widgets directly displayed on web or intranet pages. The resulting quality is unpredictable because the tools are not based on a specific domain but on all data available online. Reasonably extensive opportunities for customization do exist, however, depending on the engine.

Integrated enterprise products

Some software makers have designed their products for integration into a company's own tools (office systems, email, etc.) so that users working abroad can take advantage of machine translation to translate documents they want to write or read.

Yet with the emergence of cloud computing, this trend will undoubtedly ebb in favour of integrating language technology into SaaS solutions.

Products integrated into office systems:

- Systran Server and Desktop Applications: plugins for MS Office and Web browsers.
- Microsoft Translator: plugin for MS Office Word, widget that can be integrated into web pages.
- Softissimo's Reverso (based on ProMT): toolbar for browsers



Other enterprise solutions

Another enterprise solution involves using existing products to provide a customized solution with two separate parts:

- firstly, the development of one or several custom-designed translation engines that draw on available technologies and vendors, depending on the required language pairs, as well as existing client content, such as glossaries, translation memories and text corpora.

These customized engines must be accompanied by methods for evaluating translation quality and how effectively it meets client needs.

- secondly, the use of connectors for integrating translation engines into the company's infrastructure, as necessary: for example, plugins for messaging and office technology applications, a widget on web pages, or a type-in window for urgent translations, depending on the specific needs.

How to Make Automation Easy

Lexcelera, a languages services provider founded in Paris in 1986, has a history of combining the excellence of human translation with the productivity of automated technologies.

A pioneer in machine translation research and use, the company is a leader in linguistic engineering solutions.

Lexcelera was also the first translation provider in France to receive ISO 9001:2000 quality certification.



Glossary

- **Corpus**

A collection of text consisting of a large number of sentences. The text may be monolingual or bilingual.

- **Dictionary building**

The process of developing a rule-based translation model using tools such as the extraction of terminology and unknown words and normalization lists.

- **Hybrid Machine Translation (HMT)**

An automatic translation system that uses multiple technologies based on both rules and statistics.

- **In-domain corpus**

In statistical machine translation, a bilingual text corpus specific to a certain domain and used for training purposes, characterized by its specific domain terminology.

- **Language model**

Software based on a set of monolingual rules, developed by statistical training methods and designed to normalize the text of a specific language.

- **Machine Translation (MT)**

A software program that automatically translates a text from one language to another.

- **Normalization**

The act of modifying a monolingual text based on syntactic rules, generally with the aim of simplifying it or better adapting it to machine translation.

- **Normalization dictionary**

A monolingual glossary used in a rule-based translation model for automatically correcting the source text or modifying the target translation.

- **Out-of-domain corpus**

In statistical machine translation, a general bilingual text corpus used for training purposes. Serves as baseline data for common syntax and vocabulary.

- **Post-Editing Machine Translation (PEMT or PE)**

The process of correcting and improving a translation generated by a machine translation system.

- **Pre-editing**

The act of correcting and/or simplifying a source text, carried out automatically or manually, with the aim of better adapting it for machine translation.

- **Rule-Based Machine Translation (RBMT)**

Automatic translation using transformational grammar rules and multilingual dictionaries to build a translation engine.

- **Statistical Machine Translation (SMT)**

An automatic translation system that uses algorithms and large bilingual corpora to build a translation engine.

- **Testing corpus**

Part of the training corpus used to determine the quality of the translation output by comparing the reference translation with the machine translation of the same sentences.

- **Training**

The process of developing a statistical machine translation engine based on bilingual corpora.

- **Training corpus**

A bilingual corpus used for developing a statistical machine translation engine.

- **Translation dictionary**

A bilingual or multilingual glossary used in a rule-based translation model that includes coding of morphological and semantic information.

- **Translation engine**

See Translation model

- **Translation model**

Software based on a set of bilingual statistical or syntactic rules that translates a text from one language to another.

08

Lexcelera
We make global easy



09

Lexcelera
We make global easy

Find the solution that's right for you

Are you currently considering whether your company needs a machine translation system?

Do you already have a machine translation system in place but want to get more out of it?

Are you unsure of which machine translation system would best fit your needs?

Are you trying to decide between an internal machine translation system or Software as a Service?

Lexcelera's consulting services will help you evaluate the various machine translation solutions, compare their return on investment (ROI) and determine the best approach to meet your needs.

To learn more:

Call us: **+33 (0)1 55 28 88 00**

Email us: **info@lexcelera.com**

Visit our website: **www.lexcelera.com**

Follow us!

 **@Lexcelera**

 **www.facebook.com/Lexcelera**

 **www.linkedin.com/company/Lexcelera**



Fabrice Chabot

Computational Linguist

Fabrice, who joined Lexcelera in 2011, manages machine translation and linguistic engineering.

He began his career in the localization industry as a terminologist in 1999. He then worked in various related fields, including terminology, computer-assisted translation (CAT) management and linguistic engineering, particularly machine translation and computer-assisted review. Fabrice has a Master's degree in Linguistics and the Language Industry.

