

MetaExtract: An NLP System to Automatically Assign Metadata

Ozgun Yilmazel, Christina M. Finneran, Elizabeth D. Liddy
Center for Natural Language Processing
School of Information Studies
Syracuse University
Syracuse NY 13444
+1 315.443.5484

{oyilmaz, cmfinner, liddy}@mailbox.syr.edu

ABSTRACT

We have developed MetaExtract, a system to automatically assign Dublin Core + GEM metadata using extraction techniques from our natural language processing research. MetaExtract is comprised of three distinct processes: eQuery and HTML-based Extraction modules and a Keyword Generator module. We conducted a Web-based survey to have users evaluate each metadata element's quality. Only two of the elements, Title and Keyword, were shown to be significantly different, with the manual quality slightly higher. The remaining elements for which we had enough data to test were shown not to be significantly different; they are: Description, Grade, Duration, Essential Resources, Pedagogy-Teaching Method, and Pedagogy-Group.

Categories and Subject Descriptors

H.3.7 [Digital Libraries]: Standards, System Issues, User Issues
I.2.7 [Artificial Intelligence]: Natural Language Processing

General Terms: Measurement, Design

1. INTRODUCTION

Given the amount of time and expertise that manually creating metadata for digital content takes, efforts have been underway to automatically extract metadata [1,2]. We have developed a system to extract information from the content of educational documents to use for the metadata. Our domain is lesson plans and Web-based educational activities in math and science for grades K-12. The goal is to extract appropriate terms and phrases from the digital documents to populate item-level metadata. This method is limited to information that is contained in the document. To account for information not presented in the text of the documents, our system uses a collection-specific configuration file that enables a collection holder to specify collection-level metadata elements. In Section 2, we describe the metadata extraction system. Section 3 presents our results from an evaluation of the metadata quality as judged by users.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Conference '04, Month 1-2, 2004, City, State, Country.
Copyright 2004 ACM 1-58113-000-0/00/0004...\$5.00.

2. METADATA GENERATION

Our MetaExtract system (Figure 1) compiles the output from three distinct extraction modules, along with information from the collection-level configuration file, to automatically assign metadata. Below we describe each of these modules: eQuery Extraction module, HTML-based Extraction module, and Keyword Generator module. Some elements are extracted from both the eQuery and HTML-based extraction modules so that we can have a higher chance of populating the metadata. For these elements, we have a prioritization process to indicate which of the extractions to assign to the particular metadata element. The results of all of the modules are gathered and output as a text file.

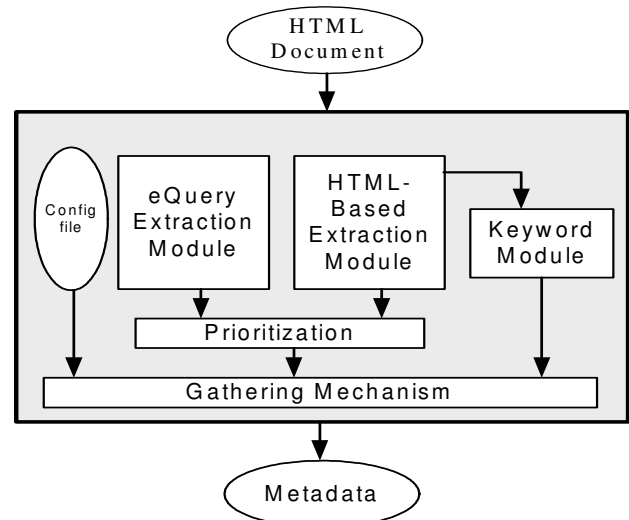


Figure 1. MetaExtract System

2.1 eQuery Extraction Module

Within the eQuery module, we use natural language processing (NLP) to extract terms and phrases found within single sentences. This module is a rule-based system that uses shallow parsing rules and multiple levels of NLP tagging [3]. We adapted the generic extraction phase of the module to extract terms and phrases to assign to the following metadata elements: Creator, Title, Date, Grade, Duration, Essential Resources, Pedagogy-Teaching Method, Pedagogy-Grouping, Pedagogy-Assessment, Pedagogy-Process, Audience, Standards, Publisher, and Relations.

2.2 HTML-based Extraction Module

The structure of the Web-based lesson plans and educational activities provide additional clues as to where the terms and phrases for some of the metadata elements can be found. The HTML-based Extraction module extracts over sentence boundaries. It uses the structure of the HTML and syntactic clues to determine where the contents of an element may be. Then it compares the text in that location to a list of clue words developed by analyzing hundreds of lesson plans. The clue words will determine which metadata element is present.

Through the HTML-based Extraction module we extract terms and phrases and assign them to the following elements: Title, Description, Grade, Duration, and Essential Resources.

2.3 Keywords Generator Module

The keywords for the educational documents are generated by computing the standard tf*idf metric on each document. Before running the documents through the keyword module, we use the HTML-based Extraction module to identify which sections of the documents to use for processing keywords.

