Referent: Martin Forst

Titel: Grammatical Machine Translation II

Abstract

In the 1990s, statistical methods revolutionized natural language processing in general and machine translation (MT) in particular.

However, traditional string-based statistical machine translation (SMT) approaches crucially rely on transfer statistics acquired from bitexts, which are limited in size and rather specialized in genre, so that it is hard to repurpose those systems to new genres. Furthermore, word n-gram language models have proven to model morphologically poor languages much better than morphologically rich languages, and even the improvements of SMT systems due to better language models for morphologically poor languages have started to level off.

In my talk, I will present a new approach to machine translation that tries to overcome the limitations of traditional SMT as well as of its other main sources of inspiration, which are the work on Grammatical Machine Translation by Riezler and Maxwell (2006) and Context-Based Machine Translation (CBMT) by Carbonell et al. (2006). Like Riezler and Maxwell (2006), we use hand-crafted deep grammars for morphosyntactic analysis and generation. Like CBMT, we shift away from transfer statistics to overcome the limitations of available bitexts. Our approach relies on phrase pairs with dummy variable words from which f-structure transfer rules can be induced automatically. These phrase pairs can be created (semi-)automatically from bilingual dictionaries and (semi-)automatically or manually from bitexts. Compared to hand-crafted transfer rules they have the advantage of being easier to develop and of automatically providing rules that are in sync with the grammars.

The actual translation process is very similar to the one in Riezler and Maxwell (2006), with the difference of transfer statistics not having a prominent role. An input string is parsed into f-structures, the n-best f-structures (according to an available parse ranking model) are transferred by means of the induced rules, the target f-structures are ranked on the basis of a combination of the source analysis probability, monolingual bilexical dependency statistics and some very general transfer statistics (number of transfer rules applied, number of features/PREDs staying untranslated, etc.). The n-best target f-structures are then used as input to the generator, and finally the generated strings are ranked by means of a combination of source and transfer statistics as well as a language model in order to identify the best target string.

Since this is work in progress and the development of phrase pairs is not fully automatic, we do not have quantifiable results yet.

Nevertheless, preliminary results look encouraging.