Master's Thesis Review

Author: Bushra Jawaid
Title: Statistical Machine Translation between Languages with Significant Word Order Differences
Opponent: RNDr. Ondřej Bojar, Ph.D.

The thesis presented by Bushra Jawaid aims at improving English-to-Urdu translation by specifically handling word reordering and attempting at better target-side morphological coherence. The overall structure of the text is well balanced. The introductory survey of MT techniques is clearly written and detailed enough. Not speaking Urdu, I was especially delighted by the introduction to Urdu alphabet and language and by the careful transliteration and explanation of all Urdu examples.

Starting from scratch, Bushra was able to collect a reasonably sized parallel and monolingual corpus, documenting all the problems of texts from individual data sources. For some data sources, Bushra even invested the labour to manually align the texts at the level of sentences and correct poor translations. This new data source alone deserves a public release. Similarly, the subsequent evaluation of available taggers is very valuable for all Urdu NLP researchers.

The core of Bushra’s thesis are techniques for improving target-side word order. The methods are carefully described, evaluated by an automatic score and also manually inspected. Bushra was able to achieve a significant improvement in the automatic score and even outperform Google Translate on most of her specific test sets. The unsuccessful experiments with factored translation are also well documented and some explanations proposed.

I have the following two minor questions to the content:

1. Could you provide some numbers illustrating how often each transformation rule in Section 3.2.1 is applied and how often the system resorts to the default rule for a particular non-terminal?
2. Regarding the mixed results of Mixed LM: could you provide the model weights learned in MERT for the various setups? My hypothesis is that e.g. the language model weight got consistently lower leading to a more literal translation.

The thesis is well written with only occasional errors in grammar (e.g. “itself” instead of “themselves”), typesetting (e.g. column labels of Tables 4.2 and 4.3 or the misleading “Urdu Tree” caption in Figure 3.1) or references (e.g. “Section 0” on page 23 paragraph 1 or the label “[Conference]” for journal papers in the list of references).

Conclusion
The reported thesis proves the author's ability to solve independently and creatively assigned tasks in the area of NLP. The thesis is written in good English, all experiments are sufficiently documented and the results are discussed. In my opinion, it complies with the requirements for Master Thesis at MFF. I recommend to accept the thesis for the defense.

Prague, August 26, 2010.

RNDr. Ondřej Bojar, Ph.D.
Charles University in Prague, ÚFAL