Speeding up the Implementation Process of a Shallow Transfer Machine Translation System

Jernej Vičič
University of Primorska, Slovenia
Petr Homola
Charles University in Prague, Czech Republic

Introduction

This is a presentation of an attempt to automate all data creation processes of a rule-based shallow-transfer machine translation system. The presented methods were tested on two fully functional translation systems Slovenian-Serbian and Slovenian-Macedonian

Language pairs

Slovenian language is spoken by 2 million people, mostly in Slovenia.
Serbian language is spoken by 9 million people, mostly in Serbia, parts of Bosnia and Montenegro (Montenegrin).
Macedonian language is spoken by 4 million people, mostly in Macedonia.
All three languages belong to Southern-Slavic language group.

The architecture of a shallow transfer RBMT system

The slightly changed architecture adopted by:
Česílko, (Hajič et al., 2000)
Apertium, (Corbi-Bellot et al., 2005)
A new module has been proposed, the local agreement module.

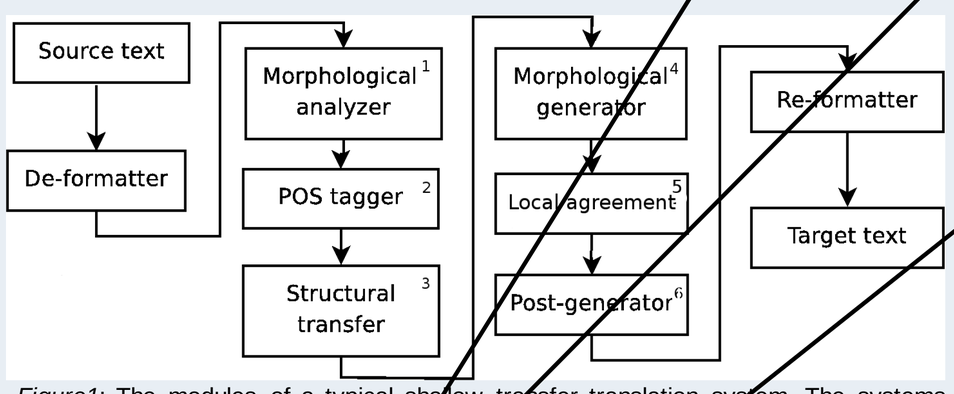

Figure1: The modules of a typical shallow transfer translation system. The systems \cite{corbi1,hajic2003} follow this design. An addition of the original architecture is the local agreement module tagged as number 6.

1) Morphological analyzer - morphologically annotates the source text
2) POS tagger - disambiguates morphological ambiguities
3) Structural transfer - transfer from source to target language using shallow transfer rules
4) Morphological generator - translates the morphologically annotated target text into target translation
5) Local agreement - enforces local agreement of adjacent lexical units
6) Post-generator - post processing of the target translations

Monolingual source and target dictionary creation

Morphological paradigm:
- typical lemma; the lemma the paradigm was constructed from,
- a stem; the longest common prefix of all words in the lemma,
- a set of all words split into stems, suffixes and Morpho-Syntactic Descriptors (MSDs)

All of the word forms of a lemma present in the corpus are grouped into a class represented by lemma. A paradigm is constructed from each class for each lemma. Two paradigms are joined together if the lemmata of both paradigms have the same POS tag and if the entries, pairs of suffix and MSD, of one paradigm present a complete subset of the compared paradigm.

POS tagger training

The POS tagger was trained using the Apertium toolkit tools. The training corpus was collected from the Wikipedia pages.

Bilingual Translation Dictionary Creation

An SMT word-to-word model was trained using GIZA++ on a parallel, sentence aligned corpus. The corpus was lemmatized and POS tagged. The dictionary was further extended.

