Diploma thesis review

Sergio Raúl Duarte Torres: Entity Retrieval on Wikipedia in the scope of gikiCLEF

In his thesis, Sergio Raúl Duarte Torres presents a system for entity search and ranking and its evaluation on the GikiCLEF 2009 data set. This shared task focuses on answering short questions that require "geographic reasoning, complex information extraction, and cross-lingual processing". In the presented system, the queries are expected to be formulated in Spanish and answers are searched in the Spanish version of Wikipedia. The system is built upon the well-known software tools, such as Lucene, Hibernate, and OpenNLP.

Chapter 1 of the thesis includes introduction and motivation for the work. In Chapter 2, the author presents quite a comprehensive and detailed overview of related literature. Chapter 3 describes the architecture of the system including methods for exploiting the external semantic resources. In Chapter 4, the results of the evaluation experiments are presented and compared with those obtained by the baseline system. The thesis is concluded and future work discussed in Chapter 5. The thesis is writen in English on 46 pages plus references and three appendices.

The work is well-structured. Both the system and the experiments are well-described. More complex explanations are provided with appropriate examples. The system evaluation is based on the test set of 50 questions and their manually assigned answers provided by GikiCLEF coordinators (in a form of a list of Wikipedia pages). The main contribution of the work is in the proposed techniques for query expansion that improved performance of the baseline system. The author also attempts to evaluate individual modules (preprocessing steps) independently and proposes techniques to overcome their insufficiency in this application (POS tagging and stemming). A detailed analysis of the results and comparison with other systems participated in the GikiCLEF 2009 is also provided.

The entire system described in this work has several parameters that must be set or optimized in some way. First, it is not clear how the IR engine score threshold (p. 35) is set. Second, the cut-off parameter k (p. 41) seems to be optimized on the test data which can be considered "cheating".

One of the preprocessing steps of the system is stemming. Is it really performed prior tagging? In this step, some information which can be potentially usefull for POS tagging is probably lost.

The author describes a method for query expansion by adding synonyms of the query words (obtained from Wordnet). This step, however, would require identification of the word senses (word-sense disambiguation) which is not a trivial step and it is not further described. How exactly is it performed? Is word-sense dismabiguation involved at all?

Conclusion

Sergio Raúl Duarte Torres has implemented quite a complex entity-retrieval system, performed a series of experiments and achieved interesting results in his work. I recommend his diploma thesis to be succesfully defended.

Praha, 7.9. 2009

Pavel Pecina, ÚFAL MFF UK