3. EVALUATION

In a prior study, we reported on a preliminary evaluation of automatically assigned metadata as an entire unit [2]. We learned that we needed to have a finer evaluation to test the quality of **each** metadata element as produced by the MetaExtract system. Because the manual metadata cannot be considered a gold standard as some elements are omitted, we had users (science and math teachers) judge how well the metadata represented the lesson plan. We designed a Web-based questionnaire where users were given a lesson plan and its associated metadata, either manually or automatically assigned. For each lesson plan/metadata set, they were asked to indicate how well each of the metadata element entries represented the lesson plan selecting: *Very Poorly*, *Poorly*, *Well*, *Very Well*, or *Unsure*. They also were given a text box to write any comments and, if the element's contents poorly represented the lesson plan, they were specifically asked to explain why. We included this textbox as a quality control measure because in our previous work [2] we noticed that some teachers were evaluating the lesson plans themselves rather than the metadata.

To prevent drop-out, we wanted to keep the instrument as short as possible. From the Dublin Core and GEM metadata element set, we selected a subset of elements to evaluate, dropping elements that would not require user expertise such as Language and Date. We evaluated 10 metadata elements: Title, Description, Grade, Keyword, Duration, Essential Resources, Pedagogy-Method, Pedagogy-Process, Pedagogy-Assessment, and Pedagogy-Grouping.

Our test collection of 35 lesson plans came from The GEM Gateway [4], as these lesson plans have manually described metadata that matches our Dublin Core + GEM element set. We had 30 subject matter experts make a total of 1,300 judgments on the quality of the entries for the 10 metadata elements, with each lesson plan/metadata set being reviewed by 2 or 3 subjects.

3.1 Results & Discussion

As the judgment data is ordinal and was not normally distributed, we used the Mann-Whitney U-test to compare the difference of each element's medians for manually and automatically assigned

metadata. The U test compares the Mean Ranks of the automatic and manual metadata, as shown in Figure 2. Note that the Mean Rank is higher for the elements that typically are present for each document as their n is larger. Title, Description, and Keywords were populated for every document while the Pedagogy elements were present less frequently. Only two of the elements, Title and Keyword, were shown to be significantly different, with the manual quality slightly higher. The remaining elements for which we had enough data to test were shown not to be significantly different; they are: Description, Grade, Duration, Essential Resources, Pedagogy-Teaching Method, and Pedagogy-Group.

Given the amount of effort manual metadata assignment takes and the fact that most automatically assigned metadata elements are not statistically different in quality, we believe that MetaExtract offers tremendous promise in solving the metadata generation bottleneck.

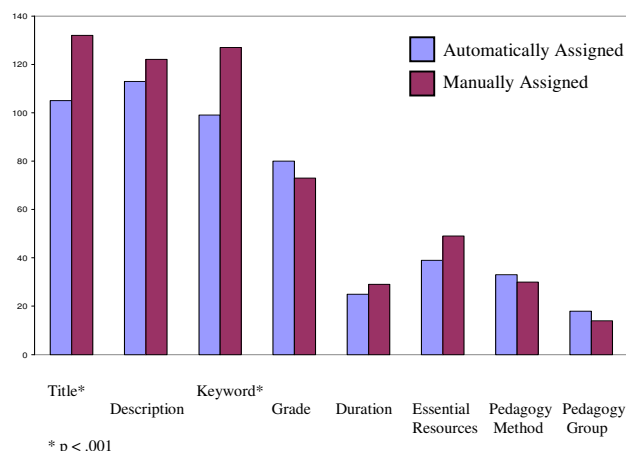


Figure 2. The Mean Ranks of Metadata Quality for Elements

4. FUTURE WORK

Work is currently underway on improving the extraction of the Title and Keywords elements. We will also be conducting an information retrieval evaluation to compare automatically assigned metadata to manually assigned metadata on their usefulness for retrieval.

5. ACKNOWLEDGMENTS

This project is funded by the National Science Foundation under the NSDL program (NSF Award Number: 0226312).

REFERENCES

- [1] Han, H., Giles, C.L., Manavoglu, E., Hongyuan, Z., Zhang, A., and Fox, E.A. Automatic document metadata extraction using support vector machines. In *Proceedings of the 2003 Joint Conference on digital libraries (JCDL 2003)* (Houston, TX, May 27-31, 2003), 37-48.
- [2] Liddy, E.D., Allen, E., Harwell, S., Corieri, S., Yilmazel, O., Ozgencil, N.E., Diekema, A. McCracken, N., Silverstein, J., and Sutton, S.A. Automatic metadata generation & evaluation. In *Proceedings of the 25th annual international ACM SIGIR Conference* (Tampere, Finland, August 11-15, 2002, 401-402.
- [3] Center for Natural Language Processing Technologies: eQuery. <http://cnlp.org/tech/equery.asp>
- [4] The Gateway to Educational Materials. <http://www.thegateway.org>