Supervisor report on PhD thesis

Jan Dedek

Semantic Annotations

My original annotation for this direction of research (before any student went this way) was following: "Semantic annotation of data from web-resources is one of the most important and most problematic steps on the way of realization of the semantic web. The semantically annotated data can be automatically integrated across different sources and such data are ready for the process of inference of additional knowledge from the data. This work is supposed to improve existing semantic annotation technologies. The research will be concentrated on technologies of automatization of the semantic annotation process. Different methods of the automatization will be concerned - linguistic methods, HTML wrapping, graphical methods (OCR), machine learning methods, rule-based systems, statistical and probabilistic methods. This work will also deal with textual data in Czech and with the possibility of employment of the Czech school of linguistics (ÚFAL Praha and NLPlab Brno) in the process of the semantic annotation. The exiting linguistic tools for Czech will be compared with the other tools (for other world languages), which were already used in the process of information extraction and semantic annotation. The secondary goal of this work is integration of the semantic annotation process with other (semantic) web technologies like: web-crawling, web content mining, RDF, OWL, semantic indexing and retrieval, ontology mapping, ontology mining, user profile mining, etc. Success in this goal will bring us a small model of semantic web. This model could be used for simulation and development of the semantic technologies and the development will be much more realistic than it could be in present times."

I have to admit, that such a problem setting is too wide for one PhD student to solve in 4-5 years. So, after Jan Dedek has shown interest in these problems, goals were narrowed during the time, as our experience on complexity of the problem increased (in fact, the whole problem is rather a problem for a team for several years).

I have also to emphasize, that the study program (in which this thesis is submitted) are Software Systems, in this case software systems for automated processing of the web content.

Applicant tried to extract the core and most difficult problem (I agree the problem has to be attacked right here from the very beginning). From the point of view of Software Systems, the use case of Semantic Web, as described by Tim Berners-Lee, points to automated processing of web content in textual form without assistance of creator (so, a third party annotation of textual web content). Most important part of such specified semantic annotation is information extraction. As already anticipated in the beginning, the use linguistic deep parsing as a support structure for this had to be tested. From the point of view of Software Systems most important was to integrate all this – PDT tools and the GATE system were selected. I agreed that goals will be specified this way.

As a result of this PhD work, new publicly available implementation of new methods was created. I value most the integration of the extraction method based on ILP into GATE framework. Linguistic tools of

Prague linguistic school were integrated into GATE framework too. This enables the use of all machine learning algorithm inside the framework and makes all experiments repeatable and comparable. Moreover, the implementation of the Fuzzy ILP Classifier is fully compatible with Weka. From Software Systems point of view these were nontrivial integration of big software packages with different philosophy, methodology and languages (implementation is backed by original modeling solution).

Large amount of new domain knowledge (especially from linguistic) had to be managed; Jan Dedek mastered many of them.

Thesis presents several novel contributions: to mention at least (by my opinion) most important, it is the first attempt to use deep linguistic parsing for information extraction. Both PDT - Prague Dependency Treebank resources and Stanford parser were used. This demonstrates language independence of this software solution. Author's concept of shareable extraction ontologies enabled the usage of a semantic web reasoner as the interpreter of extraction ontology. Information extracted from documents was used for document classification. All these show fulfillment of the task and create an original scientific contribution (in several conference publications and one journal with impact factor). Main advantages of these results are

- Nontrivial implementation and software integration enables repeatability and comparability of experiments
- Use of FILP gives extraction rules understandable by humans (in contrast to PAUM black box results)
- Results are numerically comparable to best results in literature (which are in most cases not repeatable and hence not verifiable), in few cases even better

Results exceeded my expectations, all goal were fulfilled. Written form of thesis could have been polished more. Material on problem definition and related work are not concentrated into one single paragraph and practical spread through the whole work (author quotes and really uses over 90 references; I personally like it more this way).

I fully support admittance to defense.

Prague July26th, 2012

Prof. Peter Vojtas
Supervisor