NLP Technologies for Automatic Analysis of Language Productions: A Proposal

Xiaobin Chen Björn Rudzewitz

February 14, 2017
Seminar für Sprachwissenschaft
Universität Tübingen

This is a full day workshop to introduce to language educators the Unstructured Information Management Framework (UIMA) for analyzing authentic and learner-produced texts. The UIMA framework is a software system that features highly-modularized, reusable, and scalable analysis of large volumes of unstructured information such as texts, audio and video. The audience will be introduced to the basic concept of the UIMA framework and learn how to incorporate Natural Language Processing (NLP) tools into UIMA for various tasks involving automatic text analysis, such as readability assessment, error detection and exercise generation, etc. They will also write their first analysis engine to analyze the syntactic complexity of learner productions.

Description of the Workshop

Natural Language Processing (NLP) technologies have been widely adopted for educational application development. In the field of language education, common tasks involving NLP include readability assessment, automatic essay scoring, error detection, exercise generation, short answer assessment, and complexity research, to name just a few. The proliferation of NLP tools and frameworks in recent years has greatly expanded the analysis toolkit for language educators and researchers, while at the same time becomes daunting to a lot of them, because the majority of them are not trained as programmers. However, analyzing large volumes of natural language data, be it authentic or learner-produced language, is inevitable for researchers in the educational domain. As a result, it is optimal if they could be introduced to technologies that help them incorporate NLP tools and streamline the analytic process of diverse language analysis needs.

This workshop aims at helping the participants to make sense of the general process involved in automatic analysis of language production by making use of a powerful

software system for unstructured information analysis—the Unstructured Information Management Framework (UIMA), which features modularization of analysis components and scalability of analysis capabilities. Furthermore, a large number of common analytical components have been implemented as UIMA analysis engines thanks to projects like OpenNLP, uimaFit, or DKPro. We will demonstrate how to make use of existing NLP tools implemented as UIMA components and self-implemented analysis components to construct analytical pipelines for doing complexity analysis of learner production. By the end of the workshop, the participants will have gained hands-on experience with the UIMA framework on real-life educational tasks and learned how to generalize what they will have learned for constructing customized analysis of language for various purposes.