

Examiner's Report

Machine Translation on Related Austronesian Languages

Septina Dian Larasati

In her very short master thesis the author presents a core of a machine-translation system that is to translate between closely related languages, from Indonesian to Malay. The thesis has the following parts: Language Pair Typology, Machine Translation Platform, Finite-State Morphology, Monolingual and Bilingual Dictionaries, Transfer and Additional Process, Evaluation and Future Work. The main text is accompanied by appendices comprising finite-state graphs, tables and a translation texts as results of the machine translation proposed.

In the chapter on Language Pair Typology the author presents an overview of the main properties of both languages under investigation, i.e. Indonesian and Malay. Even a person without any knowledge of these languages can understand their basic features. This part is well written; one can very quickly penetrate into the main characteristics of the languages in question. It is evident that they considerably differ from European languages, although some features are, at least conceptually, similar. Then comes the chapter on Machine Translation Platform. Therein, the architecture of the system is depicted with a simple graph containing the description of all the modules of the system. On this level it seems very simple. Moreover, resources used are also described. Then Finite-State Morphology is described as applied to both languages, and existing tools are presented: especially Lttoolbox and various problems which this tool cannot cope with and which need special treatment. Also the tagset for both languages and finite-state graphs are given.

In Chapter 5 Monolingual and Bilingual Dictionaries working dictionaries for both languages are described with entries specified in XML format. In Chapter 6 Transfer and Additional Process are described and the next chapter contains a very primitive evaluation.

The whole approach is a very primitive one. Of course, in case no MT system for these languages is available an enthusiast must start somehow „from scratch“ but even in this case the author's objective should be more ambitious. Apart from basic facts about the language pair, no interesting MT idea was formulated, just a very primitive and basic approach to MT between two very close languages. One needs more data to build up working monolingual and bilingual dictionaries, describe at least an embryonic syntactic analyzer and synthesizer concentrating on the differences between the closely related languages, no matter how syntactically simple they are. On the whole, a survey of various kinds of differences should have been presented. If close languages are to be machine-translated, the MT system must elucidate where the differences are: so the master thesis should contain a chapter comprising a survey of important differences. Some differences were described in the master thesis but a more extensive study should be included and translation of problematic parts where the languages differ should be depicted.

If the author wants to develop her system as it is described in her thesis a lot of work is still ahead: more extensive dictionaries (monolingual and bilingual), elaborated morphology and syntax (semantics can come at a later stage). One needs monolingual corpora of both languages and a parallel corpus, not just a comparable one. Without corpora, no serious Indonesian-Malay MT system can be developed.


I am aware of the fact that the objective of writing a master thesis does not consist in a complex MT system of any kindt, but the student should show a basic ability of writing a (possibly only compiled) scientific text on an appropriate level. However, only very basic features of the languages in question and a solution to their MT translation were presented. I would have expected more.

The master thesis is written in a very bad English with trivial errors, therefore some parts are difficult to understand (even illegible). It seems that – besides the author – I was the first to read it. At least the supervisor should have read the thesis in a preparatory stage (and during proofreading) at least and, of course, the text should have also been perused by a native speaker of English.

I ask for one clarification: on page 28, 4th line, there is: n="verb" but on the previous page you say that a proclitic is to be changed to an independent pronoun. I do not grasp the association.

Overall evaluation: I recommend the thesis for the defence.

In Prague August 22, 2010


doc. RNDr. Vladimír Petkevič, CSc.
Institute of Theoretical and Computational Linguistics
Faculty of Arts
Charles University