Workshop on Open Infrastructures and Analysis Frameworks for HLT at COLING 2014, Dublin, Ireland

Submitted by rene on Fri, 2014-03-28 07:39

- coling
- Dublin
- HLT
- Ireland
- NLP
- workshop

Start: 2014-08-23
Timezone: Europe/Dublin
toc_collapse=0; Table of Contents

Workshop on Open Infrastructures and Analysis Frameworks for HLT

http://glicom.upf.edu/OIAF4HLT/

At the 25th International Conference on Computational Linguistics (COLING 2014)
Helix Conference Centre at Dublin City University (DCU)
23-29 August 2014

1. Description

Recent advances in digital storage and networking, coupled with the extension of human language technologies (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 causing 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 (such as implementing a parser), 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, two main frameworks, UIMA and GATE, have emerged that aim to allow the easy integration of varied tools through common type systems and standardized communication methods for components analysing unstructured textual information, such as natural language. Both frameworks offer a solid processing infrastructure that allows developers to concentrate on the implementation of the actual analytics components. An increasing number of members of the NLP community have adopted one of these frameworks as a platform for facilitating the creation of reusable NLP components that can be assembled to address different NLP tasks depending on their order, combination and configuration. Analysis frameworks also reduce the problem of reproducibility of NLP results by formalising solution composition and making language processing tools shareable.

Very recently, several efforts have been devoted to the development of web service platforms for NLP. These platforms exploit the growing number of web-based tools and services available for tasks related to HLT, including corpus annotation, configuration and execution of NLP pipelines, and evaluation of results and automatic parameter tuning. These platforms can 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. We therefore feel that this is the moment to bring 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. This includes those engaged in development of infrastructures for HLT as well as those who will use these services and infrastructures, especially for multi-site collaborative work.

2. Workshop Objectives

The overall goal of this workshop is to provide a forum for discussion of the requirements for an envisaged open â€œglobal laboratoryâ€• for HLT research and development and establish the basis of a community effort to develop and support it. To this end, the workshop will include both presentations addressing the issues and challenges of developing, deploying, and using the global laboratory for distributed and collaborative efforts and discussion that will identify next steps for moving forward, fostering community-wide awareness, and establishing and encouraging communication among the various players.

It aims at bringing together members of the NLP community specifically users, developers or providers of components and tools for these frameworks in order to explore and discuss the opportunities and challenges in using such platforms for modern, well-engineered NLP applications.

The challenge of creating reusable and interoperable components raises particular interest and are affected by legal issues, such as potentially incompatible licenses of components and tools as well as the technical aspects of packaging and distribution of components. Also, tools are important, for example to assemble complex processing pipelines, to manage the bodies of data that are to be analysed and to visualize, explore, and further deploy the analysis results. Further challenges are involved in embedding framework based analysis within applications or using it in distributed computing scenarios, such as deployment of and access to required resources. Finally, the preservation of analysis results, their provenance and reproducibility are of particular interest to the scientific user community.

3. Topics

Workshop topics include, but are not limited to:

- processing of very large data collections: scale-out, parallelization, and performance optimization
- advanced applications driven by an NLP framework
- sophisticated tools to build and manage complex processing pipelines
- analysis of results: exploration, evaluation, visualization, and statistical analysis
- experience reports combining components from different sources, as well as solutions to interoperability issues
- experience reports combining different frameworks (e.g. GATE/UIMA/WebLicht/etc.)
- UIMA components with a special focus on genericity and type-system independence

- repositories of ready-to-use components for UIMA and/or GATE
- distribution of components: documentation, licensing and packaging
- developing for UIMA or GATE: simplified APIs, debugging, unit testing, and limitations of the frameworks
- combining annotation type systems in processing frameworks (GATE, UIMA, etc.) with standardization efforts, such as done in the ISO TC37/SC4 or TEI contexts.
- use of NLP frameworks in real-world "industry" settings
- reports on current projects and frameworks, their challenges and proposed or implemented solutions, including efforts to address interoperability
- issues and challenges of multi-site collaborative projects, including reports of implemented or proposed strategies
- pipeline management, including authentication, strategies for passing resources through disparate tools and across hosting nodes, and licensing
- development and use of evaluation environments that facilitate assessment of HLT component performance, iterative application development, and replication of results
- community awareness and implementation of open infrastructures, including how to engage the community, establish confidence in the process, and promote use

4. Dates

Paper Submission Deadline: 2nd May 2014
Author Notification Deadline: 6th June 2014
Camera-Ready Paper Deadline: 27th June 2014
Workshop: 23rd August 2014

5. Organisers

Nancy Ide
Department of Computer Science, Vassar College

James Pustejovsky
Department of Computer Science, Brandeis University

Eric Nyberg
Language Technologies Institute, School of Computer Science, Carnegie Mellon University

Christopher Cieri
Linguistic Data Consortium, University of Pennsylvania

Jonathan Wright
Linguistic Data Consortium, University of Pennsylvania

Jens Grivolla
GLiCom, Universitat Pompeu Fabra

Kalina Bontcheva
Department of Computer Science, University of Sheffield

https://www.semanticsoftware.info/event/workshop-open-infrastructures-and-analysis-frameworks-hlt-coling-2014-dublin-ireland