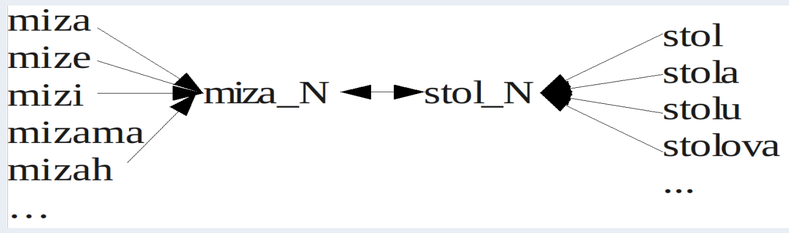

Figure2: The bilingual dictionary is based on lemmata.

Shallow Transfer Finite-State Rules Induction

The shallow transfer, finite-state type rules, were constructed using available software from the Apertium toolkit.

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

The evaluation comprised of three sets of tests:
1) following the LDC guidelines, producing separate results for fluency and adequacy
2) using the Word Recognition Rate (WRR) metrics
3) using the METEOR metrics

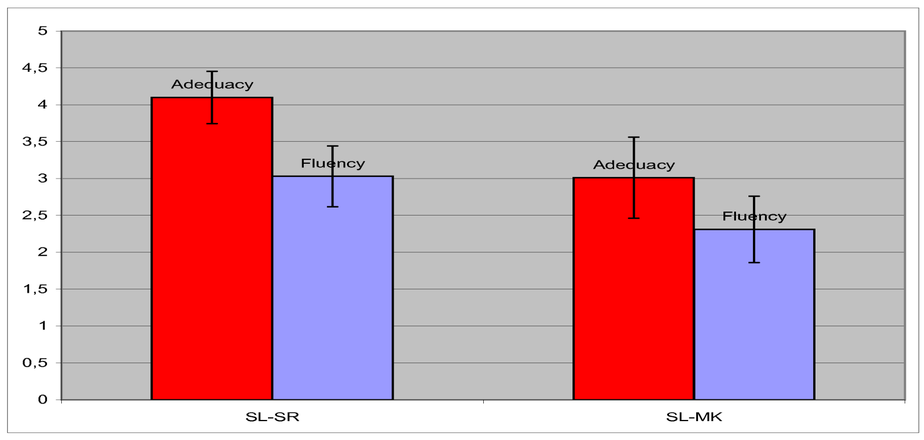

Figure 3: Evaluation results using (LDC, 2005) guidelines. Average values of four independent evaluations show high scores for adequacy and lower values for fluency.

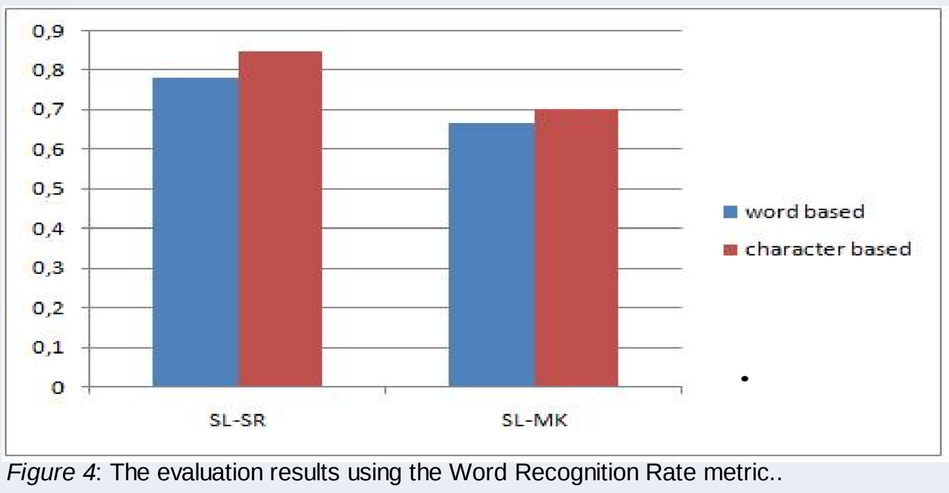

Figure 4: The evaluation results using the Word Recognition Rate metric..