Ciência sem Fronteiras at BIS

We created a **web page for applicants**: http://bis.informatik.uni-leipzig.de/csf

More offers will be added to the page around **April 22nd, 2013.**

Contact and application: **Sebastian Hellmann <hellmann@informatik.uni-leipzig.de>**

Research Group: http://aksw.org

This document contains four topics, that aim at the combination of Natural Language Processing, Semantic Web technologies and the Web of Data

Application

Good to excellent English skills beneficial.

No German language skills necessary (training will be provided in Germany).

Besides sending the usual CV and letter of motivation, please include any of the following, if available:

- your bachelor and master thesis
- a link to any open-source projects (including any mailing lists) you have been active on. (possibly via an aggregator service such as http://masterbranch.com/)
- a list of publications (either scientific or otherwise, e.g. software documentation, blog posts)
- your favorite UNIX one-line command (ideally related to RDF).

Supervisor

You will be supervised by our three-level system of supervision:

- operative level: Sebastian Hellmann - http://bis.informatik.uni-leipzig.de/SebastianHellmann
- tactical level: Dr. Sören Auer - http://www.informatik.uni-leipzig.de/~auer/
- strategical level: Prof. Klaus Peter Fähnrich - http://bis.informatik.uni-leipzig.de/en/KlausPeterFaehnrich

Topics

T1: DBpedia Português for Multilingual Natural Language Processing (NLP)

Pablo Mendes (http://pablomendes.wordpress.com) will be partially involved in the operative supervision of this thesis.

The topic aims to research the usefulness of data provided by DBpedia and other community-based data sets to extend the state of the art in NLP. The research will be in three directions:

1. Research on how to set up and integrate data and NLP tools for languages other than English. This task can focus at first on creating an NLP infrastructure around the Portuguese DBpedia (http://pt.dbpedia.org). Results should be more general, however, to be transferred to other DBpedia language editions: http://wiki.dbpedia.org/Internationalization, Journal of Web Semantics: http://svn.aksw.org/papers/2011/DBpedia_I18n/public.pdf
2. Collection (or creation) of NLP benchmarks with the aim to show which extracted data from DBpedia and LOD background knowledge can be employed to improve performance of exiting NLP tools.
3. Extend the current NIF (http://nlp2rdf.org) standard and integrate heterogeneous NLP web services to improve performance (F-measure)

The PhD student will become a member of the DBpedia Community project (http://dbpedia.org) and will optimize the infrastructure and the data output for NLP processes.

T2: A Multilingual Open WordNet as Crystallization Point for a Linguistic Linked Open Data Cloud

The explosion of information technology in the last two decades has led to a substantial growth in quantity, diversity and complexity of web-accessible linguistic data. These resources become even more useful when linked with each other, and the last few years have seen the emergence of numerous approaches in various disciplines concerned with linguistic resources.

It is the challenge of our time to store, interlink and exploit this wealth of data accumulated in more than half a century of computational linguistics of empirical, corpus-based study of language, and of computational lexicography in all its heterogeneity.

A crucial question involved here is the interoperability of the language resources, actively addressed by the community since the late 1980s, but still a problem that is partially solved at best. A closely related challenge is information integration, i.e., how heterogeneous information from different sources can be retrieved and combined in an efficient way.

With the rise of the Semantic Web, new representation formalisms and novel technologies have become available, and, independently from each other, researchers in different communities have recognized the potential of these developments with respect to the challenges posited by the heterogeneity and multitude of linguistic resources available today. Many of these approaches follow the Linked Data paradigm that postulates rules for the publication and representation of web resources. If (linguistic) resources are published in accordance with these rules, it is possible to follow links between existing resources to find other, related data and exploit network effects.

The PhD student will analyse, convert to RDF, interlink and fuse existing linguistic data with special emphasis on open WordNets (http://casta-net.jp/~kuribayashi/multi/) and Wiktionary

(http://dbpedia.org/Wiktionary)
The AKSW research group (http://aksw.org) has a leading role in the combination of Linguistic Data and Semantic Web technology. The PhD student will be responsible for the maintenance and development of the Linguistic Linked Open Data Cloud:

- http://linguistics.okfn.org/resources/llod
- http://ldl2012.lod2.eu/program/proceedings

The PhD student should have very good communication skills as well as an ability to improvise in an international open-source and open-community environment.

T3: A Methodology for Knowledge Extraction in the Enterprise Sector

A common problem in enterprises is the fragmentation of data sources due to a lack of Semantics and a common model for the available enterprise data. Recent Semantic Web research, however, has produced a plethora of tools for Knowledge Extraction. An overview and a classification is maintained by the AKSW research group (http://aksw.org) at this Wikipedia page:
http://en.wikipedia.org/wiki/Knowledge_extraction

The PhD student is to devise a methodology that aids enterprises in selecting applicable Knowledge Extraction tools based on their provided use cases. The task is complex as it requires an in-depth analyses of enterprise use cases and a unified workflow for semantic-based data integration in the enterprise data space.

The PhD will be required to learn and understand several programming languages for the consolidation and integration of the relevant software into a Knowledge Extraction tool suite to back up the developed methodology. Furthermore, the PhD student will acquire knowledge of databases, natural language processing and ontology learning.

T4: Creation of Binding of Textual content to Structured Knowledge

Pablo Mendes (http://pablomendes.wordpress.com) will be partially involved in the operative supervision of this thesis.

The topic exploits the “Conceptual Model for Semantic Enhancement ” by Pablo Mendes to advance the state of the art in Spotting (mention recognition, e.g. NER), Candidate Selection (detecting possible senses for a surface form), Disambiguation (choosing/ranking/classifying senses for a mention and Linking (deciding, if should annotate: to account for entities , not in the KB, or uninformative annotations)

Based on this initial conceptual model, the PhD student will explore in two directions:

1. Integrating more sophisticated methods into the existing DBpedia Spotlight system (http://spotlight.dbpedia.org/) . This includes higher-level annotations such as relations, discourse, sentiment.
2. Use-case driven selection of features for annotation. The PhD student will create and study several visualisation use cases that will aid users in exploring document collections based on structured knowledge. The goal is to find the right parameters to provide useful annotations for the respective user interfaces.

We are looking forward to welcome you to our research group in Germany.
All the best,
Sebastian