Review of the Master Thesis

Author: Bushra Jawaid
Title: Statistical Machine Translation between Languages with Significant Word Order Difference
Supervisor: RNDr. Daniel Zeman, Ph.D.

The goal of the thesis is tuning of a statistical machine translation system for a language pair with significant word order differences, namely English-to-Urdu translation. The work consists of training data collecting and cleaning, training the MT system, analyzing its errors and designing and evaluating techniques to improve translation quality for the given language pair. The thesis consists of five chapters. It includes a list of references and lists of tables and figures, an appendix with description of transformation rules, another appendix with sample translations and an index of terms. After Introduction, Chapter 2 describes the used corpora and the process of their preparation, Chapter 3 provides a theoretical discussion of the improvement techniques and Chapter 4 gives their experimental evaluation and discussion. Chapter 5 summarizes the contribution of the work. The attached CD contains a selection of the collected corpora (only those parts whose licensing terms permitted redistribution), the monolingual data for Urdu language model, various tools needed for the experiments (such as the web crawler, text extractor, normalization, tokenizer, Urdu stemmer and tagger), as well as configuration files for the MT system to be able to repeat the experiments.

The author's work can be divided into four areas: 1) acquiring suitable training data; 2) approaches to data cleaning and normalization; 3) experiments with various reordering models, including the transformation-based preprocessing of the source English; 4) experiments with factored translation models.

Acquisition and manual correction of training/test corpora (both parallel and Urdu monolingual) turned out to be a crucial first step because next-to-nothing had been available before. The author used her native knowledge of Urdu to manually correct and sentence-align more than 8000 of the Emilie corpus (which, despite being known for its many flaws, is the only "parallel" corpus distributed as such). The author then found other texts on the web, namely an Urdu translation of the Penn Treebank texts, plus Quran and Bible translations. These corpora required less manual editing but the author had to implement a few tools for their downloading, cleaning and normalization (some Urdu texts use European digits and punctuation, others use Arabic ones; some Urdu texts use diacritical marks for short vowels while most texts omit them).

The author uses the Moses MT system for the research. After initial experiments with various normalization schemes, the main focus of the work goes to various techniques of assessing the differences in preferred word order in the two languages. The author evaluates two existing trainable reordering models available with Moses ("bidirectional" and "distance-based"). The most interesting part however is the author's own preprocessing approach (inspired by a similarly operating block in a rule-based MT system, co-authored by the applicant with two other students during their Be. project). It first parses the English input using the Stanford parser, then Urdu-inspired transformation rules are applied to the parse tree, the resulting reordered sentence is output as plain text and used as input for Moses. Using this approach, the author has been able to significantly improve over the baseline results and insignificantly improve over the distance-based reordering results (using BLEU score as the metric), proving once again that a well-deployed linguistic knowledge can improve performance of a statistical
model. Finally, the author also describes preliminary experiments with factored translation models, aimed at the relatively rich morphology of Urdu (in comparison to English). Although no improvement of BLEU score has been achieved in this part of the research, the evaluation and discussion in the thesis gives a ground for future work. Problems can be attributed partially to the computational complexity of the factored models, partially to unavailability of high-performance tools for Urdu morphological analysis.

The work shows that the author has gained a good insight into the problem of statistical phrase-based translation; she provides the reader with an ample theoretical background and with references to relevant literature. She extensively works with existing tools (such as Moses, SRILM, Stanford Parser, CRULP Urdu tools). The text of the thesis is well structured, written in passable English (although the last two chapters would benefit from some more proofreading).

I have detected only a few minor flaws, e.g.:

- In Table 4.9, unnonnalized reference Urdu segment somehow made it into the table instead of its normalized counterpart (diacriticized “۝پ” / “phir” instead of normalized “پ” / “phr”).
- Occasionally, Romanization of Urdu examples is missing.
- On page 59, tense+number suffix is called “case ending”.
- An appendix with a description of the Urdu POS tagset would be helpful.

**Conclusion**

The thesis proves the author’s ability to independently and creatively solve problems and conduct research in the area of NLP. The work is well structured and written, all experiments are appropriately documented and the results discussed. The most important achievements are: 1. Enormous effort invested into collecting and cleaning the data; the result is an invaluable resource for the research community; 2. Detailed experimentation with reordering, successful porting of the transformation rules from a rule-based MT system to Moses and the resulting improvement. On my opinion, the thesis complies with the requirements for Master Thesis at MFF UK. I recommend the thesis to be accepted.

Jenštejn, August 24, 2010

RNDr. Daniel Zeman, Ph.D.
Institute of Formal and Applied Linguistics
Charles University in Prague