

Using Natural Language Processing Tools in Educational Data Mining

Scott Crossley
Georgia State U.
Atlanta, GA 30303
scrossley@gsu.edu

Laura K Allen
Arizona State Univ.
Tempe, AZ, 85287
LauraKAllen@asu.edu

Danielle S. McNamara
Arizona State Univ.
Tempe, AZ, 85287
dsmcnama@asu.edu

ABSTRACT

The workshop will cover the development, use, and educational data mining applications of a number of freely available natural language processing (NLP) tools such as Coh-Metrix, the Writing Assessment Tool (WAT), the Simple NLP (SiNLP) tool, the Tool for the Automatic Analysis of Lexical Sophistication (TAALES), and the Tool for the Automatic Analysis of Cohesion (TAACO). The workshop will provide the participants with an overview of the types of linguistic features that can be measured with these NLP tools. Additionally, it will describe how these features have been (and can be) used in text analyses that are of importance to the educational data mining community. Participants will receive hands-on training with the tools using data from computerized learning environments. Participants will also be shown how the output from these tools can be used to develop machine learning algorithms that can aid in predicting educational outcomes.

Keywords

Natural language processing, Data mining, machine learning

1. Description of the Tutorial Content and Themes

In this workshop, participants will be introduced to a number of freely available natural language processing tools. The primary focus of the tutorial will be to familiarize participants with the linguistic constructs measured by the tools, including lexical sophistication, text cohesion, rhetorical style, syntactic

complexity, and text organization. The constructs that will be discussed have all been shown to correlate with student success in various educational settings. The remainder of the tutorial will focus on how these constructs can be applied in educational data mining research to predict educational outcomes. The basic outline for the tutorial is an introduction to linguistic constructs, an overview of available NLP tools, a description of how the linguistic constructs are calculated in the tools, and a discussion of the links between these linguistic constructs and educational outcomes (e.g., performance, attitudes, etc.).

1.1 Justification for Importance of Topic

Natural language processing tools have been widely used to better inform research in a number of fields including cognitive science, medical discourse, literary studies, language learning, and the social sciences. However, large-scale use of NLP tools in educational data mining is much less common. Recently, the strength of NLP has begun to be recognized, as evidenced by a number of funding opportunities, and research findings in fields related to EDM (e.g., MOOCs, ITSs, etc.). This suggests that NLP is poised to become an important element of educational research, particularly when used in combination with more traditional measurements of student success (i.e., psychometric data, system interaction data, and sensor data). The convergence of readily available NLP tools, large-scale educational data sets, and data mining techniques can provide EDM researchers with new approaches to better understand and predict variables related to educational success.