The presented work deals with an important natural language processing problems focused on the area of short informal Vietnamese texts from social networks. The objectives of the thesis are to provide a method of text compression and error correction, together with mining the semantic content of the text messages by means of the so-called named entity recognition (NER) approach. The proposed n-gram based methods are tailored to provide efficient solution for specific properties of the Vietnamese language, and are also focused and tested on the domain of short Twitter messages.

The work is organized into six chapters. The first chapter presents a brief context of the research and defines the goals of the thesis. Chapter 2 is devoted to the background, preliminaries and related research. Its three parts introduce the specifics of Vietnamese language processing, including e.g., the typing methods, the basics of text compression techniques, and finally, the named entity recognition is introduced and overviewed. An extensive Third chapter deals with problems of Vietnamese text compression. Several methods are presented, namely syllable, tri-gram and n-gram approaches. All are defined and experimentally evaluated throughout the chapter. Chapter 4 introduces methods for normalization of Vietnamese texts. The focus is on spelling error correction methods suited specifically for informal messages, such as tweets, that also serve as an evaluation example for this work. Chapter 5 draws on the results presented earlier in the text in order to propose a named entity recognition workflow for Vietnamese tweets. The proposal consist of normalization phase, the part-of speech identification phase, feature extraction, and finally, the NER phase realized by the SVM classifier. Again, the approach is experimentally verified on real data. Chapter 6 concludes the thesis by summarizing the results and presenting ideas for future work in the area.

Among the main original contribution of the thesis I would like to emphasize the following:

1. The overview of named entity recognition approaches in general, and for the Vietnamese language in particular, presented in section 2.3 is well elaborated. This part provides an extensive literature survey together with examples of language specific problems, and thus, it serves as a solid base for supporting author’s approaches in the following text.

2. An extensive part on Vietnamese text compression in Chapter 3 is valuable mainly because of its experimental results. Author presents current methods based on syllables and n-grams, and provides thorough tests and evaluations of the suitability for the chosen domain. The properties, and pros and cons of these approaches are gathered and commented on in the concluding part. Author emphasizes different
criteria of suitability, such as the small size of dictionaries necessary for syllable approaches, paid by the trade-off of higher compression of n-gram methods.

3. Similarly, the normalization of Vietnamese texts, presented in Chapter 4, has a high potential impact for practical applications. The error correction methods, proposed by author have proven be very efficient for the domain at hand. Moreover, the modification of Dice coefficient can be an interesting general result which deserves further elaboration.

4. The whole thesis is nicely focused on the semantic analysis of tweets that constitutes the fifth chapter of the thesis. I really liked the way how previous chapters provided the solid domain-dependent preprocessing and feature extraction methods in order to utilize a machine learning approach for the named entities detection.

Concerning the methodology and practical usability of the approach, especially in different domains, I have two remarks:

1. The NER approach depends quite a lot on the preprocessing parts, such as the normalization, POS detection, etc. Is it true to say that when moving to different domains (such as the Q&A systems mentioned by author in the conclusions), all parts of the workflow would remain the same, with the exception of the named entity definitions and training the classifier?

2. Have you considered other machine learning methods instead of SVM? Was there some inherent property of the SVM that made it superior in your approach?

To summarize my report: The thesis present a solid work of scientific research. Its goals are well specified and fulfilled in the work. The author is well aware of the context and related work and is able to focus on a new niche in the area and to bring new, original results. These results are moreover well founded by several international peer-reviewed publications. I believe that Vu Nguyen Hong has proven to be able to create a substantial and original research, and thus, based on the presented thesis, I propose to award a PhD degree to him.

In Prague, October 25 2016

Roman Neruda,
Institute of Computer Science
Czech Academy of Sciences