5.1 MACHINE TRANSLATION

Machine translation is the automatic translation of text from one natural language (the source) to another (the target). It was one of the first application areas envisioned for computers, but it is only in the past decade that the technology has seen widespread usage. Here is a sentence from this book:

–AI is one of the ***newest fields*** in science and engineering.‖

And here it is translated from English to Tamil by an online tool, Google Translate:

AI அறிவியல் மற்றும் பொறியியலில் புதிய துறைகளில் ஒன்றாகும்.

For those who don‘t read Tamil, here is the Tamil translated back to English. The words that came out different are in italics: –AI is one of the ***new disciplines*** in science and engineering.‖ The differences are of typical accuracy: of the two sentences, one has an error (change) that would not be made by a native speaker, yet the meaning is clearly conveyed.

5.1.1 Types of Translation

Historically, there have been three main applications of machine translation.

1. *Rough translation*, as provided by free online services, gives the –gist‖ of a foreign sentence or document, but contains errors.
2. *Pre-edited translation* is used by companies to publish their documentation and sales materials in multiple languages. The original source text is written in a constrained language that is easier to translate automatically, and the results are usually edited by a human to correct any errors.
3. *Restricted-source translation* works fully automatically, but only on highly stereotypical language, such as a weather report.

Translation is difficult because, in the fully general case, it requires in-depth understanding of the text. This is true even for very simple text of one word. Consider the word –Open‖ on the door of a store. It communicates the idea that the store is accepting customers at the moment. Now consider the same word –Open‖ on a large banner outside a newly constructed store. It means that the store is now in daily operation.

The problem is that different languages categorize the world differently. For example, to translate the English word –him,‖ into Tamil, a choice must be made between the humble and honorific form, a choice that depends on the social relationship between the speaker and the referent of –him.‖

5.1.2 Machine translation systems

Some systems attempt to analyze the source language text all the way into internal knowledge representation and then generate sentences in the target language from that representation. This is difficult because it involves three unsolved problems:

- creating a complete knowledge representation of everything;
- parsing into that representation; and
- generating sentences from that representation.

Other systems are based on a transfer model. They keep a database of translation rules, and whenever the rule matches, they translate directly, at lexical, syntactic, or semantic level.

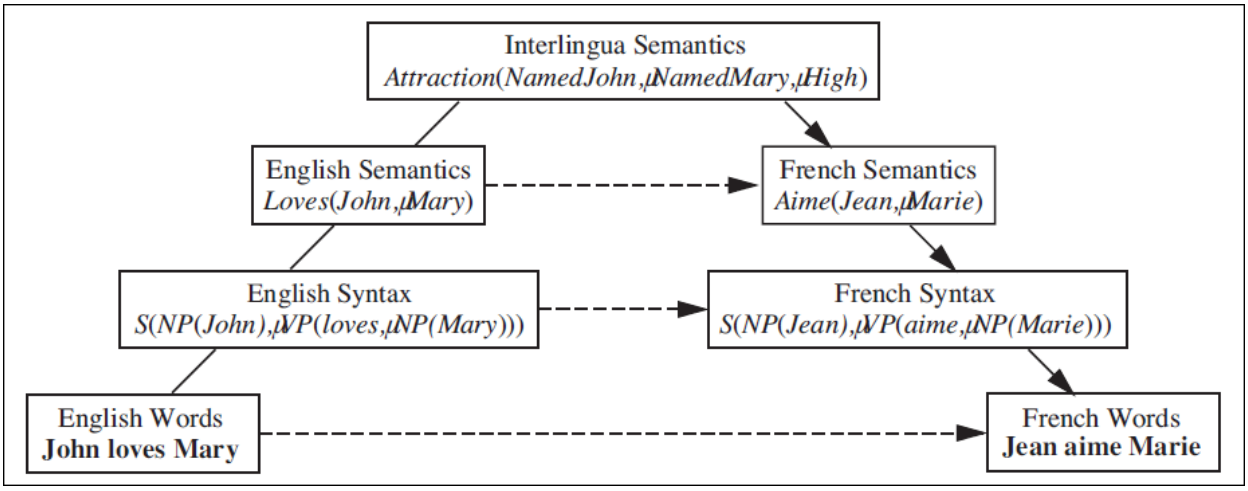

Figure 23.12 The Vauquois triangle: schematic diagram of choices for machine translation system

5.1.3 Statistical machine translation

Statistical machine translation needs data sample translations from which a translation model can be learned. To translate a sentence in, say, English (e) into French (f), we find the string of words f* that maximizes.

$$f^* = \operatorname*{argmax}_f P(f \mid e) = \operatorname{argmax} P(e \mid f)\, P(f)$$

Here the factor P(f) is the target **language model** for French; it says how probable a given sentence is in French. P(e|f) is the **translation model** based on Baye‘s rule; it says how probable an English sentence is as a translation for a given French sentence. Similarly, P(f | e) is a translation model from English to French. It is learned from a **bilingual corpus**—a collection of parallel texts, each an English/French pair. Now, if we had an infinitely large corpus, then translating a sentence would just be a lookup task. But of course our resources are finite, and most of the sentences we will be asked to translate will be novel.

Translation is a matter of three steps:

1. Break the English sentence into phrases.

2. For each phrase, choose a corresponding French phrase. We use the notation P(fi |ei) for the phrasal probability that fi is a translation of ei.

3. Choose a permutation of the phrases. For each fi, choose a **distortion** di, which isthe number of words that phrase fi has moved with respect to ei.

$$P(f, d \mid e) = \prod_i P(f_i \mid e_i)\, P(d_i)$$

5.1.4 Translation Procedure

1. **Find parallel texts**: First, gather a parallel bilingual corpus.
2. **Segment into sentences:** The unit of translation is a sentence, so we will have to break the corpus into sentences.
3. **Align sentences:** For each sentence in the English version, determine what sentence(s) it corresponds to in the French version. It is possible the order of two sentences need to be swapped, so align them.
4. **Extract distortions:** Once we have an alignment of phrases we can define distortion probabilities.
5. **Improve estimates with EM:** Compute the best alignments with the current values ofthese parameters in the E step, then update the estimates in the M step and iterate the process until convergence.