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

Text posudku:

The author worked on two topics that are both connected to computational analysis of morphology of natural languages, yet they are relatively independent on each other. The main line of research explores new ways of learning morphological processes from data so that previously unseen words can be lemmatized and tagged. The secondary line defines a rule-based finite-state transducer describing the morphology of an endangered Amazonian language, Shipibo-Konibo.

While the machine learning experiments did not lead to a state-of-the-art lemmatizer and tagger, the author's approach is interesting because it tries to directly model morphological processes, using neural networks in a manner that makes their output interpretable. There is a thorough discussion and the output of the model is evaluated from many different angles on languages from several typologically different families; all that provides valuable insights into the studied methods.

In contrast, the finite-state analyzer of Shipibo-Konibo (which is not the author's native language) uses well-established methods, yet it constitutes a solid piece of work with a very useful outcome. The morphological analyzer is fairly complete in terms of inflection patterns, hence it is a useful tool with a potential to help build other resources for Shipibo-Konibo, and to contribute to the survival of the language in the digital era.

There are 81 pages, out of which roughly 30 describe the author's own contribution (chapters 3 to 6). The text is well organized and written in very good English with negligible number of typos. Throughout the text it is quite clear what has been done and why. There is a reasonably sized review of related work and background literature.

To summarize, I believe that the present thesis complies with (even exceeds) the standards expected at the faculty, and I recommend it to the defense.

Specific questions and comments

- Page 15, equation 1.3: What is E? (E_{y|x^{(i)};θ})

- Page 15: “the search space Y(x(i)) in Equation 1.3” … in fact, the expression occurs in 1.4, not in 1.3.
- Page 28: Position can be _A (prefix), A_ (suffix), and _i_ (inner position from the left). Have you also considered learning a numbered inner position from the right?
- Page 29: “This gives the model the chance to choose another word form as next action instead of replacing the string character by character.” … Is this meant to help with irregular word forms, such as English “am” –> “be”?
- Page 43: In addition to the numeric evaluation of the accuracy, it would be nice to see some examples of the typical errors the system does on individual languages.

Práci doporučuji k obhajobě.

Práci navrhuji na zvláštní ocenění.

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

The thesis significantly exceeds the level expected from a master (Mgr.) student, both in quality and quantity of research work it describes. It is timely, innovative, and it has led to several peer-reviewed conference papers.

Datum 27.1.2020 **Podpis**