Diploma Thesis Review

Thesis title: Entity retrieval on Wikipedia in the scope of gikiCLEF
Author: Sergio Raúl Duarte Torres
Opponent: Ing. Zdeněk Žabokrtský, Ph.D.

Thesis description

The aim of the work under review is to implement and evaluate a system for generating short nominal answers to questions formulated in Spanish. More specifically, it is aimed at finding lists of named entities that could be considered as correct answers to geographically oriented questions contained in the Spanish part of the GikiCLEF 2009, which is a multilingual Information Retrieval evaluation task.

After the introductory chapter, an overview of selected related work is presented in Chapter 2: the author decided to focus on three systems that participated in a similar shared task in 2008. Chapter 3 describes the techniques he uses for the individual subtasks, outlines the system architecture and gives an overview of employed external resources (data and software tools). Chapter 4 summarizes the experiment evaluation and Chapter 5 concludes. The appendices contain Spanish and English version of GikiCLEF 2009 questions, and answers generated by the system. The thesis is written in English. The total number of pages is 54. There is no included CD-ROM, but I was provided with the created source codes readily by the author.

Remarks

The goal of the work is clearly stated. The methods used to reach the goal are reasonable, usually well motivated and often illustrated on Spanish examples translated to English. The author analyzes causes of errors at the individual substeps and proposes solutions for improving the results, such as postprocessing for the Spanish POS tagger or stemmer reimplementation. The system architecture is lucid and modular. The main information source is the Spanish version of Wikipeadia, but whenever possible the author tries to exploit also other existing data resources (DBPedia, WordNet, Yago) and software tools (Apache Lucene, Hibernate, OpenNLP). The system was applied on the GikiCLEF 2009 data. Performance results for several system configurations are shown; the presented performance gains acquired by the individual components (and by all of them in combination) over the baseline is in my opinion the most important contribution of the work.

I have an objection concerning the overall evaluation strategy. The author uses lists of correct answers extracted from the other GikiCLEF participants' submissions for evaluation purposes (p. 38). What I find wrong is that this evaluation data did influence the choice of the system parameters. First, it was the same data that confirmed that the best results are achieved when all three improvements (Yago, Query Expansion, and Category Expansion) are combined. Second, experimenting with this data led to the choice of the value of the truncating parameter k. Not only that this data was not available to the other task participants (and thus all promissing claims made about the systems' comparison become questionable), but it degrades the informativity about the best configuration performance even in isolation, as the evaluation data were in fact used also for "training". The difference is substantial, because as it follows from Figure 9, the performance decreases quicly if k differs from the value optimal for the given data. The main indicator, GikiCLEF score might have dropped to less than one half if k was not derived from the evaluation data.

In the thesis conclusion, the three methods used for query expansion are denoted as novel. However, one can find a number of published articles that use WordNet for query expansion (e.g. Buscaldi et al, 2005: A WordNet-based Query Expansion method for Geographical

Information Retrieval), as well as articles about using Wikipedia for translation. In my opinion, using lexical networks is rather the standard approach for query expansion.

What I missed in the thesis was some more technical documentation. Unfortunately I found no instructions for running the whole system, and no technical details such as speed of the whole experiment or memory requirements in the thesis.

One can find several subtle flaws in the text, such as forgotten underlining in Figure 3, typo "hyyponymy" (p. 22), missing spaces (e.g. in the caption of Figure 2, p. 27, p. 28, p. 36), the same equation appearing twice (p. 10 and p. 40), an error in token numbering in Example 2, wrong agreement "the algorithm expand ...", "name entities" instead of "named entities", "feature work" instead of "future work, "in greater detailed" instead of "in greater detail". But they have no impact on text understanding.

I have a question concerning splitting input topics (questions) into the NE type part and the restriction part. In the system, this division is guided by a simplified syntactic analysis of the topic, in which the type part is supposed to be the first noun phrase. But each natural language gives certain freedom to users formulating their questions, e.g. as in "Find Swiss citizens who are Olympic medalists" vs. "Find Olympic medalist who are Swiss citizens". Could it be useful to identify such questions in which the NE type can be alternatively derived from the second NP? Would it be possible?

Conclusion

Sergio Raúl Duarte Torres has shown that he is able to build a working software system that implements the relatively complex task of entity retrieval. He has sufficiently documented his experiments in his diploma thesis. I recommend to accept the thesis for the defense.

In Prague, August 26, 2009

Zdeněk Žabokrtský
Institute of Formal and Applied Linguistics
Charles University in Prague