GeLexi project: Machine Translation based on Total Lexicalism

Gábor Alberti, Judit Kleiber, Anita Viszket

Linguistics Dept., University of Pécs

As last year, the bacic aim of our research team is to verify that computational linguistics is worth returning to the pure theoretical (generative) linguistic basis. Our crucial argument still relies on a double (parallel computational and linguistic) chance: to use a significantly greater number of huge patterns than earlier due to the immense increase in memory capacity; and to work out a formal grammar, showing the distribution of capacity advantageous in modern computer science: "minimal processing - maximal database". This latter chance has something to do with the sweeping lexicalist turn in generative linguistics.

This year we focus on the demonstration of a new (totally lexicalist) approach to machine translation which is based on the two-way application of our parser (accepting SL sentences, generating TL sentences). At the moment we can translate English sentences into Hungarian and vice versa, on a small corpus.

Total lexicalism means that every kind of information is stored in the lexicon, and the only syntactic "weapon" is *unification*, which means that there is no need for generating phrase structure trees. In this grammar lexical items are not (fully inflected) words but morphemes (stems and affixes), which is relevant on the syntactic and semantic "level" as well.

The input of our parser is a string (a sentence), and there are several outputs: the list of the relevant lexical items, the list of the established syntactic relations, a discourse-semantic representation (based on ReALIS, which is a developed version of Kamp's DRT), and a copredicative network, which is a level between syntax and semantics.

There is an isomorphic relation between the semantic representation of the Hungarian and the English versions of a sentence, which suggests the idea of applying our semantic parser in the area of machine-aided translation. It would take only a few hours to teach an intelligent, say, English-speaking person how to interpret a semantic representation containing English names of predicators; whilst it would take years to teach her even a basic level of, say, Hungarian.

To provide "real" machine translation, we can generate sentences by using the equivalents of the relevant lexical items of the source language, and presuming variables at specific template positions (for case and agreement marking affixes). Only grammatical sentences can be generated, because all the candidates go through the same parsing mechanism as they were SL sentences.

An important innovation of this approach is that there is no need for different mechanisms to translate from and into different languages. The only task is to elaborate grammars of more and more languages to achieve translation, because the frame we propose is universal.