Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Akshay Aggarwal		
Název práce	Consistency of Linguistic Annotation		
Rok odevzdání	2020		
Studijní program	Informatika	**Studijní obor**	Matematická lingvistika

Autor posudku	RNDr. Daniel Zeman, Ph.D.	**Role**	Vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky		

Text posudku:

The goal of the thesis was to explore methods to automatically detect annotation inconsistencies in the Universal Dependencies (UD) corpora, and to see whether some of these inconsistencies can be automatically corrected. The author approaches the problem from various angles: he experiments with rule-based heuristics as well as various statistical methods such as KLcpos3 or the LISCA algorithm.

There are 137 pages, out of which roughly 83 describe the author's own contribution (chapters 4 to 7). One chapter is devoted to the overview of related literature, which is also abundantly cited throughout the text. In one of the appendices the author provides statistics of non-projective and non-planar trees in all UD treebanks – as far as I know, such a study has not been published before. The thesis is based on substantial amount of experimental work.

The structure of the text is quite standard and within the chapters (especially the four chapters devoted to the four experiments), the approach is well motivated, described and the results discussed.

The author's usage of the English language is not perfect and some sentences are somewhat complex but it is still well understandable.

To summarize, I recommend the thesis to the defense.

Specific questions and comments

- Section 4.5.3, page 38, equation 4.10: At other places, we try to set an upper limit (threshold) Θ on the value of $\theta_{pos}(A,B)$. The threshold depends on general settings, such as whether A and B contain the same genres or not, but not directly on the annotation of A and B. However, in equation 4.10, partial $\theta_{pos}(A,B)$ values, specific to the two treebanks, are first calculated, and then the threshold is based on their average. Is this well motivated? What happens if the partial values are over their partial thresholds? For example, assume that there are two genres (labeled x, y) in A, one of

them (labeled x) also appears in B. Assume that $\theta_{pos}(A_x,B_x) = 2$ (too much according to equation 4.6) and that $\theta_{pos}(A_y,B_x) = 2.2$ (too much according to equation 4.9). However, we set the threshold Θ to the average of the two measured values, i.e., 2.1. Does it mean that if there is much more data in genre x than in genre y, we may be able to declare treebanks A and B as harmonious despite the fact that none of the constituent genres are harmonious?

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 3. září 2020 **Podpis**