

**Ing. Jaroslav Toningler**

Referent pro výzkum, vývoj a doktorské studium

Fakulta aplikovaných věd

Západočeská univerzita v Plzni

Univerzitní 22, 306 14 Plzeň

**Ph.D. Thesis Review**

## **Semantic Web Search using Natural Language**

**By Ing. Ivan Habernal**

**Review by ing. Josef Neumann, Ph.D.**

I have carefully examined Doctoral Thesis of Ing. Ivan Habernal that deals with the problem of Semantic Web Search using Natural Language (SWSNL). I can confirm that the "Semantic Web Search using Natural Language" is well written thesis that clearly elaborates on and contributes to this interesting research field mainly by delivering the following results:

- It provides both theoretical background and practical evaluation of innovative approach of the ontology-based question semantics independent of the ontology for storing the domain knowledge within SWSNL domain.
- It examines the statistical model for the semantic analysis based upon supervised training.
- It provides complete evaluation of the fully functional end-to-end system with a real Web data and real user queries.

From the formal perspective I found the thesis well structured and organized written in clear and concise language. I especially like the author's style of argumentation.

The breadth and variety of theoretical research and background the thesis builds on confirms both the obvious author's real interest in the research subject and his theoretical expertise. Author examined various research method and theoretical frameworks including many statistical methods for natural language understandings that are currently available summarizing both their strong parts and short comings.

Author presents interesting and innovative approach of using OWL based semantic annotations with layer of the domain independent sentence ontology and the layer of natural language questions that are

domain-dependent. From my opinion the real contribution of the thesis comes from the unique combination of well know theoretical approaches to NLU and the new innovative OWL based semantics that is incorporated within the proposed end-to-end SWSNL system.

The fact that the author not only demonstrated but also thoroughly evaluated his proposed approach to SWSNL on the real data from the „accomodation search domain“ further proves the potential and viability of the solution.

The evaluation approach on real data sets acquired using web mining methods and using facebook based pools to get the sample of real world questions illustrates the author’s ability to come up with interesting and innovative ideas and adapt them effectively in the overall research framework to meet the goals.

In my opinion the author proved the ability to perform scientific work and to achieve meaningful results and I do recommend the Thesis for the presentation with the aim of receiving the Ph.D. degree.

Nevertheless based on the research results presented I would like to pose the following questions that should be answered:

1. One of the most critical issues mentioned is the performance of the ontology reasoning (page 96).What kind of reasoning/ ontology storage engines have been evaluated? On page 85 it is mentioned that Sesame as the ontology storage engine could be used.  
*Sesame is more or less standard framework for processing RDF data. But what about NoSQL databases like MongoDB, CouchDB or graph based databases like Neo4J? Could not these be used as an effective and efficient alternative to ontology storage and reasoning with e.g knowledge base, full text search integration on document level?*

**As this is mentioned as one of the most critical issues for practical deployment from performance perspective, could the author outline the key requirement on the reasoning/ ontology storage engines and comment how the NoSQL aproaches could fit in or what are the potential drawbacks?**

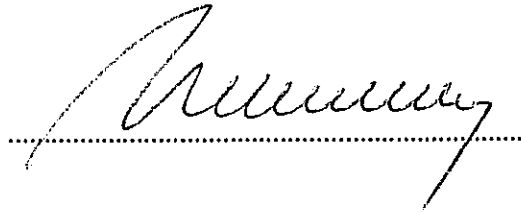
2. There are many different components mentioned in the thesis that obviously form the solid foundation of end-to-end SWSNL system.
  - a. On page 20 NLU is depicted as the key component of any system dealing with natural language.
  - b. On page 64 many different components for semantic analysis of natural language questions are introduced especially tokenizers, morphological tagger and named entity recognizers.
  - c. In chapter 7 the issues of semantic search and semantic interpretation are outlined together with the statement on the page 69 that the OWL semantic representation must be transformed into and ontology query language and that this transformation depends on the particular “back end” KB.

However there is no overall schema of the proposed architecture of SWSNL system with all key components and their relations. As per my understanding the end-to-end SWSNL system should consist of at least the following main layers:

- NLU and semantic representation of user queries
- Storage & reasoning engine for ontologies
- Knowledge base with desired domain information
- Transformation engine for user queries from ontology to particular information retrieval dialect
- Search Engine & Semantic Interpretation of retrieved results

**Could the author further clarify the proposed architecture?**

Praha, June 4, 2012

A handwritten signature in black ink, appearing to read 'Neumann', written over a horizontal dotted line.

Ing. Josef Neumann, Ph.D.

