Review of Doctoral Thesis Presented by Loganathan Ramasamy

Reviewed by Daniel Zeman.

The thesis presents results of research that the candidate conducted in the field of automatic dependency parsing (syntactic analysis) of natural language. The main focus is on transfer of necessary language resources from resource-rich languages (such as English) to resource-poor languages (five Indian languages are tested in the experiments: two Dravidian – Tamil and Telugu – and three Indo-Aryan – Hindi, Urdu and Bengali). Various scenarios are tested depending on what resources are actually available.

Syntactic parsing is a central task in natural language processing. For a few languages including English it may seem a solved task because it has reached the state where it is very difficult to bring new improvement. For other languages, the task is not solved at all because there are too few resources (annotated language data) to successfully apply the same methods as for English. State-of-the-art methods employ *supervised learning,* i.e. syntactic structures are automatically learned from large training data where the structures have been annotated by humans. There are also *unsupervised methods* (e.g. Mareček and Straka, 2013) that work without human-annotated data but they suffer from far lower accuracy. Methods studied in the thesis can be characterized as *semi-supervised* as they reuse human-annotated data but the data are from a different language and their usage is not straightforward.

On my opinion, the main contributions of the work are the following:

- The candidate created a new treebank for the Tamil language from scratch. It is a small corpus for evaluation purposes, yet it is the first freely available treebank for Tamil and as such it is an extremely valuable resource.
- The candidate also collected a parallel English-Tamil corpus and made it freely available for the research community.
- Previously published approaches by Hwa et al., 2005 (projection of dependencies across parallel corpus) and Zeman and Resnik, 2008 (delexicalized parsing) have been thoroughly tested on five Indian languages. Despite belonging to two language families, these languages share some properties that are important for parsing, such as the SOV word order and predominantly left-branching structures. These languages were chosen because they possess some minimal resources necessary for evaluation of the experiments. However, there are dozens of other related Indian languages with millions of speakers and virtually no available language resources. Thus there is a huge potential to employ the results of this work in practice.
- A novel approach has been proposed and successfully tested for the scenario that a human-translated parallel corpus is not available but an MT system for the given language pair is available (the online translation services).

I appreciate the author's effort to perform error analysis in Chapter 7, yet I find this chapter relatively weak in comparison with the rest of the thesis. I found it difficult to draw conclusions from the analysis presented, i.e. what part of the system should be improved first (and possibly how) to achieve better results. It is of course possible that some of the observations turn out inconclusive and unexplainable but then I would expect the author to state that explicitly.

The dissertation is well organized and written clearly, in well understandable English (although it would benefit from proof-reading by a native speaker, as there are grammatical errors). The cited literature is comprehensive.

To summarize, I believe that this work is a nice contribution to the field of dependency parsing of underresourced languages and that it clearly demonstrates the author's ability to conduct independent research and present its results.

Specific questions and comments

1. Page 32: Is it standard to use the “Latin” full stop for abbreviations and sentence breaks in Tamil? Are there also Tamil-specific characters for these purposes?
2. Page 34: Is the `no_space_after` attribute used only for split compounds, or also for punctuation that was originally attached to the previous word? I would expect the latter, in which case the 10% would not indicate percentage of compounds.
3. I appreciate that many figures and examples have four-way legend of Tamil in Tamil script, Tamil transliterated, English glosses and fluent English. Yet for some reason, other cases lack the translation, e.g. Figure 4.4 on page 35.
4. Page 44, Position 7. Why declare the inanimate gender, if it is unused?
5. Page 47: What is the difference between “specific” and “non-specific” indefinite referential pronouns?
6. Page 52, Figure 4.10: Why is the Tamil equivalent of “for development” classified as adverbial? From its English translation (and also from the Tamil tags, saying that this is a noun + postposition) I would say that it is an object.
7. Page 53, Figure 4.11: What are the differences between annotation of apposition in PDT and in TamilTB? This does not look like PDT (no `is_member` flags, three different children of the Apos node).
8. Page 79 and onwards: Taking the PDT-like tag from HamleDT and calling it “Interset tag” is unfortunate. These tags are in HamleDT only for convenience of us Czech developers, while the Interset features (the structure “iset”) should be the authoritative source of information. It's because the PDT-like tag is restricted to what the Czech tagset can capture. Interset features and/or feature values that are not in PDT are not visible in this tag.
9. Page 93 and elsewhere: Have you asked the question, why Bengali works well with Telugu (suggesting they are somehow similar) and Hindi works well with Tamil, despite the fact that in either of the pairs the languages are from different families? I believe it's because the Bengali and Telugu treebanks are in fact very special: they only contain chunk heads but they drop all words that are not chunk heads. Thus we do not have real Bengali and Telugu data, we have rather something that could be considered two new languages (very impoverished languages indeed). It would be better to abandon the CoNLL format (used so far in HamleDT) and use the data in their native SSF format where the words are not missing. Or alternatively, I could imagine an experiment where the Hindi data would be artificially restricted to chunk heads in the same way as in Bengali and Telugu. Then I believe Hindi would look as similar to Bengali as Telugu.
10. Page 100: The results in Table 6.6 are quite surprising and interesting.
11. Page 106, Table 6.8: What does it mean that there are dashes in four rows of the table? Are you suggesting that these languages do not have adposition-noun dependencies? If so, why do they include Slovak?
12. Page 107: “unlike other IL counterparts, `ur` does not explicitly mark the preposition/postposition at the POS level”. I think this is not true. I found a postposition *ke* (tagged PSP as in other ILs) right in the third token of the first file of the Urdu treebank:
    ```
    3      ((    NP    <fs name='NP2' drel='ccof:CCP'>
    3.1    نواک  NNPC  <fs af='نواک,n,m,sg,3,d,0,0' posn='40' name='نواک'>
    ```

```
3.2    جوکووچ        NNP   <fs af='جوکووچ,n,m,sg,3,o,0,0' posn='50' name='
جوکووچ'>
3.3    کے     PSP    <fs af='کے,psp,,,,,,' posn='60' name='کے'>
3.4    بعد    NST    <fs af='بعد,nst,m,sg,3,d,,' posn='70' name='بعد'>
       ))
```

13. Page 113, Table 6.12. You perhaps have said it elsewhere but it would be fine to repeat here, what parser was used to obtain the results. The MST parser, second order, non-projective? Trained only on parts of speech and not on other features? Or is it already delexicalized? I am asking because the numbers seem low to me, I was able to obtain higher numbers on these treebanks. And exact description of the experiment is crucial to make the numbers comparable.
14. A comment: if you do not have HamleDT-harmonized Urdu, you could have also compared all the treebanks except Tamil in their original annotation. Urdu and Hindi were created by the same team and should have compatible style.
15. Page 116 / a comment: Visible differences between predicted and gold POS tags for Bengali and Tamil could be because these two treebanks are very small and if you trained a POS tagger on them, it may not perform very well.

References

Mareček, David and Straka, Milan (2013). Stop-probability estimates computed on a large corpus improve Unsupervised Dependency Parsing. In *Proc. of ACL*, pp. 281–290, Sofia, Bulgaria.

Hwa, Rebecca and Resnik, Philip and Weinberg, Amy and Cabezas, Clara and Kolak, Okan (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Nat. Lang. Eng.*, 11(3):311–325.

Zeman, Daniel and Resnik, Philip (2008). Cross-language parser adaptation between related languages. In *IJCNLP 2008 Workshop on NLP for Less Privileged Languages*, pp. 35–42, Hyderabad, India.

In Dublin, 25 August 2014

RNDr. Daniel Zeman, Ph.D.