Review of the Master Thesis

Author: Akshay Aggarwal
Title: Consistency of Linguistic Annotation
Reviewer: doc. RNDr. Markéta Lopatková, Ph.D.

The thesis focuses on identifying potential inconsistencies and annotation errors in the Universal Dependencies (UD) Treebanks. The thesis project aims to automatically search for instances of suspect morphological and syntactic annotation, manually evaluate the results on sample data, and investigate possible automatic corrections.

The overall structure of the thesis is clear and logical – it consists of 8 chapters (3 short chapters introducing the problem, four chapters describing four (sets of) experiments with different sets of UD treebanks for different languages, one chapter summarizing possible future work, and conclusion; further, it includes lists of references, figures, tables, and abbreviations. Besides, five appendices are provided, summarizing different features of UD treebanks and languages. No CD with software, data, and the thesis text is attached; the author provides links to the GitHub repository containing the scripts, some documentation, and some results.

The thesis starts with the introduction of the UD project and with the motivation (**chapter 1**), **chapter 2** then identifies main areas that are addressed later, and **chapter 3** shortly summarizes selected already published methods that serve as a basis for the author's work. These introductory chapters are short but sufficiently detailed and clearly structured.

The core of the thesis is formed by chapters four to seven. **Chapter 4** is focused on languages with more than one UD treebanks. Akshay tests the so-called inter-treebank harmony; the proposed method is based on the coarse POS trigrams and techniques introduced in [Rosa and Žabokrtský, 2015] – a new metric, $theta_{POS}$ is suggested. Further, the author tries to empirically set a threshold serving as an indicator of "harmonicity" of these treebanks, using results of the so-called UD validator (high-quality treebanks should satisfy this threshold). This threshold is then further modified to reflect data size as well as genre distribution of available corpora. This empirical method seems interesting. However, setting the threshold just on a minimal number of corpora (two corpora for size, two corpora for genre diversity) seems to be an inappropriate generalization. I would appreciate at least some manual evaluation of the results. Do these results correlate with the rating by the UD validator?

Chapter 5 focuses on coordination. It describes experiments trying to find and correct improperly annotated conjunctions (here errors can be caused not only by the complexity of coordinated structures but also by the change in guidelines between UD versions 1 and 2). The basic idea is promising, but the experiments seem to suffer from some degree of inconsistency and (probably) last-moment changes; moreover, their description is a bit chaotic. The author focuses on non-projective structures first (Algorithms 3-4), then he proposes rules operating on all structures (Algorithms 5-7). However, it is not clear how the constraint of (non-)projectivity is defined and implemented – while Algorithm 2 should "re-hang" a conjunction to a new head only if it outputs a projective structure, some examples seem to contradict this (ex. 15b, 16b, 17b). I highly appreciate the rich exemplification of the proposed steps (on Arabic and Afrikaans); though some examples are a bit misleading (e.g., ex. 10 should focus on the conjunction *but* introducing the sentence but the author mentions preceding content words? pronouns are not typically classified as function words?). Here are several points that should be answered during the defense:

- Please explain the condition on projectivity, with a focus on "projective attachment of

a conjunction".

- Explain and exemplify the concept of a conjunction sandwich.
- Is there any reason why Algorithm 4 relies on universal POSs, while Algorithm 5 works with deprels? The choice of deprels is not clearly motivated – why do you skip obj and iobj deprels?
- How many conjunctions in the processed corpora are not attached to a node with the conj deprel? Can this simple criterion serve for revealing "suspicious" annotations?

In **chapter 6**, Akshay introduces his experiments with the LISCA algorithm [Dell'Orletta et al., 2013; Alzetta et al., 2017] ranking dependencies by their reliability – I suppose that he uses the implementation of Alzetta et al. (but it is not clearly stated in the thesis). Akshay focuses esp. on the problem with the lack of gold-standard data – to overcome it, he proposes a method of *k*-fold cross validation and discusses the impact of different *k*. He also offers an evaluation on Hindi data – he classifies and exemplifies different types of errors.

- You limit yourself to 0-score edges (with precision about 0.5). Can you somehow estimate recall (on the sample Hindi data)?

The last (except for future work and conclusion) **chapter 7** deals with distinguishing instances of auxiliary verbs (POS tag AUX) and full verbs (POS tag VERB), focusing on Hindi. The author seems to contradict himself here – he characterizes this distinction as relatively simple (p. 99 saying that "In hi [=Hindi], we can more often than not draw a clear line of distinction between auxiliary as defined by UD, and the verbs"; however, no baseline is given so we cannot assess this assumption). On the other hand, he is surprised by the high performance of the tagger (F1 score over 99) and a low number of detected problematic instances (p.102).

Summary

This text represents the second version of the thesis. The author left out the most problematic chapter dealing with non-projectivity and the method introduced as "variation nucleus". Other experiments are more elaborated and more precisely described (which I consider as very important). It must be stressed that there has been much work done. I also appreciate a rich list of bibliography Akshay has gathered and studied.

As for the language of the thesis: the text would benefit from proofreading (esp. many missing or redundant articles, some typos. However, the text is well comprehensible in most parts, the motivation for the experiments is presented, and one can typically follow the experiments without bigger problems.

Conclusion

I can recommend the current version of the diploma thesis for the defense.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

In Prague, September 2, 2020