Diploma Thesis of Amir Kamran

Hybrid Machine Translation Approaches for Low-Resource Languages

Reviewer's Review

Although the title of the thesis might be understood in a more general way, the thesis concentrates on a single language pair, namely English-to-Urdu. This of course means that the task described in the thesis is a bit easier than the title suggests – it is actually a translation from a resource-rich to a resource-poor language. Although on the one hand the bilingual data are very sparse, the source language has ample resources which may be (and they are) exploited in the experiments.

The thesis consists of six chapters, the first of which is an introduction in which the author briefly describes the main objectives of the thesis. Actually, the restriction of the broad thesis topic to a particular language pair was already envisaged in the official thesis specification. The text of the thesis is accompanied by a CD containing all relevant data and programs.

The second chapter contains a brief overview of topics important for the thesis, starting with a very brief history of Machine Translation and finishing with the description of evaluation techniques. The content of this chapter is slightly unbalanced, the author concentrates a bit too much on technical details of automatic metrics BLEU and NIST, while more than 60 years of MT history takes only a single page. This might make it more difficult for non-expert readers to understand the subsequent text of the thesis in a proper perspective.

The third chapter, on the other hand, contains a very nice description of data and tools exploited in the experiments. The description has just the right level of detail. The chapter also describes baseline experiments, in fact a plain translation model using MOSES and unmodified parallel corpora.

Chapter four starts a bit chaotically by page 27, followed by a page 28, followed by the actual first page (26) of the chapter. It first describes one of the two phenomena which the author wants to exploit in order to improve the baseline results. The discussion of the role of various markers in Urdu is well written and it makes it possible even for people who do not know the language to understand the problem and the proposed solution. On the basis of the linguistic considerations the author made a logical decision to modify the source English text prior to the translation by adding artificial markers on relevant places of source language sentences. As a side effect of this operation, the sentence lengths of the source and target languages are more equal than before (the original English sentences tend to be substantially shorter than their Urdu equivalents). This effect helps to improve the alignment, but, unfortunately, the hypothesis that it will also help to increase the translation quality, is not confirmed.

The second attempt to achieve an improvement of MT quality is presented in Chapter 5. The author the tries to reorder source language sentences so that they would adopt the word order of the target language. Two methods are presented – manual and automatic one. Both work with a set of relatively simple reordering rules and both achieve an improvement. Actually, even the second method is not purely automatic because it exploits a set of manually aligned sentences.

The final chapter then contains a discussion and some ideas for future work.

The overall impression from the thesis is good, the thesis is written in understandable, although not perfect English. Author always tries to explain all important factors of his experiments, provides a number of tables and charts summarizing and illustrating them. The tables and charts are accompanied by a discussion of the results, but the reader still may get an impression that the thesis is a bit short, especially when he reads the final pages of chapter 5 consisting almost entirely from tab les and charts.

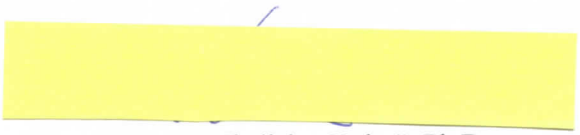
As for the content of the thesis, the experiments are interesting and well documented, the hypotheses being tested are linguistically based and natural, but something is still missing. Both hypotheses are tested on various sets of data, but only individually. Even though the inserted markers didn't bring almost any improvement, they might help in combination with the reordering, this is an experiment which is definitely missing. The second important aspect neglected in the thesis concerns the evaluation. Although the author mentions on several places the imperfection of BLEU and NIST metrics (especially when they are used with a single reference translation only), these metrics are the only ones used for the evaluation. The author mentions in the conclusions, that the random test of

translated sentences showed some improvement from the point of view of human understanding, this should have been investigated in the thesis.

It is also necessary to mention a very long list of references (although their style varies between using full first names of the authors and their abbreviations, some references are also incoplete, as, e.g., "FJ Och. Statistical machine translation: from single-word models to alignment templates. Germany, 2002". Apart from this, the thesis fulfills all formal criteria of a master thesis.

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

In Prague, August 30, 2011



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