

# Review of Thesis Proposal Presented by Bushra Jawaid

Reviewed by Daniel Zeman, 25 January 2013.

The research proposal focuses on two main issues encountered during machine translation between certain pairs of languages (here represented by English to Czech and English to Urdu): long-distance reordering of words, and rich morphology on the target side.

The author intends to apply and extend the two-step translation scenario (Fraser 2009, 2012). Besides work on machine translation, this also involves preparing tools for source-side word order transformations (builds upon the author's previous work during her undergraduate study) and for target-side morphology (the author claims to have almost finished a tool for Urdu; tools for Czech are already available).

The proposed research is indeed interesting and novel. The timeline seems reasonable, although I am afraid that building the classifier (planned for May – July 2013) may turn out to be a difficult part. I did not attempt to evaluate the language aspect of the proposal (there are occasional typos but overall it's acceptable). Nevertheless, for the final text of the thesis, I would advocate for a better way of marking the various two-step setups. Codes like “tF-LOF+MOT=tLOF+MOT-F” are very difficult to decipher. As a minimum, the operators should be described in depth, including examples. But perhaps changing some of the operators would help too. If “tF-LOF” means “we *translate* source form to intermediate lemma-or-form then I would prefer something visually more prominent than the hyphen, e.g. an arrow surrounded by spaces: “tF → LOF”.

## Questions

In Section 3.2.1, you say that “we will replace phrase-based Moses on the first step with either Joshua or Moses-chart.” I did not see this in your timeline. Are you using a hierarchical decoder in your experiments already? Do you have any comparison of results between baseline Moses and a hierarchical decoder?

In Section 3.2.2, I did not understand the way you intend to “produce multiple reorderings and later feed them into Moses.” What component will be responsible for producing multiple reorderings? You mention “the transformation system”. Do you mean the preprocessor taken from your master thesis that reorders the input English sentence, both in the training and in the test data? Just one section before, in 3.2.1, you say that this transformation system will produce 1-best reordered output, which seems confusing.

If it is true that the preprocessing transformations will produce two or more reorderings per input sentence, how will you tell Moses that they represent the same sentence? And what about the translation phase? Will you make Moses to somehow combine or compare the  $N$  reorderings and select the best one for translation, or will you translate each of the reordered inputs separately and postpone the 1-best selection for a later stage?

In Figure 3.1, if I understand it correctly, your different MOT scenarios always take a fixed subset of feature positions in the Czech morphological tag. I would have expected that for different parts of speech, different subsets of the morphosyntactic features are useful. Have you considered that?

In Section 3.4.1, you propose to train a maximum-entropy-based classifier for modeling target-side inflection, and to use it for the second step in the two-step scenario, instead of Moses. Why do you need a classifier when your intermediate data already contain both the lemma and the tag? Ain't this a job for a deterministic morphological generator? Or is this intended to reconstruct the information missing from the reduced tags? What sort of ambiguity do you assume?

In Table 3.1, you state that Urdu has two cases of its own (ergative, oblique), and in addition to that all the seven cases that are also present in Czech (you show these cases in the column labeled

“Both”). Am I correct in assuming that this is an error? I thought that Urdu had two cases for nouns (direct and oblique) and possibly lots of other cases for pronouns (because of joint pronoun-postposition tokens, such as *unke*) but this would probably be a different set than the Czech one.

In Section 3.5.5, in discussion of Table 3.3, you say that “we see no improvement over the ... baseline (BLEU of 11.84)”. More than half of the scores in the table are higher than 11.84, some of them significantly. Can you explain that please?