Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Adédayọ̀ Olúòkun
Název práce	Creation of a Dependency Treebank for Yoruba using Parallel Data
Rok odevzdání	2018
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 author implemented and evaluated known techniques of projecting annotation from a resource-rich language to a resource-poor language via aligned parallel data. The target language, Yorùbá, is a truly low-resource language despite its large speaker population, hence the research is also actual and potentially useful. As an important side result, the author also created a small manually tagged and dependency-annotated evaluation corpus of 100 Yorùbá sentences; this dataset is now publicly available as part of the Universal Dependencies collection.

The thesis shows that simple projection techniques do not outperform a tagger and parser trained on just 50 manually annotated sentences. Nevertheless, the evaluation of the projection techniques itself is valuable. The author also discusses frequent errors and their possible causes, and concludes that Yorùbá-specific postprocessing rules would be needed to make the projection competitive (this is labeled as potential future work). Another observation is that the available source languages are linguistically too distant from Yorùbá; but once we have basic resources for Yorùbá, the same techniques could be used to tackle other, more closely related African languages.

There are 64 pages, out of which roughly 35 describe the author's own contribution (chapters 3 to 5). The text is reasonably well written and organized, although it could be better: there are occasional typos, unusual formulations ("these sentences were trained" instead of "these sentences were used to train a model") and oversights (e.g., Table 1.5 is about the Human vs. Non-Human category but the column headers say "Countable" and "Uncountable"). Some sections are too concise or appear at places where it is not clear from previous text why we are now switching focus to a new topic (section 3.2.3 about UDPipe). References to literature should be in brackets including the authors' names if they are not part of the syntactic structure of the host sentence. Nevertheless, these reservations are not fundamental, as the contents of the thesis remains clear and understandable.

The author has demonstrated knowledge of related work and literature. There is a number of experiments that methodically explore possibilities of introducing tags and trees to Yorùbá,

and the experiments are followed by short discussion and analysis.

To summarize, I believe that the present thesis complies with the standards explected at the faculty, and I recommend it to the defense.

Specific questions and comments

- Page 26: It is true that Vietnamese is an analytical language like Yorùbá. But it is also true that its UD treebank is relatively small and the parser accuracy is relatively low (see http://universaldependencies.org/conll17/results-las.html#vi and http://universaldependencies.org/conll17/results-upos.html#vi). Therefore other languages may be more suitable for the task. Page 27: "The parallel data in these other languages were automatically tagged with UD POS using UDPipe." … It would be very useful to cite some published results about accuracy of the UDPipe models on these languages, so that the reader knows what to expect.
- Page 30: "The union of word types … 5,967 words." … Are these Yorùbá word types? Why are some of them without POS? Are these words that occur in aligned sentences but are not word-aligned? Table 3.4 shows that there are 6,127 word types in the Yorùbá-Ancient Greek corpus alone. Why is the union of the four language pairs smaller?
- Page 35: Watchtower tagged by voting projections is taken as training data to train a UDPipe model. Bible is taken as test data – but this is not manually annotated; the tags here have been obtained by the voting projection too, right? How can we know that the annotated test data are correct? This could explain the higher accuracy (87.86%) rather than "testing … on more data … might be the reason we got a higher percentage" – there is no causality at all, you could have much worse accuracy on a large test set.

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 4.9.2018 **Podpis**