Review of the Doctoral Thesis

TITLE: NAMED ENTITY RECOGNITION AND TEXT COMPRESSION

AUTHOR: VU NGUYEN HONG

The focus of the thesis is the processing of the Vietnamese language from three different points of views. The first is the compression of the Vietnamese text using syllables, trigrams, and n-grams. The second is the normalization of the text written in the informal language, and the third is the recognition of the named entities. The description of these topics follows after the introductory chapter and the chapter with basic definition and state of the art. The chapter 2 is very well written and describes the Vietnamese language and its specifics in an excellent way, and even the non-native speaker can understand the problems in automatic Vietnamese processing. The formal style of the thesis is good but one discord which is that the section titles are followed immediately by the subsection titles without any text. The quality of the English is also excellent. I did not find any typo or error and only very few strange sentences.

The chapter that describes compression of the Vietnamese text is long and exhaustive. In short, all three approaches presented in the chapter belong to the dictionary based compression with static dictionaries. First, the SRILM is defined in the Subsection 4, but it is first used in Subsection 3. The static dictionary approaches suggested and evaluated in this chapter are compared with the standard compression algorithms such as WinRAR, ZIP, and Huffman encoding. The comparison is a little bit strange. The standard algorithms are dynamic compression algorithm that does not need any external datasets or dictionaries. On the contrary, the suggested approaches require large dictionaries extracted from the massive corpora. The results of the suggested algorithms are good but not surprising due to the presence of the large dictionary that models the context of the sentences.

The chapter that describes the normalization of the text is short, but the proposal of the modified Dice coefficient for bigrams comparison is interesting. The results confirm the idea and clearly show that the modification is valid and may be used in similar problems.

The last chapter describes the major topic - extraction of the named entities from the short text messages (Twitter messages concretely). The experimental part is very short, but the described experiments confirm that he suggested modification are useful. The only problem I see is that the normalization used as the preprocessing step (capitalization) uses the part of the algorithm that is described and developed later. Therefore the knowledge is not fully known in the phase of preprocessing.
The overall evaluation of the thesis:
The thesis deals with processing of the Vietnamese language from three different points of view. The compression of the Vietnamese text is described in high details. The experiments presented are sufficient. The results achieved during experiments are not surprising, but the achieved compression is interesting. The normalization chapter is good and presents an excellent idea. The last chapter about named entity recognition is also interesting. The presented approach is novel and promising, but the experimental phase is short, and not all aspects are studied. However, the presented results are interesting.

The publication activities of the student contain four papers on the conferences and two articles in journals that were not published yet.

When I take into account all aspects of the presented work and include all my remarks I think that the work done by the students is very interesting and promising for the future research and

**I recommend the thesis for defense.**

Questions:
The gazetteers used in the capitalization classifier for the named entity recognition is the same as the one used later, or it is different? Why is it not mentioned in Figure 5.1?

Ostrava, 25.10.2016

doc. Ing. Jan Platoš, Ph.D.