New Challenges for NLP Frameworks

May 22, 2010

ABSTRACTS

Workshop Organisers/Editors:

René Witte, Concordia University, Montréal, Canada
Hamish Cunningham, University of Sheffield, UK
Jon Patrick, University of Sydney, Australia
Elena Beisswanger, University of Jena, Germany
Ekaterina Buyko, University of Jena, Germany
Udo Hahn, University of Jena, Germany
Karin Verspoor, University of Colorado Denver, USA
Anni R. Coden, IBM T.J. Watson Research Center, USA

New Challenges For NLP Frameworks Programme

9:15 – 9:30 Welcome

9:30 – 10:30 Invited Talk: Jean-Marie Favre

10:30 – 11:00 Coffee break

11:00 – 13:00 Talks

Kalina Bontcheva, Hamish Cunningham, Ian Roberts and Valentin Tablan: *Web-based Collaborative Corpus Annotation: Requirements and a Framework Implementation*

Cartic Ramakrishnan, William A. Baumgartner Jr., Judith A. Blake, Gully APC Burns, K. Bretonnel Cohen, Harold Drabkin, Janan Eppig, Eduard Hovy, Chun-Nan Hsu, Lawrence E. Hunter, Tommy Ingulfsen, Hiroaki Onda, Sandeep Pokkunuri, Ellen Riloff, Christophe Roeder and Karin Verspoor: *Building the Scientific Knowledge Mine (SciKnowMine): a community-driven framework for text mining tools in direct service to biocuration*

Adam Funk and Kalina Bontcheva: *Effective Development with GATE and Reusable Code for Semantically Analysing Heterogeneous Documents*

Manuel Fiorelli, Maria Teresa Pazienza, Steve Petruzza, Armando Stellato and Andrea Turbati: *Computer-aided Ontology Development: an integrated environment*

13:00 – 14:30 Lunch break

14:30 – 15:30 Invited Talk: Michael Tanenblatt

15:30 – 16:30 Poster Session

Ralf Krestel, René Witte and Sabine Bergler: *Predicate-Argument EXtractor (PAX)*

Radim Řehůřek and Petr Sojka: *Software Framework for Topic Modelling with Large Corpora*

Ninus Khamis, Juergen Rilling and René Witte: *Generating an NLP Corpus from Java Source Code: The SSL Javadoc Doclet*

Nicolas Hernandez, Fabien Poulard, Matthieu Vernier, Jérôme Rocheteau: *Building a French-speaking community around UIMA, gathering research, education and industrial partners, mainly in Natural Language Processing and Speech Recognizing domains*

Elena Beisswanger and Udo Hahn*: JULIE Lab's UIMA Collection Reader for WIKIPEDIA*

16:00 – 16:30 Coffee Break

16:30 – 17:30 Panel Discussion: New Challenges for NLP Frameworks

17:30 – 17:45 Conclusions

Effective Development with GATE and Reusable Code for Semantically Analysing Heterogeneous Documents

Adam Funk and Kalina Bontcheva; Department of Computer Science, University of Sheffield, Regent Court, Sheffield, S1 4DP, UK

We present a practical problem that involves the analysis of a large dataset of heterogeneous documents obtained by crawling the web for unstructured and semi-structured human-readable documents (HTML, PDF) related to web services as well as their machine-readable WSDL files. The analysis uses natural language processing (NLP), information extraction (IE), some specialized techniques for WSDL analysis, and various approaches to classifying web services (defined by sets of documents). The results of the analysis are exported as RDF for use in the back-end of a portal that uses Web 2.0 and Semantic Web technology. Triples representing manual annotations made on the portal are also exported back to our application to evaluate parts of our analysis and for use as training data for machine learning (ML) to improve and evaluate the service classification. This application was implemented in the GATE framework and successfully incorporated into an integrated project, and included a number of components shared with our group's other projects.

Building the Scientific Knowledge Mine (SciKnowMine): a community-driven framework for text mining tools in direct service to biocuration

Cartic Ramakrishnan[3], William A. Baumgartner Jr.[1], Judith A. Blake[2], Gully APC Burns[3], K. Bretonnel Cohen[1], Harold Drabkin[2], Janan Eppig[2], Eduard Hovy[3], Chun-Nan Hsu[3], Lawrence E. Hunter[1], Tommy Ingulfsen[3], Hiroaki 'Rocky' Onda[2], Sandeep Pokkunuri[4], Ellen Riloff[4], Christophe Roeder[1], Karin Verspoor[1]; [1]University of Colorado Denver, PO Box 6511, MS 8303, Aurora, CO 80045, USA, [2]The Jackson Laboratory, 600 Main Street, Bar Harbor, Maine 04609 USA, [3]Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, CA 90292, USA, [4]University of Utah, 50 S. Central Campus Drive, Rm 3190 MEB, Salt Lake City, UT 84112-9205

