Review of the Master Thesis

Author: Naila Ata
Title: Analyzing Errors and Chances to improve English to Urdu Phrase-based Translation
Opponent: RNDr. Daniel Zeman, Ph.D.

The goal of the thesis is to set up an MT system for English-to-Urdu data machine translation, to collect small parallel corpus for that language pair, to thoroughly manually analyze the output of the MT system, propose techniques to tackle the identified issues and evaluate them by both automatic metrics and human judgments. The thesis comprises 49 pages, out of which 9 pages contain formal stuff (headers and bibliography) and the rest is roughly divided half-and-half between a theoretical survey and the description of the author's own work. There are 9 chapters (plus Bibliography) with the following contents: 1. introduction, 2. survey of statistical MT, 3. survey of factored MT, 4. error analysis, 5. data, 6. experimental framework, 7. improvement techniques, 8. results (automatic metrics), 9. conclusion. The thesis is not accompanied by any CD, although there are data and software resources that the author created and that could be potentially very useful for the research community (see below).

Main contributions:

1. Adaptation of Vilar et al.'s error hierarchy and its usage to manually analyze en-ur MT output.
2. Collecting new data sources: multiple translations of Quran (3 English and 4 Urdu), religious texts Sahih Bukhari and Sahih Muslim, the dictionary from urduseek.com, monolingual Urdu blog data. Sadly the data is not attached to the thesis on a CD.
3. The rule-based tag postprocessing (case marking) tool and the dependency reordering tool would be an interesting resource for the research community, if they were attached to the thesis on a CD.

The text could be structured better. Especially conducting the error analysis of the system output before introducing the data, the system setup and the BLEU score results is strange. The usage of the English language is in general quite good, although there are some grammatical errors, typos and typographical problems (e.g. missing spaces after punctuation). Since Urdu is written using a non-Latin script, all Urdu examples should be accompanied by transliteration to make them more illustrative for the non-Urdu speaking readership; this is unfortunately not the case. Moreover, even the examples in the original script seem to have problems, see below for details. Also see below for separate comments on Bibliography.

It is natural that the contents of the survey part of the thesis has been adapted from existing literature; however, based on the various mistakes and inconsistencies I found in the text, I am unsure about to what extent the author actually understands the text she is presenting.

Regarding the experiments reported, these are very poorly documented (see the main comments). It is difficult to draw any conclusions from them, not to mention reproducing the work and building upon it.

Main Comments and Questions on the Text

- In the example that probably shall be (but is not labeled so) Figure 2.2, "many-to-one word alignment" is probably meant to be denoted by underlining a *word sequence on*

the Urdu side. However, the usual and more easily understandable way is to use several arrows originating at the same English word. Also, is the phrase "مارا سے خنجر" / "mārā se xnjr" in the example in correct word order? What is the verb here? Shouldn't a noun go before the postposition and the verb after them?

- The last (unnumbered) equation on page 14: what argument is maximized in the argmax function here? F? A? Or something else? What is the value range of the parameter θ?
- Page 15: Using the same index i for both the French and the English side seems misleading. Or is it intentional? That would indicate that there are same numbers of phrases on both sides and no reordering at all.
- Page 16: From the two equations on top of the page it follows that $\frac{P(e_1^I)}{P(f_1^I)} = 1$. So are both these probabilities equal? Why?
- Page 16: The next unnumbered equation uses unexplained symbols $N, \alpha, x_{i(E|F)}$. If the explanation of same-named symbols in the next section, Log-linear Model, should apply to this equation, it would have to appear **before** the equation, or immediately after it at the latest.
- Page 16: What is e_I^I?
- Page 17, Equation 2.6: Could the author explain "$\prod_{i=1}^{N} \alpha_i^{x_1(E|F)} \alpha_i^{x_2(E|F)}$"?
- Page 17, last two unnumbered equations: Why is the argmax missing from the second one?
- Page 20: "For instance, representation of read in English will be surface-form angry | lemma angry | part-of-speech VBD." – What does it mean?
- Page 20: What tagset is being used here? The Penn tagset? If so, why is "angry" tagged as "past tense verb" (VBD)? Isn't it an adjective?
- Page 22: Integrated Recasing: The author does not explain, what is the purpose of recasing in Urdu, where there is no distinction of uppercase and lowercase letters.
- Page 25: This formulation suggests that the correct word order in "King of Mankind" is "بادشاہ لوگوں کا" / "bādšāh logoñ kā". I may be mistaken but have some doubts whether this is correct. It would be better if the author showed the whole phrase in the correct word order.
- Page 27, first example: The reference translation is very strange. Is it possible that the words are in reversed order, i.e. left-to-right? (While the characters within each word are written correctly right-to-left?) Not only the text lacks any transliteration for the readers who don't read Urdu, it also encrypts the examples for those who do. It is especially annoying in the TST sentence where the Urdu text is mixed freely with some English comments and the parentheses (part of the translated text) are placed just randomly. I strongly advise against putting on the TST line anything that has not been directly output by the system! Of course, the comments are necessary, but a clearer way must be found for binding them to the output text.
- Page 27, second example: the comments in the TST line are cryptic. What word/phrase does each comment apply to? If this shall be error annotation, consider

