Master's Thesis Review

Author: Franky
Title: Methods for Creating Subjectivity Lexicon for Indonesian
Supervisor: RNDr. Ondřej Bojar, Ph.D.

The thesis submitted by Franky aims at creating a lexicon of evaluative (positive and negative) Indonesian expressions, the so-called subjectivity lexicon. Such lexicons are available for several languages and serve in sentiment analysis, i.e. automatic estimate of whether a text describes an entity in a positive or a negative way.

Franky applied four means of automatic translation to several subjectivity lexicons available for English (one machine translation system was of his own). The resulting set of possible lexicons in Indonesian was evaluated in a task of sentiment analysis for a small set of sentences randomly selected from Indonesian reviews. Additinally, a machine learning method was wrapped around the baseline method, to allow for the inclusion of features apparent in the source sentence. The lexicon-based methods of sentiment analysis were compared with a data-driven benchmark, where two thirds of the test set were taken out and used as the training data to extract evaluative expressions from. An iterative approach to refine these seed data-driven lexicons using unannotated data was also attempted.

While the test set that Franky had to manually construct may be too small to assess the performance of the methods with a high confidence, the main results are indisputable: data-driven (and thus domain-specific) methods for estimating the subjectivity of a sentence are much more precise than lexicon-based techniques. Better recalls are achieved with the generic translated lexicons. The machine learning approach stays on the level of the baseline and it would deserve a more thorough search for useful features, but it is important to mention that it was not required in the assignment. If possible, I would like to see the results of the machine learning method for NonLexFeats (Figure 4.6) adding one feature at a time; it could shed some light on what mainly contributes to the baseline.

The main aim of the thesis was clearly achieved: Franky's thesis provides several subjectivity lexicons for Indonesian, with various trade-offs of precision and recall. The thesis is well written. The structure naturally leads us from overview of the task, through available data to the examined methods. The experiments are documented in sufficient detail and the results are presented using many charts and carefully discussed. The thesis contains only a small number of grammatical or typesetting errors. Overall, the language of the thesis is very good and the reader can spot somewhat disfluent sentences only rarely.

To conclude, the thesis documents that Franky was able to explore the topic of sentiment analysis and subjectivity lexicons at a sufficient depth, bring in his own suggestions, empirically test them and clearly describe and discuss the results. I suggest the thesis to be accepted as a M.Sc thesis at Charles University in Prague.

Prague, August 19, 2013.

RNDr. Ondřej Bojar, Ph.D.
Charles University in Prague, ÚFAL