Bachelor Thesis Review

Faculty of Mathematics and Physics, Charles University

Thesis author	Yuliya Yamalutdinova
Thesis title	Detection of contradictions in pairs of texts in Kazakh
Year submitted	2019
Study program	Computer Science
Study branch	General Computer Science
Review author	Mgr. Rudolf Rosa, Ph.D. Advisor
Department	ÚFAL MFF UK

Overall

	good	OK	poor	insufficient
Assignment difficulty	X			
Assignment fulfilled	X			
Total size *... text and code, overall workload*	X			

Contradiction detection is a hard and current problem of natural language processing, even when operating on English or another high-resource language. As Kazakh is a low resource language, for which the necessary tools and resources are limited or even non-existent, trying to solve this task for Kazakh is an ambitious goal.

Nevertheless, the author managed to fulfill the assignment, devising and evaluating a solution that performs the task with a low but respectable accuracy; as there was no prior contradiction detection system for Kazakh, I take the contribution of the author as significant.

The work goes quite beyond the original assignment by having a rather extensive research component. While devising one solution to fulfill the task would in my opinion be sufficient for a Bachelor thesis, the author actually devised and evaluated a wide range of different solutions for both of the tasks (monolingual sentence alignment, contradiction detection), including an exploratory fully rule-based approach, a heuristical approach based on various word or sentence embeddings (Word2vec, Fasttext, BERT), as well as several machine learning approaches (Scikit, TensorFlow), for which the author also had to deal with the subtask of cross-lingual resource transfer – the training data had only been available for English and had to be automatically translated to Kazakh, which also included addressing some issues with data handling and cleaning, etc.

I would also like to note that for this work, the author had to acquire a range of skills and knowledge far beyond the scope of Bachelor studies, following several Master courses and studying relevant literature, as well as learning a new programming language and mastering a range of tools.

Thesis Text

	good	OK	poor	insufficient
Form *... language, typography, references*	X			
Structure *... context, goals, analysis, design, evaluation, level of detail*		X		
Problem analysis	X			
Developer documentation		X		

User Documentation		X		

The thesis is written using quite good English and well structured. It includes an overview of the theory behind the tools and approaches used in the thesis (machine learning and embeddings), and cites a range of relevant works. It explains the approach taken in sufficient detail, although I believe more could have been said about the development process – the author has a tendency to describe only the final solution, without the interesting intermediary steps that led to it. Nevertheless, the evaluation is quite extensive and trustworthy, evaluating a wide range of setups and motivating the final choice of hyperparameters.

I appreciate that the text contains several annotated examples, which makes it easier to understand, especially given the fact that the focus language is quite exotic.

The user documentation is not part of the text of the thesis, but is contained in the submitted solution.

Thesis Code

	good	OK	poor	insufficient
Design *... architecture, algorithms, data structures, used technologies*	X			
Implementation *... naming conventions, formatting, comments, testing*		X		
Stability		X		

The code is contained in several Python scripts, which are quite nicely written, with brief comments; a high-level overview of their operation is contained in the text of the thesis. There is also a range of unit tests.

While the source code itself is quite brief and simple, it uses many tools, resources and libraries (FastText, Word2vec, BERT, SciKit, TensorFlow, UDPipe, scrapy), which the author managed to master and incorporate into the solution; as some of these tools are quite complex and advanced, this is a considerable achievement. The solutions seems to be stable, even though it relies on several external tools and libraries, for which it was sometimes necessary to specifically sanitize the data (e.g. by removing special characters that the tools cannot handle).

Overall grade Výborně

Award level thesis Ano

Date

Signature