Barbora Vidová Hladká – Habilitation Thesis

"Creating and Exploiting Annotated Corpora"

Report

The research activity of Barbora Vidová Hladká, described in her Habilitation Thesis "Creating and Exploiting Annotated Corpora", revolves around the creation and exploitation of annotated linguistic corpora, covering the acquisition of language data, their linguistic/theoretical analysis and annotation, and their use in Computational Linguistics applications. This is the background context of research focused on many research questions. What I consider the most significant and interesting characteristic of her work is in fact that she succeeds to cover and bring together many different corpus-related aspects, from more theoretical to more practical and applicative. This is something that not many researchers are able to do. But this is a very generic definition of her work. I'll mention only some of these different aspects in the following.

Globally, her research shows the impressive amount of methodologies, guidelines, theoretical frameworks, data and tools that are part of the Prague school, from the first pioneering years to the most novel methodologies of these days. In her Thesis she rightly provides an overview of the major Czech annotated corpora. This properly sets the context. Moreover, she always puts her research in the relevant historical context with respect to previous research. It is very clear that she succeeds in keeping always updated with new developments in the world of corpus annotation and use, a world that has undergone relevant transformations in the last decades.

She shows a very coherent career covering all the many steps related to Czech annotated corpora. She coordinated many projects related to Czech corpora, involving many other researchers in the projects.

Reading her thesis you feel her passion for what she was/is doing, which is an important characteristics of a good researcher. Moreover you feel that she is completely participating in the research atmosphere of the Prague school.

The main focus of her research is related to morphological and syntactic annotation. She participated to the Prague pioneering work in annotating corpora. She is always able to formulate the right research questions. Just an example is when she speaks about the Information Extraction and related parsing systems and their performance in relation to different annotation types. And how to exploit a parser trained on a newspaper corpus reusing it on a domain specific corpus and testing its performance focusing on the idea of splitting the parsing process. As a domain adaptation problem. Also this is not frequently done.

Not only she covered many aspects around corpus processing. She also gives practical examples of the importance of reusing existing language resources, adapting them to new formats to get more data for better research applications. Also interesting examples of reusing guidelines for different types of corpora. She also studied the possibility to reuse the extraction strategy from one domain to another.

Her research is devoted mainly to written corpora but she studies also problems related to spoken corpora.

She is also interested in novel usages and applications of annotated corpora, outside their original context, an aspect that many corpus researchers do not touch. Some tools are simple but used in real applications. She developed, with a colleague, a system for information extraction, used in different domains. She creates, analyses and uses also Legal corpora, mainly to evaluate the performance of an information extraction pipeline. Specifically she is interested mainly in the use of annotated corpora for computer-assisted learning. A very interesting and useful application is the development of the Corpus-based Exercise Book of Czech. It is an unusual application of annotated corpora.

She also used crowdsourcing procedures as an alternative way to increase the quantity of annotated data. And did this in a very thorough manner from the first design up to evaluation. An interesting part is the exploration of a different strategy for annotation, in the form of online games. She performed two different types of crowdsourcing: the more usual one in the form of games (different games for collecting different types of annotated data) and another, less usual, using students and tasks performed at school (at the advantage of both students and annotation!). Even if it seems that the accuracy in this case was not as good as the one of a parser.

The thesis is very well structured, with a good integration among the various perspectives. She is able to go through and describe systematically all the various aspects and perspectives of her research while retaining a global vision of the main objective and focus.

Finally, the chapter on future perspectives perfectly summarises both the major achievements so far (quite substantial and varied) and how to move on in the future. Among these I just mention the use of deep learning methods, the wonderful opportunity of use of NLP in the field of Digital Humanities, with a special interest in Optical Character Recognition, and Education as an interesting application area where the wonderful data and tools developed in Prague may find a really useful application.

She has a very rich and varied set of publications, very well cited, publishing in very respected journals, books and conferences.

In conclusion, my evaluation of the thesis is excellent.

I therefore highly recommend the Thesis to be accepted for Habilitation of Barbora Vidová Hladká.

Pisa, 4 January 2020

Nicoletta Calzolari Zamorani

[redacted]