How to Add a New Language on the NLP Map: Building Resources and Tools for Languages with Scarce Resources

Rada Mihalcea
University of North Texas
rada@cs.unt.edu

Vivi Nastase
EML Research gGmbH
Vivi.Nastase@eml-r.villa-bosch.de

Abstract

Those of us whose mother tongue is not English or are curious about applications involving other languages, often find ourselves in the situation where the tools we require are not available. According to recent studies there are about 7200 different languages spoken worldwide – without including variations or dialects – out of which very few have automatic language processing tools and machine readable resources.

In this tutorial we will show how we can take advantage of lessons learned from frequently studied and used languages in NLP, and of the wealth of information and collaborative efforts mediated by the World Wide Web. We structure the presentation around two major themes: mono-lingual and cross-lingual approaches. Within the mono-lingual area, we show how to quickly assemble a corpus for statistical processing, how to obtain a semantic network using on-line resources – in particular Wikipedia – and how to obtain automatically annotated corpora for a variety of applications. The cross-lingual half of the tutorial shows how to build upon NLP methods and resources for other languages, and adapt them for a new language. We will review automatic construction of parallel corpora, projecting annotations from one side of the parallel corpus to the other, building language models, and finally we will look at how all these can come together in higher-end applications such as machine translation and cross-language information retrieval.

Biographies

Rada Mihalcea is an Assistant Professor of Computer Science at the University of North Texas. Her research interests are in lexical semantics, multilingual natural language processing, minimally supervised natural language learning, and graph-based algorithms for natural language processing. She serves on the editorial board of the Journal of Computational Linguistics, the Journal of Language Resources and Evaluations, the Journal of Natural Language Engineering, the Journal of Research in Language in Computation, and the recently established Journal of Interesting Negative Results in Natural Language Processing and Machine Learning.

Vivi Nastase is a post-doctoral fellow at EML Research gGmbH, Heidelberg, Germany. Her research interests are in lexical semantics, semantic relations, knowledge extraction, multi-document summarization, graph-based algorithms for natural language processing, multilingual natural language processing. She is a co-founder of the Journal of Interesting Negative Results in Natural Language Processing and Machine Learning.