COLING 2014

# The 25th International Conference on Computational Linguistics

# Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT

August 23rd, 2014
Dublin, Ireland

ISBN 978-1-873769-38-6
Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT)
Nancy Ide and Jens Grivolla (eds.)

# Preface

The Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technology (HLT) provides a forum for discussion of requirements for an envisaged open “global laboratory" for HLT research and development. Recent advances in digital storage and networking, coupled with the extension of HLT into ever broader areas and the persistence of difficulties in software portability, have led to an increased focus on development and deployment of web-based infrastructures that allow users to access tools and other resources and combine them to create novel solutions that can be efficiently composed, tuned, evaluated, disseminated and consumed. This in turn engenders collaborative development and deployment among individuals and teams across the globe. It also increases the need for robust, widely available evaluation methods and tools, means to achieve interoperability of software and data from diverse sources, means to handle licensing for limited access resources distributed over the web, and, perhaps crucially, the need to develop strategies for multi-site collaborative work.

For many decades, NLP has suffered from low software engineering standards, resulting in a limited degree of re-usability of code and interoperability of different modules within larger NLP systems. While this did not really hamper success in limited task areas, it caused serious problems for building complex integrated software systems, e.g., for information extraction or machine translation. This lack of integration has led to duplicated software development, work-arounds for programs written in different (versions of) programming languages, and ad-hoc tweaking of interfaces between modules developed at different sites.

In recent years, several efforts have been devoted to the development of frameworks to to allow the easy integration of varied tools through common type systems and standardized communication methods for components analyzing unstructured textual information. These include two frameworks, UIMA and GATE, which have been widely adopted within the HLT community to facilitate the creation of reusable HLT components that can be assembled to address different HLT tasks depending on their order, combination and configuration. At the same time, major projects in the US, EU, and Asia have recently undertaken the development of web service platforms for HLT, in order to exploit the growing number of web-based tools and services available for HLT-related tasks including corpus annotation, configuration and execution of tool pipelines, and evaluation of results and automatic parameter tuning. These platforms may also integrate modules and pipelines from existing frameworks such as UIMA and GATE, in order to achieve interoperability with a wide variety of modules from different sources.

Many of the issues and challenges surrounding these developments have been addressed individually in particular projects and workshops, but there are ramifications that cut across all of them. This workshop brings together participants representing the range of interests that comprise the comprehensive picture for community-driven, distributed, collaborative, web-based development and use for language processing software and resources, including developers of HLT infrastructures as well as those who will use these services and infrastructures, especially for multi-site collaborative work. The program includes presentations describing approaches to the range of challenges posed by such development, including legal issues concerning licensing of components and tools; the technical aspects of packaging and distribution of components; means to assemble complex processing pipelines and manage large bodies of data; means to visualize, explore, and further deploy analysis results; and issues surrounding the preservation of analysis results, provenance documentation, and evaluation and reproducibility. The overall goal is to work toward eliminating the fragmentation and duplication of effort that currently characterizes the field by establishing the basis of a community effort to develop and support the global laboratory for HLT development and research.

# Workshop Committee

**Program Co-Chairs**

Jens Grivolla, Universitat Pompeu Fabra (Spain)

Nancy Ide, Vassar College (USA)

**General Organizers**

Kalina Bontcheva, University of Sheffield (UK)

Christopher Cieri, Linguistic Data Consortium (USA)

Eric Nyberg, Carnegie Mellon University (USA)

James Pustejovsky, Brandeis University (USA)

Jonathan Wright, Linguistic Data Consortium (USA)

**Program Committee:**

Al Asswad, Mohammad, Cornell University
Nuria Bel, Universitat Pompeu Fabra
Steven Bethard, University of Alabama at Birmingham
Philipp Cimiano, Universität Bielefeld
Joan Codina, Universitat Pompeu Fabra
Kevin Cohen, University of Colorado School of Medicine
Azad Dehghan, University of Manchester
Leon Derczynski, University of Sheffield
Richard Eckart de Castilho, TU Darmstadt
Nicolai Erbs, TU Darmstadt
Thilo Götz, IBM Deutschland
Mark A. Greenwood, University of Sheffield
Nicolas Hernandez, University of Nantes
Yoshinobu Kano, Japan Science and Technology Agency
Peter Klügl, University of Würzburg
Marie-Jean Meurs, Concordia University
Yohei Murakama, Kyoto University
Kamel Nebhi, University of Geneva
Renaud Richardet, École Polytechnique Fédérale De Lausanne
Carlos Rodríguez-Penagos, Barcelona Media
Horacio Saggion, Universitat Pompeu Fabra
Bahar Sateli, Concordia University
Chunqi Shi, Brandeis University
Keith Suderman, Vassar College
Michael Tanenblatt, Thomas J. Watson Research Center
Martin Toepfer, Universität Würzburg
Katrin Tomanek, VigLink Inc.
Marc Verhagen, Brandeis University
Karin Verspoor, University of Melbourne
Graham Wilcock, University of Helsinki
René Witte, Concordia University
Torsten Zesch, University of Duisburg-Essen

# Table of Contents

# Conference Program

**Saturday, August 23, 2014**

8:45–9:00 Opening Remarks

9:00–9:30 *A broad-coverage collection of portable NLP components for building shareable analysis pipelines*
Richard Eckart de Castilho and Iryna Gurevych

9:30–10:00 *Integrating UIMA with Alveo, a human communication science virtual laboratory*
Dominique Estival, Steve Cassidy, Karin Verspoor, Andrew MacKinlay and Denis Burnham

10:00–10:30 *Towards Model Driven Architectures for Human Language Technologies*
Alessandro di Bari, Guido Vetere and Kateryna Tymoshenko

10:30–11:00 Coffee break

11:00–11:30 *The Language Application Grid Web Service Exchange Vocabulary*
Nancy Ide, James Pustejovsky, Keith Suderman and Marc Verhagen

11:30–12:00 *Significance of Bridging Real-world Documents and NLP Technologies*
Tadayoshi Hara, Goran Topic, Yusuke Miyao and Akiko Aizawa

12:00–12:30 *A Conceptual Framework of Online Natural Language Processing Pipeline Application*
Chunqi Shi, James Pustejovsky and Marc Verhagen

12:30–14:00 Lunch

14:00–14:30 *Command-line utilities for managing and exploring annotated corpora*
Joel Nothman, Tim Dawborn and James R. Curran

14:30–15:00 *SSF: A Common Representation Scheme for Language Analysis for Language Technology Infrastructure Development*
Akshar Bharati, Rajeev Sangal, Dipti Misra Sharma and Anil Kumar Singh

15:00–15:30 Coffee break

15:30–16:00 *Quo Vadis UIMA?*
Thilo Götz, Jörn Kottmann and Alexander Lang

**Saturday, August 23, 2014 (continued)**

16:00–16:30 *Integrated Tools for Query-driven Development of Light-weight Ontologies and Information Extraction Components*
Martin Toepfer, Georg Fette, Philip-Daniel Beck, Peter Kluegl and Frank Puppe

16:30–17:00 *Intellectual Property Rights Management with Web Service Grids*
Christopher Cieri and Denise DiPersio

17:00–17:30 *EUMSSI: a Platform for Multimodal Analysis and Recommendation using UIMA*
Jens Grivolla, Maite Melero, Toni Badia, Cosmin Cabulea, Yannick Estève, Eelco Herder, Jean-Marc Odobez, Susanne Preuß and Raúl Marín

17:30–18:00 Discussion : How to build a global community