CEE ES Autonomy Lead  
HP Enterprise Services Central and Eastern Europe  
Hewlett - Packard s.r.o.  
1/1410 Vyskocilova | Prague 4 | 140 21 | Czech Republic



LEHRSTUHL FÜR  
MUSTER-  
ERKENNUNG

Lehrstuhl für Mustererkennung (Department Informatik)

Ing. Jaroslav Toninger  
Dekanat FAV ZCU Plzen  
Univerzitni 22  
CZ-30614 PLZEN  
Tschechische Republik



2.5-06-2012

FRIEDRICH-ALEXANDER  
UNIVERSITÄT  
ERLANGEN-NÜRNBERG  
TECHNISCHE FAKULTÄT

Computer Science Department 5  
Pattern Recognition Lab  
Informatik 5  
Lehrstuhl für Mustererkennung  
Prof. Dr.-Ing. Elmar Nöth  
Martensstraße 3, 91058 Erlangen  
Room 09.136  
Telephone +49 9131 85-27888  
Fax +49 9131 303811  
noeth@cs.fau.de

Your reference  
Your message from  
Our reference Haber01

Erlangen, June 4, 2012

Expert opinion on the doctoral thesis

### **"Semantic Web Search Using Natural Language"**

of Ing. Ivan Habernal.

The thesis deals with the subject of accessing the Semantic Web with Natural Language. This subject is of high interest since standard keyword search as used by established search engines proved to be an efficient interface to unstructured data in the net. However, with the rise of the Semantic Web, search in structured data – where keyword search is not effective - becomes more important and other interfaces to the content have to be provided. One such interface is Natural Language which allows especially the novice and casual user to access both structured and unstructured content in a unified way. The thesis of Mr. Habernal addresses this research topic and is thus an important contribution to the field.

The thesis consists of 9 chapters. After an introduction which introduces the research field, provides an overview over the thesis, and defines the terminology in a compact way, a literature overview is provided in chapter 2 and 3.

Chapter 2 describes the concept of the Semantic Web and ontologies and describes existing state-of-the-art Natural Language Interfaces to the Semantic Web. Also, the problem of missing standardized evaluation measures and training, development and test sets is discussed.

Chapter 3 deals more general with Natural Language Understanding (NLU) and Spoken Language Understanding (SLU) and gives a good and compact overview over existing

Hausanschrift  
Martensstraße 3  
91058 Erlangen

Telefon  
+49 9131 85-27883  
Telefax  
+49 9131 303811

Internet  
[www5.cs.fau.de](http://www5.cs.fau.de)

Bankverbindung  
Staatsbank Landshut  
Bayerische Landesbank München  
Konto 30 127 92 80 (BLZ 700 500 00)

approaches. Starting with the “classical”, grammar-based approaches, statistical approaches are introduced, which were mostly created in the SLU scenarios. After a discussion of evaluation measures, existing corpora are described. The chapter finishes with an overview over existing systems.

The rest of the chapters deals with the own research, taking the literature described in chapter 2&3 into account. Chapter 4 describes the domain of the implemented system. The design requirements lead to choosing the domain “Accommodation”, i.e. finding a hotel room, in the Czech language. A carefully designed small corpus of initial inquiries was collected via facebook. The example inquiries impressively show the need for careful design of a data collection in order to get serious inquiries. The construction of the knowledge base (an accommodation database downloaded from the internet) and the construction of the hand-annotation are described.

Chapter 5 describes the formalism that is used to annotate the natural language inquiries. The advantages of the formalism which is based on Semantic Web standards are discussed. The corpus of the inquiries is annotated with this formalism.

Chapter 6 proposes a statistical model for the analysis of natural language questions. The representation of a sentence as a sequence of semantic triplets is explained. Based on this formalism, the statistical model is introduced. The estimation of the model parameters is explained. Finally, several named entity recognition approaches are introduced.

Chapter 7 is a crucial chapter that puts the building blocks, described in chapter 4-6 together to a complete system. First the transformation of the semantic annotation into an ontology query language is described. The query is passed on to an ontology reasoner. The presentation of the search results to the user completes the end-to-end-system.

Chapter 8 presents the experimental results. First the Named Entity Recognizers are compared. Based on the experiments, a fusion of the modules is presented. The semantic analysis module is evaluated next. Using a perfect named entity recognizer an upper limit for the semantic annotation is given. It is shown that the size of the training set influences the results. Then the experiments are rerun with the results of the Named Entity Recognition. The results are significantly worse, but still acceptable. It also has to be taken into account that some questions can just not be answered due to the

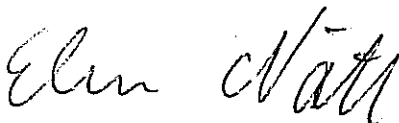
knowledge base. This is shown when an end-to-end performance with correct semantic annotation is tested. When using the semantic annotations of the system, it is shown that the system clearly falls back compared to perfect semantic annotation but significantly outperforms the baseline keyword search. This result is also demonstrated on two other corpora (the Czech ConnectionsCZ and the English ATIS corpus) and shows that – while a lot of work needs to be done – the proposed system provides significant improvement over the baseline.

Finally, chapter 9 summarizes and concludes the work. Open issues are discussed, directions of future are provided and the goals of the thesis are summarized and evaluated.

To sum up,

- the subject of the thesis is clearly relevant to current needs of the scientific community,
- the main objectives of the work have definitively been fulfilled,
- the methods used in this thesis have been very appropriate,
- the main results and contributions of the work are a carefully designed,
- the work is clearly important for the further development of science,
- the thesis satisfies with no doubt the conditions of a creative scientific work.

The author of the thesis proved to have the ability to perform research and to achieve scientific results. I do recommend the thesis for presentation with the aim of receiving the Degree of Ph.D.



Prof. Dr.-Ing. Elmar Nöth