CMPT 825: Natural Language Processing **Spring 2006**

Lecture 1 — Jan 10, 2006

Lecturer: Anoop Sarkar *Scribe: Gholamreza Haffari*

In this part of the course, we consider Machine Translation (MT) task which is one of the important topics in Natural Language Processing (NLP). The first MT system was designed to translate Russian to English as a summer project! but the project still is in progress. In MT, the goal is to automatically translate a text from a *source* language $\mathcal{F}$ into a text in *target* language $\mathcal{E}$. The initials $\mathcal{F}$ and $\mathcal{E}$ can be memorized as we always translate from French to English.

The MT task can be done in different levels. In the lowest level, the translation can be carried on in the *word* level. For translating a sentence, we simply translate each of its words (in the source language) to their corresponding words (in the target language). However there are some problems: what does happen if a word in the first language has several corresponding words in the second language? What has to be done in the cases where the length of the translation is different from the length of the original sentence? What does happen if some parts of the original sentence have to change their positions in the translation?

The last issue points to this fact that not all of the sequence of words in the second language are valid sentences (syntactically and semantically). For a sentence to be valid in a language, it should respect the syntactic structure of the language, and moreover it should be a meaningful sentence. Therefore, in the next level of translation, *phrases* are translated with the hope that syntactic structure of the target language be respected:

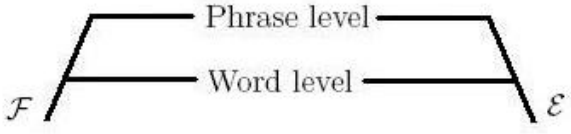

Figure 1.1. Translation in the word and phrase levels.

Still there might be some phrases which must be displaced to get a valid

syntactical sentence. This motivates us to consider the structure along the whole complete sentence, so the structure of the complete sentence is represented by a tree and the translation is done in the *tree* level. There is still a problem: a sentence in the source language can be translated to several syntactically valid sentences in the target language, but one (or a few) of them may exactly be equivalent to the original sentence in terms of the meaning. Therefore, it is better to the translation in the *meaning* level which is the ideal level of working with sentences:

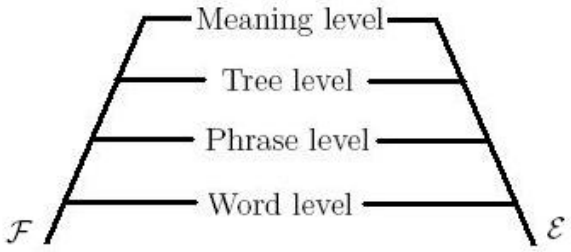

Figure 1.2. The pyramid showing different levels of translation. The mapping between the source and target languages is one-to-one at the top. As we go from the top to the bottom, the one-to-one property of the mapping becomes weaker.

The pyramid emphasizes this fact that in the top there exist only one meaning, and the mapping of the sentences is one-to-one. As we go to the bottom of the pyramid, the mapping becomes one-to-many in a higher degree.

The *ambiguity* problem is said to exist when there are different candidate sentences for the original sentence. Bar-Hillel is one of the first researchers who has dealt with ambiguity problem in MT. One of the techniques which can be used when ambiguity exists is *ranking*: to sort the candidates *somehow* and select the best one.

A *baseline* system is the most stupid system that we can build (often very fast) without using any complicated method. We always compare our designed system with a baseline system to justify the effort that we have devoted in building the complicated system.