Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Christian Cayralat		
Název práce	Orthography Standardization in Arabic Dialects		
Rok odevzdání	2021		
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 addressed the problem of computational processing of Dialectal Arabic (DA). While commonly termed dialects, the spoken variants of Arabic are so different from each other and from the Modern Standard Arabic (MSA) that they can be considered separate languages. These languages are rarely written (except in social media), which has two important consequences: 1. very little data is available to train computational models; 2. the orthography is spontaneous rather than codified, so the available data suffers from disunity. The thesis under evaluation mainly addresses the second issue by exploring ways of possible semi-automatic normalization of DA spelling. Experiments are done with one particular instance of DA, the Lebanese Arabic.

The first approach is to use the pre-existing MADAR CODA corpus to train a neural model that converts spontaneous orthography to a standardized form. The results are not bad when tested directly on held-out set from MADAR CODA; unfortunately, the corpus is quite small and too clean in comparison to real utterances from social networks, so it performs poorly on outside data; also, it is difficult to study various types of spontaneous orthography on such a clean corpus.

Therefore the author reorientated himself to acquiring data suitable for the task, and organized annotation of the Shami Corpus with spelling inconsistencies, as well as tags and lemmas. This new dataset, together with the error/inconsistency taxonomy is, from my point of view, the most useful outcome of the thesis. As the author puts it in the conclusion, "one of the biggest achievements is giving the task of spontaneous orthography standardization a clearer shape" – I totally agree with that claim.

The final chapter provides new experiments with neural models on the Annotated Shami Corpus, reports better results and discusses them. The existence of the new data arguably opens other possible ways of extending and improving the models (such as data augmentation), which is (quite justifiably) left for future work.

The author's work is truly interdisciplinary, like the whole field of computational linguistics. There is some significant experimentation with neural models, as well as corpus design and annotation, which requires organizational skills (the author formed and steered a team of annotators) and also application software development (the annotation infrastructure contains a significant contribution by the author of the thesis, although it is not exclusively his work). Finally, the design of the annotation scheme for Lebanese Arabic, which has no codified grammar (but is the author's native language) required a fair amount of linguistic expertise.

On my opinion, the author proved very capable in all of these areas.

There are 91 pages (plus 10 pages of appendices), out of which roughly 55 describe the author's own contribution (chapters 2 to 5). 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. Observations from data are duly discussed and experimental results analyzed. 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.

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 1. září 2021 **Podpis**