StandardConnection: Correlating Educational Resources in Digital Libraries to Content Standards

Stuart A. Sutton
The Information School
University of Washington
Box 352840
Seattle, WA 98195
+1 206 685 6618
sasutton@u.washington.edu

Elizabeth D. Liddy
Center for Natural Language Processing
School of Information Studies
Syracuse University
Syracuse, NY 13244
+1 315 443 4456
liddy@syr.edu

John Kendall
Mid-continent Research for Education and Learning
2550 South Parker Road, Suite 500
Aurora, CO 80014
+1 303 632 5527
jkendall@mcrel.org

Categories and Subject Descriptors:
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing---Linguistic processing
I.2.7 [Artificial Intelligence]: Natural Language Processing---Discourse, Language parsing and understanding, Text analysis

General Terms: Design, Experimentation

Keywords: Educational Resources, Content Standards, Metadata

The goal of our two year NSF National Science Digital Library-funded project is to develop Natural Language Processing technology that will automatically produce metadata values that correlate individual educational resources in digital libraries to content standards. The goal is to assign this metadata to the descriptive metadata records for resources in support of standards-based discovery and retrieval. The project will utilize the Achieve/McREL Compendix, a comprehensive knowledgebase of K-12 content standards derived from over 137 state, national and international content standards documents. The test collection of educational resources being analyzed is drawn from the more than 400 Web-based collections represented in the Gateway to Educational Materials catalog.

The significance of this project in terms of the Digital Library movement is that high-quality automatic correlation of educational resources to content standards is essential to meet the demands for searching and retrieving such resources based on those correlations. This demand will increase as the national focus on greater accountability in our K-12 institutions increases. While human correlations of resources to content standards characterize current practice, it is clear that the scale of the need for such correlations calls for sophisticated means for automatic mapping. This project is intended to provide an NLP-based solution to the problem.

Briefly, our NLP approach in this project is to analyze language utilizing all the levels through which humans extract meaning—morphological, lexical, syntactic, semantic, discourse, and pragmatic. The extent to which an individual technology includes these levels, particularly the higher-level ones determines the capability and sophistication of the resultant application. Having incorporated each of these levels into our baseline NLP document-processing module, we are extending the system's capabilities in this project to the task of learning the linguistic features that can be relied on to indicate what content standard an educational resource supports.

We are applying a sublanguage analysis framework to automatically identify clues that can be recognized in the mathematics and science educational materials to indicate to which standards the resources apply. Based on the discourse model, the system learns from recognizing these linguistic clues in the training set. The system will then be able to process new resources as they are added to the digital library and appropriately assign to the metadata for those resources the learning standards to which they are applicable.

This work is a continuation of our NSF NSDL project "Breaking the Metadata Generation Bottleneck" where we were successful in processing text to automatically assign metadata tags for the descriptive and subject aspects of educational resources.

JCDL'02, July 13-17, 2002, Portland, Oregon, USA.
ACM 1-58113-513-0/02/0007.