aligning the annotation to the annotated phrases vertically, i.e. place comments above or below the phrases, clearly indicating what belongs together.

- Page 28: "In this examples, literal translation of "comes" is present, which..." does not say the important thing, i.e. that it is present in the TST sentence while it is missing from the REF sentence.
- Page 29, Table 1: The numbers of the different types of errors are useless if we don't know the size (sentences and tokens) of the analyzed data. Also, how many sentences are understandable to some extent (if any)?
- Page 30: The EMILLE corpus is not "available on the web". It has to be acquired from ELRA/ELDA. Also, for CRULP data to be used, a license for the English texts of the Penn Treebank is needed. The crucial point however: references to these corpora are missing, both bibliographical and URL.
- Page 32: It never occurred to me that the German umlaut "ö" comes from traditional Arabic literature. Does the symbol really look like this?
- Page 33: "Therefore, brackets were removed from the Urdu text." — Only the brackets, or including the text inside? Does the English text lack the brackets only, or the text as well?
- Page 33: "For English,left3word-wsj model accompanied with tagger" — What tagger was used? Any references, literature, URL?
- Page 33: The English factored example is not tokenized (tokens "beneficent," and "merciful."). It should be tokenized before translation.
- Page 33: In the Urdu factored example, the stem factors do not correspond to the word forms. They are taken from other tokens. Why?
- Section 6.2: What was the setting of the software tools? What were the Giza++ alignments computed on? Whole word forms? Stems?
- Page 36, Urdu Tagging: Missing reference (citation, URL) for the Stanford POS tagger. Was the tagger trained on the 4K CRULP sentences? Has the case-marker tag postprocessing been implemented by the author? If so, it would be a valuable contribution for the research community and should accompany the thesis on a CD. If not, then the source of the postprocessing tool should be mentioned.
- Page 37: "helping verbs" are usually called "auxiliary verbs" in English.
- Page 37: The table of case markers omits the genitive. The "example of case markers" above is useless, it is probably an Urdu sentence or two without any explanation, not to mention translation.
- Page 39: Missing reference (citation, URL) for the Stanford dependency parser.
- Page 39, the example "She has been writing a novel **since** last month." – The word "since" is missing from both the dependency tree and the reordered English sentence.
- Page 39: The dependency reordering tool operating over the output of the Stanford parser could be potentially useful for others. It would be nice if it was distributed with the thesis on a CD.
- Page 40: Reordering of "Weather of Northern areas in summer" results in "Northern areas of summer in weather". However, the word order of the Urdu sentence

corresponds to "summer in Northern areas of weather". Why does reordering of the "X of Y" phrases stop half-way and ignores the Y member? BTW, Urdu examples such as this one would benefit from glosses, i.e. literal word-by-word English translation.

