Architectures of machine translation systems

Different strategies have been adopted by different researchers at different times in the history of machine translation. The choice of strategy reflects one side of the depth and linguistic diversity but also the grandeur of ambition on the other side. There are generally two types of architecture for machine translation, which are:

1. Linguistic Architecture

In the linguistic architecture there are three basic approaches being used for developing MT systems that differ in their complexity and sophistication. These approaches are:

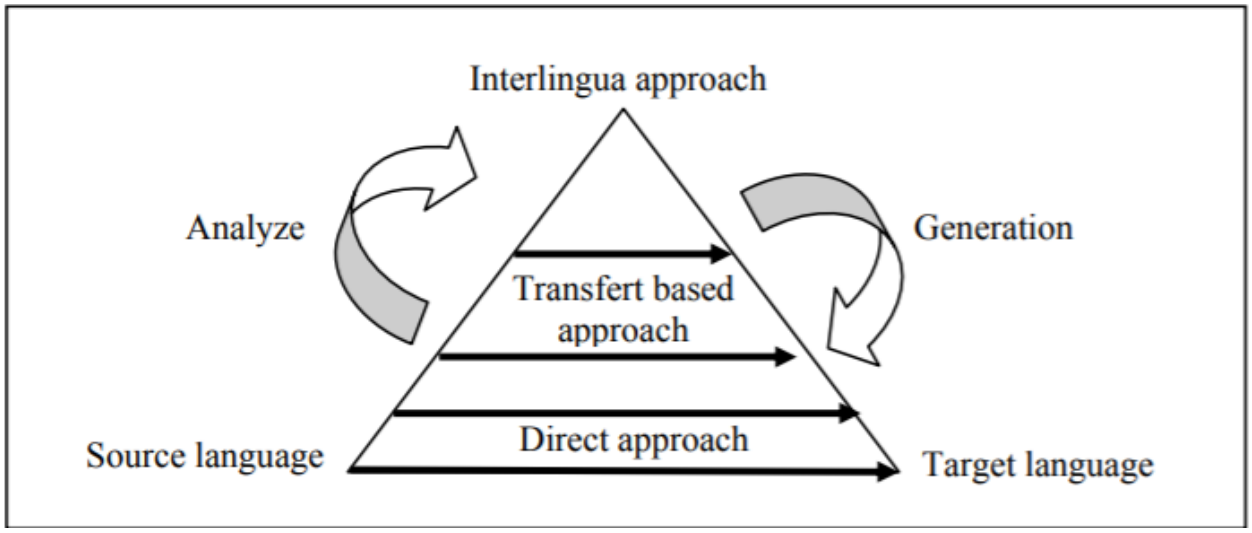


Figure 1. The Vauquois triangle [1]

- Direct approach

In direct translation, translation is direct from the source text to the target text. The vocabularies of SL texts are analyzed as needed for the resolution of SL ambiguities, for the correct identification of TL expressions as well as for the specification of word order in TL. This approach involves taking a string of words from the source language, removing the morphological inflection from words to obtain the base forms, and looking them up in a bilingual dictionary between the source and the target languages. Components of this system are a large bilingual dictionary and a program for lexically and morphologically analyzing and generating texts.

- Transfer-based approach

In the Transfer approach, translation is completed through three stages: the first stage consists in converting SL texts into an intermediate representation, usually parse trees; the second stage converting these representations into equivalent ones in the target language; and the third one is the generation of the final target text.

In the transfer approach, the source text is analyzed into an abstract representation that still has many of the characteristics of the source, but not the target, language. This representation can range from purely syntactic to highly semantic. In the syntactic transfer, some type of tree manipulation into a target language tree converts the parse tree of the source input. This can be guided by associating feature structures with the tree. Whatever representation is used, transfer to the target language is done using rules that map the source language structures into their target language equivalents. Then in the generation stage, the mapped target structure is altered as required by the constraints of the target language and the final translation is produced.

- Interlingua approach

The Interlingua approach is the most suitable approach for multilingual systems. It has two stages: Analysis (from SL to the Interlingua) and Generation (from the Interlingua to the TL). In the analysis phase, a sentence in the source language is analyzed and then its semantic content is extracted and represented in the Interlingua form representation, where an Interlingua is an entirely new language that is independent of any source or target language and is designed to be used as an intermediary internal representation of the source text. The analysis phase is followed by the generation of the target sentences from the Interlingua representation. An analysis program for a specific SL can be used for more than one TL since it is SL-specific and not oriented to any particular TL. Furthermore, the generation program for a particular TL can be used again for translation from every SL to this particular TL since it is TL-specific and not designed for input from a particular SL. [1]

2. Computational Architecture

In the Computational architecture there are different approaches being used for developing MT systems that work in different ways. These five common most types of approaches are:

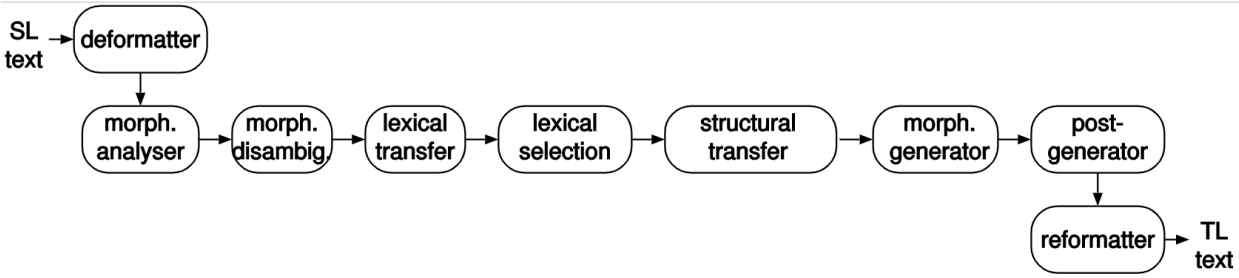
- Rule-Based Machine Translation (RBMT)

A rule-based system requires experts' knowledge about the source and the target language to develop syntactic, semantic and morphological rules to achieve the translation.

