Review of the Doctoral Thesis "Semantic Annotations" by Mr. Jan Dedek

Summary

This thesis essentially describes the use of information extraction techniques for analysing Czech text, automatically inducing extraction rules and developing independent extraction ontologies from these rules. The main foundation of the information extraction technique is the use of deep language parsing. Additionally, the extracted information is combined with Fuzzy ILP techniques for some experimental document classification. I have a slight issue with the title of the work, in that I don't think it accurately reflects the contents. This thesis is not really about semantic annotation, but about the use of deep language parsing for the generation of extraction rules.

While information extraction and document classification have been extensively studied, and many techniques proposed, there are a number of novel aspects to the methodology used in this work:
1. in the past, the use of deep language parsing has not typically been associated with information extraction, which usually relies on much shallower techniques;
2. the automatic induction of extraction rules has been little studied until now;
3. new fuzzy ILP techniques have been developed for document classification;
4. Czech is a language which has not been studied nearly as extensively as many other languages, such as English, for the purposes of information extraction and document classification.

These new techniques are all interesting. In particular, the use of deep language parsing is of interest because typically shallower techniques are used, such as POS tagging and chunking. This is partly because deep analysis techniques can be very slow and may not work well except on very specific kinds of formal text (similar to that on which they have been trained), and partly because such tools are not available for many languages. The authors make use of an existing tool for parsing Czech. The choice of Czech as the language is also timely, as traditionally, research on Information Extraction and other NLP applications has focused on more mainstream languages; however, as an official EU language, amongst other reasons, there is particular interest in tools for its analysis. The development of Information Extraction tools for Czech is important as many other applications may depend on this: for example, question answering, sentiment analysis, information retrieval, ontology creation and so on. Adaptation of these tools and techniques to other neighbouring languages would also be an interesting extension.

In general, the thesis is well written and clearly presented and explained, but unfortunately the information presented is very sparse. In my opinion, there is not nearly sufficient detail of most of the work that has been done. For a PhD thesis, I would expect far more detail and discussion than is present. This lack of detail makes it difficult to understand what has really been done and the nature of the methods used, so there are many unanswered questions.

My other main concern is the quality of the results produced. The evaluations are detailed but unfortunately for the most part, do not appear to show any real progress in the state of the art -- what they show is that the methodology does not actually produce improved results on current methods, which is rather disappointing. While negative results are not necessarily bad as long as they prove a point, these do not really do so. The candidate should get credit for not trying to dissimilate the results and for showing the failings of the system, but on the other hand, there is not much to be learnt if we do not have any improvements on current techniques. The work really needs to be continued in order to find some way in which improvements can be made over existing techniques. It is difficult to therefore draw any conclusions about the work in the thesis: indeed, the concluding chapter is very short and simply makes the observation that other possibilities could be tried to improve the work, but this is all rather vague.

Detailed report

The introduction and motivation are clearly written, but require more detail - in particular the section on new ideas, models and methods is very brief and does not mention much in the way of novelty. I would expect to see more than 3 pages of text for an introductory chapter of a thesis.

Chapter 2 is again only 6 pages long, which is rather short, and needs to define the problem better. For example, information extraction is a well researched topic and the candidate needs to explain what the current issues are that other tools do not resolve, with a brief outline of how his work resolves these. The whole section needs much more detail, as it currently demonstrates only a very superficial understanding of the problem and task. The task of Information Extraction should be clearly defined, making clear what it involves, why it is needed, and why it is hard. For example, how does Information Extraction solve the Semantic Web bottleneck? What do you mean by complexity of information? The sections on IE components such as entity recognition, relation extraction, event extraction and so on are much too short -- I would expect these to be several pages each rather than several sentences. Similarly, the half page about the use of machine learning for IE is terribly sparse, as is the section on document classification. For a PhD thesis I would expect these issues to be covered and discussed in depth, with appropriate references (these issues have all been discussed many times in the literature), not merely mentioned in a few sentences.

The Related Work chapter is also terribly sparse - 6 pages is really not sufficient to show a detailed knowledge about the state of the art in this area. The candidate barely touches on the plethora of methods and tools currently available for information extraction and its subcomponents. For a piece of work entitled "Semantic Annotation", the related work section on semantic annotation should definitely be more than 5 sentences. The idea of extraction ontologies also needs to be explained early on, as it is not clear until much later in the thesis exactly what these are.

The chapter on Third Party Tools and Resources is generally good and well-written. However, I would like to see more information on why the particular tools used were chosen. For example, there are many different frameworks other than GATE which could have been used. Similarly, the discussion of Named Entity recognition tools and so on needs a lot more detail - for example, precisely which tools did you use? What components does ANNIE consist of? Which ones were used in your system?

In Chapter 5, I am missing a clear description of the set of steps taken (other than Figure 5.1, which is onot very detailed). It is also not clear how the various components fit together. The data model for extraction and repreentation of information also seems to be missing, plus details of the manual annotation performed for the machine learning phase, who did it, how it was done, what guidelines were used, and so on.

In Chapter 6, I would like to see more details about the various rules and rule types, such as examples of text matching each rule type. How was the set of rules determined? Which tool actually produces these rules, and how much of this effort was already existing (i.e. did you modify the tools or was this out of the box?). Section 6.2.4 also needs much more detail - how many documents were annotated? How does the number of documents annotated impact on the training? How was the manual annotation performed? I would expect to see some discussion on the effects of this.

In Chapter 7 also, I would like to see more details about the data -- who did the annotation? Was it single or double annotated? If single, why not double in order to remove bias? If double, did you measure IAA in order to assess the difficulty of the task (both for humans and as an upper limit)? What tool was used for annotating? Why? I would like to see examples of all the different

annotation types. How did you choose which types to annotate? Why did you not evaluate all of them? Some of them occur so infrequently that they are likely to skew results -- I would suggest using a larger dataset in this case.

A good number of evaluations and experiments have been carried out in this work, which is good to see. But again, the description of these is rather minimal, and there is very little discussion about the results. In particular, in view of the low quality of the results, some justifications and explanations should be made about this, with sugestions for improvement. From the results shown, it seems that the methodology is actually quite naive, which is rather problematic. The qualitative scores are worryingly low. The datasets used in the evaluation of the manual rules are rather too small to be statistically significant also. With the evaluation of the learned rules, it is clear that the method performs generally worse than existing methods such as PAUM -- there should be more discussion and some justification of why this might still be satisfactory as a result (currently, there is no real explanation given). The problem overall with the evaluation is that it does not show how there is any added value in this research to the state of the art -- if the candidate could find a way to demonstrate this then it would make the approach seem more believable. While disappointing results are not necessarily a sign of failure, there should at least be some glimmer of hope provided, but there is rather little of this to be found.

Finally the conclusion could do with further discussion and a continuation of the defence of this work in view of the rather low evaluation results. More details of how the work could be improved would go a long way towards this.

In summary, the work described is clearly presented and is well thought out and executed. However it is marred by the very sketchy information provided in the thesis itself. I feel sure that the candidate has plenty of background knowledge in the area, but it is not shown here. Similarly, the techniques could easily be described and motivated in much more detail. It is unfortunate that the results of the experiments are not better, but with better discussion and justification, this in itself would not be too problematic. It is currently a little hard to see what insights could be taken from this work given the results. My recommendaton is thus that while the majority of the work is sound and clearly shows creative scientific development, it requires some further development and a major rewrite with much more detail than is currently given. It would then be an excellent piece of research.

Dr Diana Maynard