- Page 41, experiment with only Quranic data: What language model was used? What data was it trained on? What was the order of the LM?
- Page 42: "Resultant corpus comprised of 20708 lines..." – so what combinations of those 3 English Quran translations and 4 Urdu translations have been added here?
- Page 43: "With **Tafseer**" – What is this?! Section 5 does not mention such resource. There should be a description and a URL reference, possibly also citation of a relevant publication.
- Page 43: "Inclusion of this book into the corpus decreased the BLEU score..." – Is the test data still Quran-only? So possibly the BLEU drop is because of very different domains. What domain is Tafseer? When was it written, what is its subject?
- Page 43 and 44, "Effect of Large Language Model": Again, what is the order of the large LM? What LM was used in the earlier experiments. Development and test data are now different size, where are they taken from? All these questions must be answered so that we know what the indicated BLEU score of 0.1617 is to compare to. (I.e. another question: what is the BLEU and NIST score of exactly the same training, dev and test data, only with the smaller LM?) Without that, the concluding sentence that *"BLEU score **confirmed** that using a large monolingual data increased the quality of translation."* is an invalid claim.
- Page 44, factored translation: "Development set consisted of 800 sentences and test set comprised of 200 sentences." – What domain? Where are the sentences taken from? Are they the same as in the previous LM experiment?
- Table 8.9: Why is the test BLEU score so low now? Since we are only using one factor with surface forms, shouldn't the score be equal to one of the previous results in unfactored experiments?
- Page 45: "we decided to map Urdu PoS tag set on a smaller set of tags" – This is very interesting. It would deserve an appendix to the thesis with the list of the mappings, accompanied by an explanation of the meaning of the tags.
- Page 45: So what is the "generalized tagging"? Tagging with the reduced tagset? Or this together with the rule-based case marking?
- Page 46: There is no table showing the impact of the rule-based case marking on the BLEU score.
- Very important point for Chapter 8: The manual analysis of errors, as presented for the baseline system in Chapter 4, should be repeated here for the outputs of the modified systems, especially those that hurt the BLEU score. If the author designed a methodology for error analysis and then she proposed adjustments based on such analysis, it is only logical to expect that the adjustments will also be evaluated using same analysis. Otherwise how can we reliably tell what happened to the factored translation that it did not help? Poor performance of the tagger (see also Conclusion of the thesis) is a likely *hypothesis* but where is the evidence that supports it?

Bibliography

- The bibliography is disastrous. There are missing years (first Koehn et al.), misspellings (Carpaut, Kochnn, Habbash), a mysterious author named "—" and falling alphabetically between two Och's. For about 15 bibliographic items I found no mention in the text of the thesis. On the other hand, the following **20** (!) publications, cited in the thesis, are not listed in bibliography:

    1. Brown et al. 1990 (p11)
    2. Brown et al. 1991 (p11)
    3. Gale & Church 1991 (p11)
    4. Shannon 1948 (p11)
    5. Weaver 1949/1955 (p12)
    6. Yamada & Knight 2001 (p13)
    7. Dempster et al. 1977 (p14)
    8. Falagan et al. 1994 (p23)
    9. Toutanova & Manning 2000 (p33)
    10. Ahmed et al. 2007 (p37)
    11. Al Onaizan & Papineni 2006 (p38)
    12. Koehn et al. 2005 (p38)
    13. Tillmann 2004 (p38)
    14. Galley & Manning 2008 (p38)
    15. Zollmann et al. 2008 (p38)
    16. Quirk et al. 2005 (p38)
    17. Wu 1997 (p38)
    18. Collins et al. 2005 (p38)
    19. Xia et al. 2004 (p38)
    20. Xu et al. 2009 (p38)

Technical Comments

- Page 9: "Vilar [et**.el**., 2006]"
- Page 11: "In 1993, Brown described" → "Brown **et al.** (1993) described"
- Figure 2.1: Swapped E and F contrast not only with the customary notation in the MT community but also with the text immediately below ("considering a French sentence F as an encoded English sentence"). Introductory examples (including the text of this thesis) usually assume that F = French = foreign = source, and E = English = target. Bad formatting of the figure makes the F on the right partially invisible.
- Equation 2.2: "$E' = P(E).P(F|E)$" misses the $\mathbf{argmax}_E$ function. The translation cannot be the product of two probabilities, unless we agree that French "*le chien*" is translated into English as "0.00000000000273".

