

Review of the Master Thesis

Author: Dinh Lê Thành

Title: Question and Answer Classifier for Closed Domain Interactive Question Answering

Supervisor: RNDr. Markéta Lopatková, Ph.D.

The goal of the thesis is a question type classification of a given closed-domain user-collected data using machine learning classifiers. The thesis consists of seven chapters. It includes list of references and lists of tables, figures and abbreviations as well as an appendix dealing with installation and tools descriptions. The attached CD contains text of the thesis, a short user manual, used tools (parser, tagger), scripts written by the author and data used in the project.

The author focuses on two basic tasks, (i) question classification and (ii) identification of topic and follow-up questions. He has chosen two effective classifiers, Naïve Bayes Classifier and Decision Tree Classifier. The whole work is a part of a large student project of question-answering library system conducted at Free University of Bozen-Bolzano.

For the first task, question classification, the author has re-implemented the system used by (Li & Roth, 2002) on so called TREC data (<http://l2r.cs.uiuc.edu/~cogcomp/data.php>), widely used corpus for evaluating question answering systems. This system uses Naïve Bayes Classifier (implemented in Perl, using Laplace smoothing parameter equal to 10). The author clearly describes extracted features (morphological, syntactic and semantic information is used). The same setting was then used for classification of questions collected from the BOB library system developed at Free University of Bozen-Bolzano (<http://www.unibz.it/EN/LIBRARY/ABOUT/PROJECTS/bob-project.html>). As the taxonomy used for the open-domain TREC data is not suitable for the closed-domain BOB corpus, the author has proposed his own taxonomy suitable for library domain. The author then compares the results of his system on BOB data (with newly adapted taxonomy) with its performance on the TREC coarse-grained and fine-grained taxonomies.

The second task consists in classification whether a given question is a topic question (i.e., whether it introduces a new topic in a dialog) or whether it is so called follow-up question (i.e., a topic continuation question). For this work, a Decision Tree Classifier was chosen as in (Yang et al, 2006). The author re-implemented their system (using Weka tool). He used this system with the same setting on TREC data and then on BOB data. Similarly as for the first task, author describes extracted features and compares results on the TREC corpus and the BOB corpus.

The work shows that the author has a good insight into the problem of question answering systems; he provides a reader with an ample survey of existing work in this field. The work on the thesis gained him practical knowledge of the area of machine learning. He extensively works with existing tools and data (esp. TreeTagger, Stanford Parser, WordNet and packages for counting Similarity, TREC data). He also spent considerable time on preparation of BOB data, proposal of adequate taxonomy for BOB data as well as manual annotation to obtain training and testing data. I also appreciate that the reported work is a part of a large student project.

I have detected only minor errors, especially insufficient references or citations with formal flaws. Examples:


- The author refers to (Pinto et al, 2002) in the text, p. 16, but this reference does not appear in the list of references;

- TreeTagger: the reference (Schmid, 1995) appears in chapter VI. concerning topic and follow-up classification but not in chapter V. concerning question classification; it is not included in the list of references;
- In Appendix, the Stanford Parser is mentioned but I did not find any notice of this tool in the text;
- References are formally inconsistent, there are missing years (e.g., Sunghad)

Conclusion

The reported thesis proves the author's ability to solve independently and creatively assigned tasks in the area of NLP. The thesis is written in good English, all experiments are sufficiently documented and the results are discussed. In my opinion, it complies with the requirements for Master Thesis at MFF. I recommend to accept the thesis for the defense.

Prague, September 7, 2009



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