RBMT examples:

SYSTRAN is one of the oldest Machine Translation company. It translates from and to around 20 languages. SYSTRAN was used for the Apollo-Soyuz project (1973) and by the Europian Commission (1975). It was used by Google's language tools until 2007. See more at its Wikipedia article or the company's website. With the emerge of STM, SYSTRAN started using statistical models and recent publications show that they are experimenting with the neural approach as well. The OpenNMT toolkit is also a work of the company's researchers.

Apertium is open-source RBMT software released under the terms of GNU General Public License. It is available in 35 languages and it is still under development. It was originally designed for languages closely related to Spanish. The image below is an illustration of the Apertium's pipeline.



GramTrans is a cooperation of a company based in Denmark and a company based in Norway and it offers machine translation for Scandinavian languages.

Advantages:

- No bilingual text required
- Domain-independent
- Total control (a possible new rule for every situation)
- Reusability (existing rules of languages can be transferred when paired with new languages)

Disadvantages:

- Requires good dictionaries
- Manually set rules (requires expertise)
- The more the rules the harder to deal with the system. [2]

- Statistical Machine Translation (SMT)

This approach uses statistical models based on the analysis of bilingual text corpora. It was first introduced in 1955, but it gained interest only after 1988 when the IBM Watson Research Center started using it.

SMT examples:
Google Translate (between 2006 and 2016, when they announced to change to NMT)
Microsoft Translator (in 2016 changed to NMT)
Moses: Open source toolkit for statistical machine translation.

Advantages:

- Less manual work from linguistic experts
- One SMT suitable for more language pairs
- Less out-of-dictionary translation: with the right language model, the translation is more fluent

Disadvantages:

- Requires bilingual corpus
- Specific errors are hard to fix
- Less suitable for language pairs with big differences in word order. [2]

- Syntax Based Machine Translation (SBMT)

The goal of Syntax-Based Machine Translation techniques is to incorporate an explicit representation of syntax into the statistical machine translation systems. Syntax-based translation is based on the idea of translating syntactic units, rather than single words or strings of words (as in phrase-based MT), i.e. (partial) parse trees of sentences/utterances.

The idea of syntax-based translation is quite old in MT, though its statistical counterpart did not take off until the advent of strong stochastic parsers in the 1990s. Examples of this approach include DOP-based MT and, more recently, synchronous context-free grammars. One of the challenges of the syntax-based approach is translation speed. Improvements in translation quality have been noted with the use of syntax-based translation, but the speed of translation is significantly less than other approaches. [3]

- Neural Machine Translation (NMT)

The neural approach uses neural networks to achieve machine translation. Compared to the previous models, NMTs can be built with one network instead of a pipeline of separate tasks.

In 2014, sequence-to-sequence models were introduced opening new possibilities for neural networks in NLP. Before the seq2seq models, the neural networks needed a way to transform the sequence input into computer-ready numbers (one-hot encoding, embeddings). With seq2seq, the possibility of training a network with input and output sequences became possible.

NMT examples:
Google Translate (from 2016)
Microsoft Translate (from 2016)
Translation on Facebook
OpenNMT: An open-source neural machine translation system.

Advantages:

- End-to-end models (no pipeline of specific tasks)

Disadvantages:

- Requires bilingual corpus
- Rare word problem. [2]

- Hybrid Machine Translation (HMT)

Hybrid Machine Translation is a method of machine translation that is characterized by the use of multiple machine translation approaches within a single machine translation system. The motivation for developing hybrid machine translation systems stems from the failure of any single technique to achieve a satisfactory level of accuracy.

Although there are several forms of hybrid machine translation such as Multi-Engine, statistical rule generation and multi-pass, the most common forms are:

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. This is also known as statistical smoothing and automatic post editing. This is more of a "Band-Aid" approach to machine translation where there is an attempt to improve lower quality output from an RBMT engine rather than addressing the root cause of issues.

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. Many issues can be addressed at their root causes through rules that go beyond the capabilities on a statistical only approach. [4]

References

1. Cheragui. (2012). Theoretical Overview of Machine Translation. Proceeding ICWIT, 160-169.

2. Gergely D. Németh. Machine Translation: A Short Overview. URL: https://towardsdatascience.com/machine-translation-a-short-overview-91343ff39c9f.

3. Amr Ahmed, Greg Hanneman. Syntax-Based Statistical Machine Translation: A review // Computational Linguistics. – 30 p.

4. What is Hybrid Machine Translation? URL: https://omniscien.com/faq/what-is-hybrid-machine-translation/.