Review of Doctoral Thesis Presented by Nathan David Green

Reviewed by Daniel Zeman, 27 July 2013.

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 combination of multiple parsers into an *ensemble system* that should achieve better parsing quality than the best individual parser. In the last section, attention is paid to more realistic evaluation of parsing quality: besides the standard dependency accuracy, parsers are also evaluated indirectly by the applications that use their output (machine translation in this case).

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; yet there is still enough to improve. For other languages, the task is not solved at all because there are too few resources (annotated data) to successfully apply the same methods as for English.

Using an ensemble of diverse systems that solve the same task by different approaches (thus doing different errors) is nowadays a well established technique used across the natural language processing field in order to improve quality of output. It has also been applied to dependency parsing, as the author acknowledges. Nevertheless, never before was the technique investigated and evaluated in such a depth. The previous attempts were limited to voting with simple weights, without much insight into how the approach could be futher improved. So the main contribution of this part of the thesis, on my opinion is:

- in-depth comparison of several voting configurations and scenarios
- performing error analysis, implementation of a method that identifies parts of speech that are particularly difficult for a particular parser, and makes use of that knowledge
- demonstrating how an SVM classifier can be trained that tells for a given configuration which of the individual parsers should be trusted
- evaluating the technique on five different languages, including two underresourced ones

I also very much appreciate the final section where parsing quality is evaluated indirectly via machine translation. Syntactic parsing is rarely an end-user application. In a typical case it is one of the early steps in a language-processing pipeline. Thus for an end user, the impact on the end application is much more interesting than pure parsing accuracy. There are probably few researchers that would disagree; yet such indirect evaluation is seen very rarely.

An important by-product of the research is new human-annotated data for English: the manually corrected dependencies of the section 23 of the Penn Treebank, and dependency annotation of the English side of WMT 2012 test data. I share the author's belief that this new piece of data will be very valuable to the research community.

The dissertation is well organized and written (apart from a few typos). The language is, for the most part, clear and understandable (the exception is one section where I had some difficulty understanding what was done; see below). The cited literature is comprehensive.

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

Specific questions and comments

- Table 3.3: The best English UAS (probably the best of the whole thesis) is 92.58. However, later on p. 48 the author claims that 92.54 is the best. Why? And the best number in Table 3.6 is 92.55. Are all these numbers mutually comparable?
- Section 3.1.1.4: Although there does not seem to be any other plausible explanation, I would appreciate if you were more specific about what you mean by a "POS error". Am I right in assuming that e.g. an NN error is an instance of a node whose part-of-speech tag is NN and whose parent was determined wrongly by the parser in question? And when you say "correctly predicted POS's tags" as on p. 46, you actually mean "correctly predicted parents for nodes with POS tags"? (And by the way, are the POS tags that you use to classify the errors gold-standard tags?)
- Section 3.1.2: It took me some effort to understand what was clustered and why. There are multiple different sets of "weights" that do not seem to be distinguished terminologically. It should be made clear right in the beginning of the Section, otherwise it is an annoying reading. Could you confirm (or reject, and explain) that:
    - Part-of-speech tags are assigned to clusters. Clusters are fuzzy, which means that every tag more or less belongs to any number of clusters. The extent of the bond between the tag T_i and cluster C_j is expressed by what you call "weight of cluster for POS tag" in Table 3.8.
    - The measure of similarity of tags, that determines their affiliation with clusters, is defined so that members of the same cluster are similarly difficult for a particular parser to process?
    - For every parser (model) P_k we measure its accuracy of assigning parents to nodes with tags from every cluster C_j. The weight of this parser in the ensemble system (i.e. how much we trust this parser on that particular type of node) will be proportional to the accuracy. This is what you call "cluster weights for each model" in Table 3.7.
- Page 52: "the [Italian] ensemble system ... does reduce error for each individual POS." I deduce that it reduces the overall error rate, thus it increases the UAS. However, according to Table 3.12 (as you also admit on p. 50), none of the combinations were able to achieve as high a score as the best individual model, i.e. no increase occurred, right? Am I missing something?
- There should be a table with explanations of the part-of-speech tags that occur throughout the thesis (the Penn English set and the Italian set) so that the reader can better understand why some of them are more difficult for the parser than others.
- Figure 3.7: The diagram would be easier to understand if there were arrows to make it a directed graph.
- Page 64 and elsewhere: Why is the "improvement over the *average* dependency model" frequently mentioned? I would think that an ensemble system is only interesting if it beats the best individual parser. What benefit is there in beating the average?
- Figure 3.10: For both curves the final UAS after 10 iterations is same or worse than before self-training. But Table 3.20 below mentions a 1.1% improvement after 12 iterations (not shown in Figure 3.10). The best UAS in the figure is achieved after iteration 3. Can we know how many iterations we should perform in any particular configuration so that we achieve the best UAS?
- Section 4.1: You measure BLEU and NIST changes of syntax-based MT with parsers that differ in their treatment of deep noun phrase structure. If you retrain the parser on data that

differs in deep noun phrase structure, it is possible that the newly trained parser will deliver slightly different structures also outside noun phrases (just because the model is different). Have you looked into this to see whether it occurs at all and if so, to what extent?

- Conclusion: The author says that the SVM meta-classifier is the most valuable achievement because it is linguistically independent. However, I would not say that the other ensemble techniques are much language-dependent. The only dependence I see is that there must be *some* part-of-speech classification available.

In Jenštejn, 27 July 2013

RNDr. Daniel Zeman, Ph.D.