Review of the Master Thesis

Author: **Maria Ximena Gutierrez Vasques**
Title: Quantifying Determiners from the Distributional Semantics View
Supervisor: doc. RNDr. Markéta Lopatková, Ph.D.

The goal of the thesis is to model logical relations like entailment using distributional semantic models. It focuses on noun and adjectival phrases, with the main stress on phrases with quantifiers like *all, some, many* or *both*, and entailment like *many dogs* |= *some dogs*. The author repeats and further enriches the experiments described in (Baroni, Bernardi, Do and Shan, 2012). The entailment relation is modeled by an SVM classifier trained on semantic vectors representing pairs of quantifying phrases that are in an entailment relation. The main contribution presented in the work consists in the development of models with less complex polynomial kernels and optimized parameters.
The thesis consists of five chapters; it includes a short appendix describing data and technology used, a sufficient list of references and a list of tables. The attached CD contains the text of the thesis, scripts performing the experiments, and a list of words extracted from the corpus that form the semantic space.

After a short introduction (chapter 1), which describes the goal and structure of the thesis, the second chapter serves as an introduction to the approach and methods adopted. It describes the basic features of distributional semantic models, the principle of compositionality, the concept of entailment and general principles of machine learning (including evaluation methods). Then the author concentrates on the theoretical background of the Support Vector Machines, she also describes effects of a parameters estimation and of a selection of different types of kernels. She uses standard handbooks and articles, with correct citations (I have only some doubts about the origin of several figures, e.g. Fig. 2.1, 2,3, 2,4) and – up to my understanding – with no substantial shortcomings.
Chapter 3 deals with entailment in distributional semantics. The author describes in detail the experiments made by Baroni, Bernardi, Do and Shan (2012), namely detecting noun entailment (AN |= N and N_1 |= N_2) and detecting quantifier-noun sequence entailment (Q_1N |= Q_2N). This description serves as a basis for her own experiments described in chapter 4.
The core chapter 4 gives an overview of the experiments conducted by the author. She repeated the QN entailment experiments using SVN classifiers with less complex kernels (linear and quadratic, in comparison to the cubic ones from the original experiments) and she performed the optimization of the parameter C defining "costs" for misclassified training examples. She showed that the SVM classifiers with adjusted settings over-perform the original results described by Baroni, Bernardi, Do and Shan (2012). The results were analyzed and explained, confidence intervals were tested.
The last short chapter summarizes the results and mentions possible future work.

The thesis is clearly written, I have not detected any serious errors or other formal imperfections. Only a limited number of typos appear in the text, which do not disturb a reader or prevent an understanding of the text.

The thesis was co-supervised by dr. Raffaella Bernardi and successfully defended at the Free University of Bozen-Bolzano.

Conclusion

The reported thesis summarizes the approaches to the problem, repeats the older experiments, and suggests settings that over-perform the these experiments. It proves the author's ability to solve assigned tasks in the area of NLP and to clearly formulate her goals, used methods and analyze the results.

The thesis complies with the requirements for Master Thesis at MFF. I recommend to accept the thesis for the defense.

Prague, January 13, 2013

doc. RNDr. Markéta Lopatková, Ph.D.
Institute of Formal and Applied Linguistics
Charles University in Prague