LREC 2014 Tutorial

Multilingual Semantic Processing with BabelNet

Roberto Navigli and David Jurgens
Sapienza University of Rome

Motivation

The tremendous growth in multilingual text has significantly increased the need for multilingual resources in many research areas. However, many Artificial Intelligence and Natural Language Processing resources and algorithms have focused almost entirely on English. Multilingual lexical knowledge is indispensable for implementing the next step towards multilingual Natural Language Processing in which multilinguality is not a barrier, but an opportunity for identifying and leveraging linguistic phenomena common to multiple languages. Recently, new multilingual resources have emerged that provide a direct bridge from existing lexical and semantic resources such as WordNet to multilingual representations. In this tutorial, we highlight one such resource, BabelNet, to show how to transform existing monolingual semantic processing into the multilingual setting.

Summary

This half-day tutorial introduces concepts of Multilingual Semantic Processing and provides participants with an in-depth look at several case studies that incorporate multilinguality in their semantic tasks using BabelNet. The tutorial contains five sessions, designed to help participants familiarize themselves with the necessary concepts and extend their current monolingual semantic processing to the multilingual setting. Each session comes with hands-on examples that are done together, take-home exercises for those interested in working on more involved tasks, and a list of references to related research to further deepen their knowledge on a particular topic.

Outline

1. Foundations (.5h)

This introductory session will provide the necessary motivation and background in semantic processing needed to fully participate in the tutorial. Topics include word senses and ontologies, as well as practical issues such as data preparation and encoding.

2. Constructing a multilingual semantic resource (1.5h)

This session introduces techniques for building a common multilingual representation using existing knowledge bases and collaboratively constructed resources. We describe a case study from BabelNet which merges two common semantic resources, Wikipedia and WordNet, and also highlight recent efforts to merge other semantic resources such as OmegaWiki and Wiktionary. This session provides insight into issues such as aligning semantic representations across multiple languages and resources. We also describe the challenges of creating resources that cover large numbers of languages, such as BabelNet 2.0, which includes more than 50 languages.

3. Identifying multilingual concepts and entities in text (.5h)
Some multilingual texts may refer to an identical set of concepts and entities, which motivates linking them to a common, multilingual semantic representation. In this session, we introduce the challenges of linking ambiguous, multilingual text to specific concepts and entities. We formalize the problem as Word Sense Disambiguation and Entity Linking, and then provide a case study in how BabelNet can address both challenges.

4. Adding multilingual semantics to Information Extraction (.5h)
This session introduces techniques for adding multilingual semantics to Information Extraction so that the extracted relations connect concepts rather than text. We examine the example of WiSeNet, which builds a rich semantic graph on top of BabelNet by using Open Information Extraction techniques to harvest large numbers of semantic relations from text.

5. Annotating Multilingual Data through Gamification (.5h)
A main bottleneck for multilingual semantic processing has been a lack of annotated data for both training and evaluation. This session introduces the challenges of annotation with multilingual semantic representations and demonstrates ideas on how to design game-like tasks to improve annotation quality and quantity.

As time allows, we will finish with a group discussion for encouraging future collaborations and discussing open problems in the area.

Tutorial Speakers:

Roberto Navigli, associate professor in the Department of Computer Science at the Sapienza University of Rome. He is the recipient of an ERC Starting Grant in computer science and informatics on multilingual word sense disambiguation (2011-2016) and a co-PI of a Google Focused Research Award on Natural Language Understanding. His research interests lie in the field of Natural Language Processing.

David Jurgens, research scientist in the Department of Computer Science at the Sapienza University of Rome. His research interests focus on issues in Natural Language Processing with an emphasis on Semantics and Evaluation He has co-organized multiple research workshops and his research has recently been featured in the MIT Technology Review.