Although there exist many high-performing text-mining tools to address literature biocuration, the challenge of delivering effective computational support for curation of large-scale biomedical databases is still unsolved. We describe a community-driven solution (the SciKnowMine Project) implemented using the Unstructured Information Management Architecture (UIMA) framework. This system's design is intended to provide knowledge engineering enhancement of pre-existing biocuration systems by providing a large-scale text-processing pipeline bringing together multiple Natural Language Processing (NLP) toolsets for use within well-defined biocuration tasks. By working closely with biocurators at the Mouse Genome Informatics (MGI) group (http://www.informatics.jax.org/) at The Jackson Laboratory in the context of their everyday work, we break down the biocuration workflow into components and isolate specific targeted elements to provide maximum impact. We envisage a system for classifying documents based on a series of increasingly specific classifiers, starting with very simple surface-level decision criteria and gradually introducing more sophisticated techniques. This classification pipeline will be applied to the task of identifying papers of interest to mouse genetics (primary MGI document triage), thus facilitating the input of documents into the MGI curation pipeline. We also describe other biocuration challenges (gene normalization) and how our NLP-framework based approach could be applied to them.

JULIE Lab’s UIMA Collection Reader for WIKIPEDIA

Elena Beisswanger and Udo Hahn; Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Fürstengraben 30, 07743 Jena, Germany

WIKIPEDIA, a huge, collaboratively built Web encyclopedia, is gaining increasing importance as a lexico-semantic resource for a large variety of natural language processing tasks. However, other than ‘well-defined’ and pre-formatted resources such as WORDNET, the ease of usability of its articles for text analytics is severely hampered due to underspecified document structure descriptions. To overcome this shortcoming, we here introduce a JAVA-based collection reader for WIKIPEDIA articles that is fully integrated in the Unstructured Information Management Architecture (UIMA). It imports articles from a WIKIPEDIA database, parses their raw text, composes a cleansed document text version and retains the original document structure in terms of UIMA annotations. We describe the structure and design of the WIKIPEDIA Reader and introduce the tools we incorporated, viz. UKP Lab’s JWPLDataMachine for setting up the database and the JWPL parser for parsing the wiki markup. In addition, we briefly introduce a scheduling system (in which the WIKIPEDIA Reader is integrated) that enables running several NLP pipelines in parallel, each with its own instance of the reader.

Web-based Collaborative Corpus Annotation: Requirements and a Framework Implementation

Kalina Bontcheva, Hamish Cunningham, Ian Roberts, Valentin Tablan; Natural Language Processing Group, Department of Computer Science, University of Sheffield, 211 Portobello, Sheffield S1 4DP, UK

Current tools demonstrate that text annotation projects can be approached successfully in a collaborative fashion. However, we believe that this can be improved further by providing a unified environment that provides a multi-role methodological framework to support the different phases and actors in the annotation process. The multi-role support is particularly important, as it enables the most efficient use of the skills of the different people and lowers overall annotation costs through having simple and efficient annotation web-based UIs for non-specialist annotators.In this paper we present Teamware, a novel web-based collaborative annotation environment which enables users to carry out complex corpus annotation projects, involving less skilled, cheaper annotators working remotely from within their web browsers.It hasbeen evaluated by us through the creation of several gold standard corpora, as well as through external evaluation in commercial annotation projects. Teamware is based on GATE: a widely used, scalable and robust open-source language processing framework.

Computer-aided Ontology Development: an integrated environment

Manuel Fiorelli, Maria Teresa Pazienza, Steve Petruzza, Armando Stellato, Andrea Turbati; ART Research Group, Dept. of Computer Science, Systems and Production (DISP) University of Rome, Tor Vergata, Via del Politecnico 1, 00133 Rome, Italy

In this paper we introduce CODA (Computer-aided Ontology Development Architecture), an Architecture and a Framework for semi-automatic development of ontologies through analysis of heterogeneous information sources. We have been motivated in its design by observing that several fields of research provided interesting contributions towards the objective of augmenting/enriching ontology content, but that they lack a common perspective and a systematic approach. While in the context of Natural Language Processing specific architectures and frameworks have been defined, time is not yet completely mature for systems able to reuse extracted information for ontology enrichment purposes: several examples do exist, though they do not comply with any model nor architecture. Objective of CODA is to acknowledge and improve existing frameworks to cover these gaps, by providing: a conceptual systematization of data extracted from unstructured information to enrich ontology content, an architecture defining the components which take part in such scenario, and a framework supporting all of the above. This paper provides an overview of the whole picture, and introduces UIMAST, an extension for the Knowledge Management and Acquisition Platform Semantic Turkey, that implements CODA principles by allowing reuse of components developed inside UIMA framework to drive semi-automatic Acquisition of Knowledge from Web Content.

