Master's Thesis Review

Author: Naila Ata
Title: Analyzing Errors and Chances of Improving English to Urdu Phrase-Based Translation
Supervisor: RNDr. Ondřej Bojar, Ph.D.

The thesis submitted by Naila Ata aims at analyzing errors and chances of improving English-to-Urdu machine translation.

The first three chapters provide a brief survey of MT, statistical MT and factored phrase-based MT, respectively. Chapter 4 is devoted to error analysis of a baseline system, concluding that morphological choice is the most frequent flaw. Chapter 5 lists data sources used to build a parallel English-Urdu corpus and chapter 6 continues with text normalization. Chapter 6 also briefly lists software used and some steps of the "standard Moses training pipeline". Chapter 7 introduces Naila's ideas for tackling the observed types of errors and Chapter 8 empirically evaluates some of the configurations. The conclusion is given in Chapter 9.

The highlights of the scientific content of the thesis are:

- manual analysis of errors in English-to-Urdu translation,
- the idea to reuse multiple translations of Quran to increase the corpus size,
- the inclusion of Urduseek dictionary in the training data,
- the idea to reduce Urdu tagset to increase tagger accuracy (Section "Impact of Generalized tagging for Urdu" within Section 8.2.).

The weak points are:

- incomparable experiments in Section 8.1 due to different test set sizes,
- false or misleading claims, e.g.:
  - "BLEU score confirmed that using a large ... increased the quality" on pg. 44 while the score is actually lower,
  - "Second main source is the missing connection words" on pg. 28 while it is content words that are marked as missing more often in Table 1, pg. 29,
  - "BLEU score increased by 0.571" on pg. 45; this probably reports the difference on the development set (instead of the test set) and moreover includes a typo or miscalculation, the actual difference is 0.0569,
- too scarce comments and discussion of the results (some of the results reported in Section 8 may actually need a re-evaluation),
- lack of technical details on reordering rules mentioned in Chapter 7 and evaluated at the end of Section 8.2,
- wrong formulas (e.g. Equation 2.2, pg. 12),
- confusing references, e.g. "The scores are nearly identical to the baseline in table 1" on pg. 44,
- missing references (e.g. corpora sources on pg. 30-31, Moses and MERT on pg. 33),
- somewhat messy structure (see e.g. the contents of Chapter 7 or the various unnumbered parts of Section 8.2),
- unclear statements (e.g. "...considering a French sentence F as an encoded English sentence implies that the probability ... can be expressed using Bayes's rule" on pg. 12).

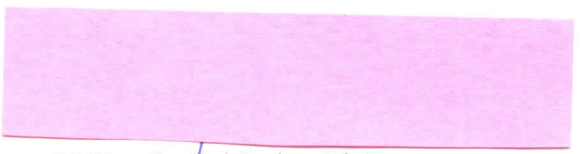
Additionally, the text is hard to read for readers not familiar with Urdu due to the lack of English glosses for Urdu examples.

From the formal point of view, the presented document unfortunately contains many mistakes, starting with the abstract consisting of the original assignment instead of a summary of Naila's work. Despite our efforts, many severe errors remained in the presented document, including poor sentence formation, English grammar, typos (e.g. imp9rovements, Kohen or many typos in the bibliography: Entropry, Koehnn, Statisitcal) or typesetting issues (e.g. Figure 2.1, pg. 12; the figures on pg. 24 and 25).

While I would like to emphasize that Naila has made a huge leap in her knowledge of MT and she has invested a lot effort in manual analysis of MT errors, data preparation as well as MT experiments, the current version of the text is not an adequate evidence of that.

To conclude, the presented thesis is unfortunately below the standard of M.Sc theses at Charles University. I recommend the thesis to be rejected and re-submitted after a significant improvement in the experimental and discussion section and thorough editing of the whole text.

Prague, August 27, 2010.



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