- Missing Figure 2.2, or missing caption of Figure 2.2 (as it is probably the example immediately after referring to the figure).
- Page 14: Why are some equations unnumbered?
- $\prod_{J=1}^{J}$... looks like a cyclic definition ("let's iterate J from 1 to J"), should be $\prod_{j=1}^{J}$.... The same happens later on the page with S.
- Page 15: "null words" are mentioned without previously explaining them. MT-aware readers could be expected to know what the author is talking about but, well, for such readers the whole introduction here is useless anyway. So I take it, a less advanced reader is aimed at here.
- Page 15: "**Kohen** et el (2003)" → "**Kochn** et al. (2003)". Besides, the reference is ambiguous. The Bibliography contains (**Kochnn** and Knight, 2003) and (Kochnn, Och and Marcu, 2003). The latter is the correct one here, probably. The bibliographic items in the same year should be distinguished somehow, e.g. "Koehn et al. (2003b)".
- Equations 2.4, 2.5 and 2.6: "$\arg\max_e$" → "$\arg\max_E$"
- Page 26: "lrig doog" ... unlike Urdu, in English it is customary to write from left to right.
- Section 6.1, page 32: "Due to unavailability of English-Urdu parallel corpus, we searched web..." — explaining this right after Chapter 5 is superfluous — "... Details of data acquisition are discussed in chapter 7, where..." — and this is simply not true. It is described in Chapter 5, while Chapter 7 is about "searching for better phrase based machine translation".
- Page 33: strange citation "(et el Toutanova and Manning. 2000)"
- Page 34, Tuning refers to "Section 2.4". There is no such section.
- Page 39: The phrase "A quick brown fox" appears suddenly without warning. It should be labeled as an example, and accompanied by an explanation that the second line is its possible dependency structure.
- Page 44: "scores are nearly identical to the baseline in table 1" – Table 1 is on page 29 and contains no BLEU scores at all.

English Language Errors and Typos

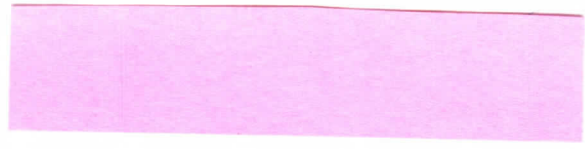
- Page 9: "To identify most frequent errors and **its** source."
- Page 11: "systems which forms"
- Page 12: "to tackle translation problem" → "to tackle **the** translation problem"
- Page 13: "sentences are break down" → "sentences are broke down"
- Page 14: "Let A denotes set of alignments" → "Let A denote set of alignments"
- Page 15: "can also be model by" → "can also be modeled by"
- Page 17: "probability is **knows** as distortion"
- Page 19: "Section 3.1 provided" ... past tense is strange given that this appears *before* Section 3.1.

- Page 20: "lemma and morphological information is translated separately and combine" → "lemma and morphological information are translated separately and combined"
- Page 21: "scoring functions, which helps" → "scoring functions, which help"
- Page 22: "traning" → "training"
- Page 23: "ranked error types based on its effect" → "ranked error types based on their effect"
- Page 26: "according the tense" → "according **to** the tense"
- Page 26: "In the above two sentence" → "In the above two sentences"
- Page 26: "agreement of number, gender and **number**"
- Page 28: "In this examples" → "In these examples"
- Page 28: "Urde sentence" → "Urdu sentence"
- Page 28: "reveals that there two main sources" → "reveals that there are two main sources"
- Page 30: "pair of language" → "pair of languages"
- Page 31: "all English version has" → "all English versions have"
- Page 38: "imp9rovements"
- Page 39: "therefore dependency structure is also swapped here" → "therefore **the** dependency structure is also swapped here"
- Page 41: "with different corpus" → "with different corpora"
- Page 42: "by crawling **a** urduseek.com" → "by crawling urduseek.com"
- Page 42: "converted in to multiple one" → "converted into multiple ones"

Conclusion

The topic of the thesis is very interesting and the thesis hints that the author has done some interesting work in en-ur corpus acquisition, error analysis, Urdu tagging, Urdu tagset mapping and English dependency structure reordering. However, the work is not described well and some more evaluation and error analysis is also desirable. The text has been patched together in a rush, numerous small technical problems such as misformatted figures, missing bibliographic references etc. could be corrected easily. I believe that the thesis would benefit greatly if the author rewrote it and resubmitted, bearing the above comments in mind.

Jenštejn, August 27, 2010



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