Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers

Barbora Šmahlíková (xsmahl00)
Timotej Ponek (xponek00)

Statistical machine translation is a method that automatically maps sentences from one human language to another. In order to solve the problems of long-distance reordering between the source and target language text, the models can use some syntactic information. We can divide these techniques into two main categories: pre-ordering techniques and tree-based decoding techniques. Pre-ordering techniques parse and reorder the source text and then translate it into the target language. Tree-based decoding techniques, on the other hand, expect a tree or a forest as an input, and then they perform both the reordering and the translation simultaneously.

The advantage of pre-ordering is that the sentence is reordered before the translation begins, so the translation can then be performed with a conventional translation pipeline using a standard phrase-based decoder. However, this is not the case for the second approach, tree-based decoding, because it is necessary to have a decoder that can work with unordered sentences.

In this presentation, we are going to give an overview of two approaches for tree-to-string translation (using either synchronous context-free grammars or tree transducers) and compare them. Then, we introduce Travatar, an open-source toolkit for forest-to-string translation using tree transducers. We will briefly describe how this decoder works and then interpret the results achieved using this toolkit.