A method to restrict the blow-up of hypotheses of a non-disambiguated shallow machine translation system

Jernej Vičič
University of Primorska, Slovenia
Petr Homola, Vladislav Kuboň
Charles University in Prague

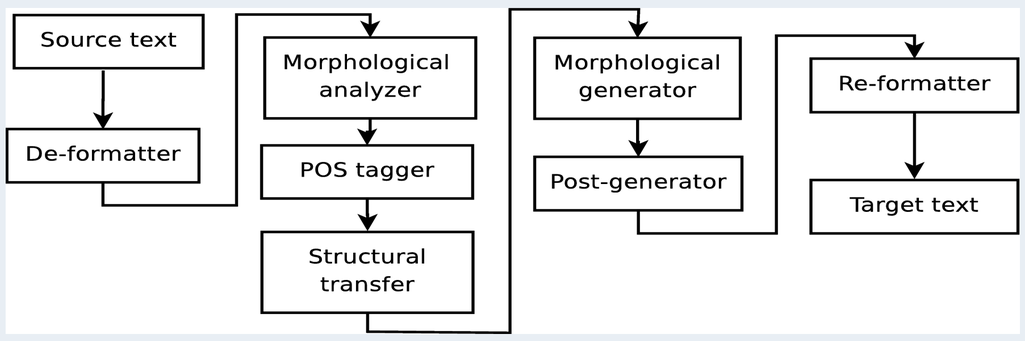
System Apertium

The open source shallow transfer MT system Apertium was originally designed for the Romance languages of Spain at the University of Alicante.

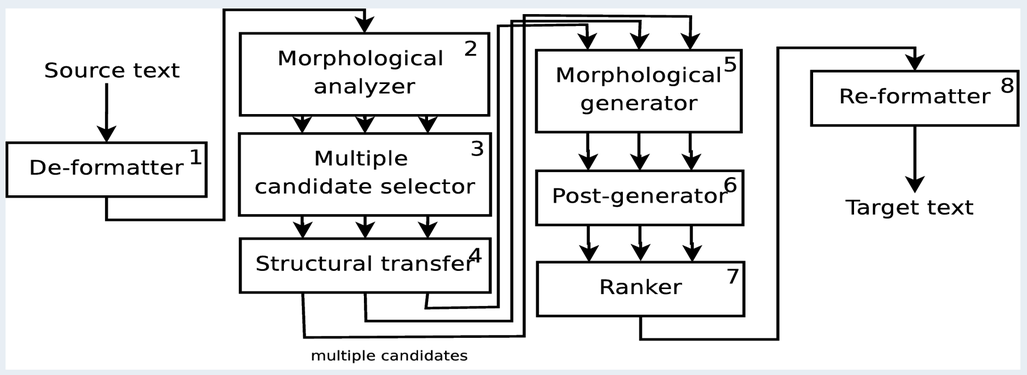
The dictionaries of Apertium contain single equivalents as well as multiword expressions. Transfer rules, which handle, for example, the rearrangement of clitic pronouns, have the form pattern-action and there are approx. 90 of them. The system is capable to process about 5,000 words per second.

The original architecture of the system described below contains a morphological disambiguator, tagger. As one of the initial processing stages it has a crucial role for the translation quality - the tagging errors may be carried over to the subsequent stages and they may negatively influence the translation.

The Schemes of Apertium



The original scheme



The modified scheme

1) De-formatter - encapsulates the formatting
2) Morphological analyzer - morphologically annotates the source text
3) Multiple candidates selector - produces an n-best set of possible translation candidates
4) Structural transfer - transfer from source to target language using shallow transfer rules
5) Morphological generator - translates target morphologically annotated text into target translation
6) Post-generator - post processing of the target translations
7) Ranker - selection of the best translation candidate from the list of produces translations
8) Re-formatter - introduces the formatting encapsulated by the first module

The Multiple Candidate selector

All posiible ambiguous candidates are constructed after the morphological analysis thus producing a lot of possible candidates. The multiple candidate selector discads the improbable candidates using the automatically generated set of rules. The number of translation candidates is reduced, but some source sentences can still yieald an exponential number of translation candidates.
The arbitrary number of translation candidates is obtained by using a stochastic ranker that selects the n-best set of candidates.

The Stochastic Ranker

The multiple candidate selector discads the improbable candidates still leaving an n-best set of candidates The task of selecting the best result is left till the end of the processing chain, to a stochastic ranker of generated target language sentences.
The stochastic post-processor aims at selecting one particular sentence that is best in the given context. A simple trigram-based language model (trained on word forms without any morphological annotation) sorts out ``wrong'' target sentences(these include grammatically ill-formed sentences as well as inappropriate lexical mapping).The current model has been trained on a corpus of approx. 9 million words which have been randomly chosen from the Serbian Wikipedia pages.

Automatic induction of local agreement rules

The automatic induction of the local agreement rules produces the same format as used by Structural transfer module. Trigrams and bigrams with morphological descriptions are extracted from source language corpus. Each bigram and trigram was checked for agreement among tags of different words, the tags and their positions were free.
The POS tags of the source bigram or trigram present the pattern part of the rule, the action part of the rule is constructed from all the morphosyntactic tags with agreement information. The rule candidates were grouped according to the pattern and action definitions.

Evaluation and results

System	All	All cand.	Used
original	57	57	57
all	44 million	44 million	34,526
rules	44 million	936,326	4,284
ranker	44 million	1,325,216	6,569
rules + ranker	44 million	437,123	4,041

1) System - Name of the tested system, each system is presented in this section.
2) All - the number of all candidates produced from tested sentences.
3) All cand. - the number of all translation candidates entering the last translation phase, the ranking phase.
4) Used - the number of unique candidates entering the last translation phase, the ranking phase.

System	E. D. char	E. D. word
original	0,848	0,778
all	0,892	0,826
rules	0,896	0,829
ranker	0,888	0,824
rules + ranker	0,896	0,829

1) System - Name of the tested system, each system is presented in this section.
2) All - the number of all candidates produced from tested sentences.
3) All cand. - the number of all translation candidates entering the last translation phase, the ranking phase.
4) Used - the number of unique candidates entering the last translation phase, the ranking phase.