Diploma Thesis of Septina Dian Larasati

Machine Translation on Related Austronesian Languages

Supervisor's Review

Although the title of the thesis might be understood in a more general way, the main task of this thesis was building a machine translation system for a particular pair of related Austronesian languages, namely for Indonesian and Malaysian. Given the fact that the amount of available resources for these languages is extremely limited, the author was supposed to use an existing MT system (Česílko, Apertium) and to concentrate on gathering all necessary data (morphology, mono and bilingual dictionaries, transfer rules etc.) and adapting the system if any properties of the source or target languages will not fit into the existing mechanism.

The thesis consists of eight chapters, the first of which is an introduction in which the author briefly summarizes previous experiments with almost direct MT of related languages carried mostly on European languages. She also gives an outline of the thesis.

The second chapter introduces both languages, Indonesian and Malaysian, and gives an overview of their syntactic and morphological properties which play an important role in a simple direct MT system for related languages.

The third chapter introduces an MT platform used for the experiment. Out of the two systems for shallow MT which have a very similar architecture (Česílko, Apertium), the author has decided to use Apertium mainly due to its open source nature.

Given the fact that shallow MT systems usually rely on syntactic similarity of related languages, the quality of the resulting translations depends mainly on the quality of morphological tools (analyzer, synthesizer). This is reflected in the fourth chapter of the thesis which describes problems and solutions related to the implementation of Indonesian and Malaysian morphology in the Apertium framework. The author has discovered that a standard Ltoolbox tool used in Apertium is not sufficient for handling certain properties of Austronesian langugaes as, e.g., reduplication. She has therefore decided to build the Indonesian morphology on the basis of an existing morphological analyzer exploiting the XEROX technology. However, neither this tool turned out to be fully adequate for the given task, therefore the author had to modify it. In cooperation with the Apertium community she also managed to incorporate the modified morphological analyzer for Indonesian into the Apertium framework.

The fifth chapter is devoted to dictionaries. Due to the lack of available resources the author could not use wide coverage mono or bilingual dictionaries. In this chapter she explains her approach concentrating on providing small but reliable dictionaries which are easily extendable in the future. The process involves certain amount of manual work which is necessary for creating core dictionaries which might be expandable in the future by the application of more automatic methods. The author also compares the initial morphological analyzer with the modified one demonstrating that a substantial simplification and reduction of the number of lemmas brings only a relatively small loss of coverage on 1000 randomly chosen sentences from Indonesian Wikipedia. This is a good sign for future extensions of the dictionary which may substantially increase the coverage by the addition of a relatively small number of new lemmas.

The sixth chapter briefly deals with transfer phase which is very simple due to the close relationship between the source and target language. It also presents a temporary solution to the problem of handling compound words in Apertium where the issue is so far handled in an unsatisfactory way. The language pair used in this thesis thus provides a valuable input for the whole Apertium community pointing out issues which need a special attention.

The sevenths chapter presents a simple evaluation using Word-Error-Rate metric measured against post-edited output provided by two independent post-editors. Due to the close relatedness of both languages the author uses untranslated text as a baseline and achieves a substantial improvement in the translation exploiting a hand-picked dictionary. On the other hand, if the dictionary is not manually refined, the results fall below the baseline. This clearly indicates that the quality of the dictionary is crucial for the quality of the translation. The tests have also revealed the necessity to pay special attention to idiomatic expressions in the future which are pretty frequent in certain types of texts (e.g. newspaper articles etc.).

The thesis ends with an outline of the future work, three appendices and bibliography.

From the point of view of the thesis supervisor I can confirm that the author did exactly what she was supposed to do – to build a pivot implementation of an MT system for closely related languages using existing technology. She completed the main task – to identify problems arising from the fact that the rather simplistic MT technology of Apertium had been previously tested only on typologically similar European languages. The author not only identified certain problems at a morphological level, she also actively cooperated with the Apertium community in suggesting and testing a more adequate solution. Also the linguistic part of the thesis is very valuable and it provides a good starting point for future research on MT between the two related Austronesian languages.

The overall impression from the thesis is negatively influenced by two factors. The first one is the lack of available data and resources which did not allow neither building an MT prototype with a reasonable dictionary coverage nor wider scale evaluation of the results. Although the author spent a long time searching for at least some resources, they turned out to be nonexistent not only on the web, but also among the scientific community dealing with both languages under investigation. This fact was confirmed during the author's participation at a Malindo conference in Jakarta where she had a chance to present a paper describing some parts of her thesis which has also been published in the conference proceedings.

The second negative factor is the quality of the English. The author did not have a choice of writing the thesis in her mother tongue, but nevertheless, if she would ask a native speaker for proofreading the thesis, it would very much improve the overall impression. Especially the second part of the thesis which was apparently written under time pressure is in certain parts very difficult to understand. The thesis also appears to be very short, what is partially due to the compact layout and very small font used throughout the thesis.

Given the facts mentioned in the review, I can recommend this thesis for the defence.

In Prague, August 27th, 2010

[redacted]

RNDr.Vladislav Kuboň, Ph.D.
ÚFAL MFF UK