Building a French-speaking community around UIMA, gathering research, education and industrial partners, mainly in Natural Language Processing and Speech Recognizing domains

Nicolas Hernandez, Fabien Poulard, Matthieu Vernier, Jérôme Rocheteau; LINA (CNRS - UMR 6241), University of Nantes, 2 rue de la Houssinière, B.P. 92208, 44322 NANTES Cedex 3, France

We report on the efforts the UN-LINA has made to build a UIMA French-speaking community both in Natural Language Processing and Speech Recognizing domains that would bring together researchers, industrials and educational interests. The intentions of building this community are twofold: to encourage the French-speaking academic and industrial organizations which have not yet adopt a middleware solution to use UIMA as a common development framework and middleware architecture for their research and engineering projects; to improve the collaborative development of common UIMA-based NLP tools and components for processing French.We present the services we set up as well as the resources we distribute freely under open licences to accomplish this objective. Most of them are currently available on the uima-fr.org Web Portal. They consist of: A web portal to discuss and exchange information about UIMA; A bundle of scripts and resources for automatically installing the whole of the Apache UIMA SDK; A bundle of UIMA-based components including some French NLP preprocessing components, a type mapper and a semantic rule-based analyser; A bundle of UIMA tools including an Analysis Engine Apache Maven archetype and an advanced web rest server; Course and training materials.

Generating an NLP Corpus from Java Source Code: The SSL Javadoc Doclet

Ninus Khamis, Juergen Rilling, and René Witte; Department of Computer Science and Software Engineering, Concordia University, Montréal, Canada

Program source code contains a large amount of natural language text, particularly in the form of comments, which makes it an emerging target of text analysis techniques. Due to the mix with program code, it is difficult to process source code comments directly within NLP frameworks such as GATE. Within this work, we present an effective means for generating a corpus using information found in source code and in-line documentation, by developing a custom Doclet for the Javadoc tool. Our SSLDoclet is able to generate a rich corpus using information found in source code and in-line documentation, including both sytactic and semantic information. The generated corpus encodes knowledge extracted from source code such as parent/child relations between classes and interfaces, methods and fields of classes, as well as return types and parameter lists and their in-line documentation, including author information provided through Javadoc. The developed XML schema is specifically designed to be easily processable by NLP applications, which allows language engineers to focus their efforts on text analysis tasks, like the automatic quality control of source code comments. The SSLDoclet is available as open source software from semanticsoftware.info.

Software Framework for Topic Modelling with Large Corpora

Radim Řehůřek and Petr Sojka; Natural Language Processing Laboratory, Masaryk University, Faculty of Informatics, Botanická 68a, Brno, Czech Republic

Large corpora are ubiquitous in today’s world and memory quickly becomes the limiting factor in practical applications of the Vector Space Model (VSM). In this paper, we identify a gap in existing implementations of many of the popular algorithms, which is their scalability and ease of use. We describe a Natural Language Processing software framework which is based on the idea of document streaming, i.e. processing corpora document after document, in a memory independent fashion. Within this framework, we implement several popular algorithms for topical inference, including Latent Semantic Analysis and Latent Dirichlet Allocation, in a way that makes them completely independent of the training corpus size. Particular emphasis is placed on straightforward and intuitive framework design, so that modifications and extensions of the methods and/or their application by interested practitioners are effortless. We demonstrate the usefulness of our approach on a real-world scenario of computing document similarities within an existing digital library DML-CZ.

Predicate-Argument EXtractor (PAX)

Ralf Krestel[1], René Witte,[2] and Sabine Bergler; [1]L3S Research Center, Leibniz Universität Hannover, Germany, [2]Department of Computer Science and Software Engineering, Concordia University, Montréal, Canada

In this paper, we describe the open source GATE component PAX for extracting predicate-argument structures (PASs). PASs are used in various contexts such as detecting textual entailment, question answering, automatic summarization, or knowledge representation to represent relations within a sentence structure. Different ``semantic" parsers extract relational information from sentences but there exists no common format to store this information. Our predicate-argument extractor component (PAX) takes the annotations generated by selected parsers and transforms the parsers' results to predicate-argument structures represented as triples (subjec t-verb-object). This allows downstream components in an analysis pipeline to process PAS triples independent of the deployed parser, as well as combine the results from several parsers within a single pipeline. Currently we support MiniPar, RASP, SUPPLE, and the Stanford Parser. In addition, we can extract PAS out of noun phrases making use of the output of a noun phrase chunker. We show the results using the different parsers on an exemplary news article.