The thesis describes the application of three natural language processing methods to samples of Japanese and English texts. The methods, concerned with text mining, include: (i) topic-based clustering of Japanese fiction texts, (ii) sentiment analysis of articles from several English-language newspapers dealing with a specific issue, and (iii) summarization of English and Japanese articles. The introductory parts, the experiments as the core of the thesis, and the summary are all framed within a general question of whether such methods are useful in the specific field of Japanese studies.

Description of the experiments is preceded by an overview of the main text mining methods and how they can be used in practical applications, and by another chapter on preprocessing of texts with a focus on the Japanese language, including a section on how to build the Vector space model, representing a sketch of the meaning of one or more texts. The experiments are described in the following three chapters - Ch. 3: Text mining of Japanese proletarian literature, Ch. 4: Sentiment analysis of newspaper articles addressing the problem of Japanese officials’ visits to Yasukuni shrine, and Ch. 5: Automatic summarization of academic articles. A specific research question is formulated for each experiment. The chapters also include introductions into the subject of the texts, detailed descriptions of the methods (as far as this is possible within the scope of the thesis) and discussions of the results. Owing to the availability of abstracts and keywords produced by humans, the third experiment allowed for a formal evaluation.

A summary of the three experiments in the Conclusion states that the results are not equally satisfactory in all the cases. The clustering experiment (Ch. 3) was a clear success, while the summarization experiment (Ch. 5) seems to have fallen behind expectations (only two of the four generated summaries gave better results than a random selection of sentences). The second experiment (Ch. 4) was restricted to English-language press (4 newspapers published in the UK, China, Korea and Japan) and the results were not confirmed across all articles published in the same newspaper about the topic, but seems to have yielded a surprising picture of the British and Korean dailies, which tended to use a more negative language than their Chinese and Japanese counterparts. An extensive bibliography and an
appendix conclude the thesis, showing the texts in the summarization experiment, with parts selected for the manual and/or automatic summary highlighted.

The thesis shows an impressive linguistic and computational expertise of the author. It is well structured, both in the overall setup and in the description of individual experiments. At least to my understanding, the introductory chapters strike the right balance in the presentation of the NLP methods for the audience with a background in the humanities. The results are discussed fairly and to an appropriate detail.

There were surely many options how to go about an assessment of usability of NLP methods for the study of a specific language and culture. Maybe the arguments for the specific decisions to opt for this way rather than another could have been made explicit. For example: why text mining and not, say, parsing or machine translation? Why exactly these specific methods of text mining? Why exactly these texts? My guess would be that these choices were made because they were found to be most useful for someone involved in the Japanese studies and demonstrating the options in the best way, but I wish I could read that in the thesis.

I understand that some other desirables may be well beyond the scope of a bachelor thesis, but I would like to mention at least two: (i) My main concern is that experiments like this should be described in the context of other comparable state-of-the-art methods, tools and results, for English and – if available – for Japanese. There are a few occasional mentions in the text, but the comparison of the methods and results used should be systematic. The impressive bibliography seems to provide enough resources for doing this, but the field is developing fast and up-to-date research may only be reported in individual papers. I admit that to follow developments in such a field is a very demanding task itself, so rather than insisting that the author should have done that I would like to hear what she considers to be the main merit of the case studies. (ii) This is more of a side remark – if this were not a bachelor thesis but a book aimed at those involved in Japanese studies, perhaps with a penchant for NLP, I would love to see, in addition to the persuasive case studies, also pointers to tools and other practical recommendations.

Apart from that, the text is easily understandable, correct content-wise, written in a fairly good English. However, the number of formal problems is not negligible (typography, typos, omissions, grammar, even in the title of the thesis). It would not take more than proper proofreading to fix that.

Despite the desirables and formal blemishes, this bachelor thesis satisfies all necessary requirements I recommend it for defense.

Otázky k obhajobě:

• Could you comment on the two paragraphs above, the one starting “There were surely many options...” and “I understand that some other desirables...”?

Datum: 14. 6. 2019

Ing. Alexandr Rosen, Ph.D.
Ústav teoretické a komputační lingvistiky
Filozofická fakulta Univerzity Karlovy v Praze
alexandr.rosen@ff.cuni.cz
tel.: 721451239