Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Ronald Ahmed Cardenas Acosta		
Název práce	Universal Morphological Analysis using Reinforcement Learning		
Rok odevzdání	2020		
Studijní program	Informatika	**Studijní obor**	Matematická lingvistika

Autor posudku	David Mareček	**Role**	oponent
Pracoviště	Ústav formální a aplikované lingvistiky		

Text posudku:

The goal of Ronald's thesis is to develop a universal system for context-aware morphological analysis (lemmatization and morphological tagging) using transducing actions grouped by Byte-Pair Encoding that have the potential to resemble morphological processes. The analyser is tested on a typologically diverse set of languages and especially on a low-resourced Peruvian language Shipibo-Konibo.

The thesis is divided into 6 chapters. After the introduction, Ronald describes the Shipibo-Konibo language and its morphological properties. This is followed by a review of related work. In the third chapter, a rule-based morphological analysis of Shipibo-Konibo is proposed. The unsupervised method for obtaining transducer actions and description of neural network architectures used for training morphological analysers are given in Chapter 4. The datasets used, baseline models and evaluation metrics are shown in Chapter 5. The results are shown and discussed in Chapter 6. The last chapter concludes and suggests future work.

The thesis itself has 77 pages including 12 pages of attachments. It is written in very good English, clearly structured, and, with a few exceptions, easily understood. I like the rule-based lemmatization system, which reached excellent results compared to the neural model. However, I miss a detailed description of rules, for example how the system works and how many rules are there. Neural models for lemmatization and tagging were tested in several different settings, nevertheless, the results achieved on the Sigmorphon shared task were quite a bit below the baseline for the vast majority of languages. However, all the results were properly discussed and the limitations of the proposed models were outlined. Other more detailed comments and questions follow:

1. The process of obtaining gold action operation sequences in Section 4.1.2 is not very clear, even though it seems to be a core thing of the thesis. The Damerou-Levenshtein distance is not cited and the "transpose" operation is not explained. What is the "position"? Is it

the position in the original word or in the current state? And after you obtain the set of operations from DL, you first perform BPE and then sorting? Could you show an example of BPEed sequence of actions? Do the joined sequences of actions resemble morphological operations? I cannot see any joined actions in visualisations in the Attachment A.3.

2. Your results on the Sigmorphon shared task are below the baseline. The baseline was quite strong. However, you should set up a different baseline so that you show that your system is doing something reasonable. For example, the baseline lemmatizer could use the word-form as the lemma or perform only some basic rules (e.g. deleting -s, -ed, and -ing for English) Another baseline for comparison lemmatizers could be a machine translation system working on characters and translating word-forms to lemmas. A baseline for a morphological tagger could be predicting individual morphological labels by binary classifiers.

3. The operations with a specific number as position (e.g. “subs._9_-sk”) seem to be very random and highly dependent on the length of the word. Do you have any evidence where it works? Wouldn’t it be better to use also a position relative to the end of the word (e.g. “subs._-2_-sk”)?

4. How many rules you used for the rule-based system of Shipibo-Konibo? How do the rules look like? How much work it was to build such a system?

Overall, I find the thesis very good. Even though the results are not as anticipated, there has been a lot of work done and everything was properly discussed. I recommend the thesis to be defended.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 29. 1. 2020

Podpis: