Standards Alignment for Metadata Assignment

Anne R. Diekema, Ozgur Yilmazel, Jennifer Bailey, Sarah C. Harwell, and Elizabeth D. Liddy
Center for Natural Language Processing
245 Hinds Hall
Syracuse, NY 13244
1 (315) 443 - 5484
{diekemar,oyilmaz,jabail01,scharwel,liddy}@syr.edu

ABSTRACT
The research in this paper describes a Machine Learning technique called hierarchical text categorization which is used to solve the problem of finding equivalents from among different state and national education standards. The approach is based on a set of manually aligned standards and utilizes the hierarchical structure present in the standards to achieve a more accurate result. Details of this approach and its evaluation are presented.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries - *standards*

General Terms: Algorithms, Measurement, Experimentation

Keywords: Automatic Metadata Assignment, Educational Standards, Hierarchical Text Classification, Natural Language Processing, NSDL, Machine Learning

1. INTRODUCTION
The National Science Digital Library (NSDL) provides resources for teaching and learning in the areas of science, technology, engineering, and math (STEM). STEM teachers have expressed the desire to search the materials in the digital library by the educational standards of their home states [1]. To satisfy this demand, state-specific standards information should be added to resources' metadata records. Previously, an effort was made by some of the participating libraries to add national standards as metadata. The current work aims to leverage prior standards assignments by automatically providing equivalent standards to the ones already assigned. This way, relevant standards from all states can be automatically added to the metadata, or to the query at search time Each way assures that the user can find relevant material for his or her subject and state, irrespective of what state standards have been assigned to the resource originally.

Finding equivalent standards is a challenge because, in some cases, dissimilar standards contain identical terms, whereas equivalent standards may not share any terms. To determine standard equivalency we use Natural Language Processing (NLP) in combination with a Machine Learning technique called hierarchical text classification.

JCDL'07, June 17–22, 2007, Vancouver, British Columbia, Canada.

2. STANDARDS
In the early 1980s, experts perceived a crisis in the American education system which encouraged the creation of national standards by professional subject-area organizations such as the National Council for Teachers of Mathematics. These national standards aimed to clearly define what students in certain grade levels are expected to know in core subject areas [4]. All states (except Iowa) followed suit and published their own educational standards, often using the national standards as a guideline.

The prolific production of standards has led to efforts by McREL to provide a synthesis of national standards documents in the Compendium of K-12 Standards [3]. Align to Achieve (A2A), no longer in operation, manually aligned some of the state and national standards to McREL's Compendium to produce a database of K-12 standards.

It needs to be pointed out that both state and national standards are typically organized in hierarchical fashion, starting at the apex, with the general subject (Math), the areas within that subject next lower (e.g. Number and Operations Standard), the third level standard (understand meanings of operations and how they relate to one another) next lower, and finally the benchmark (understand the effects of adding and subtracting whole numbers) as the leaf level.

The research in this paper describes a machine learning approach to the standard to standard alignment problem at the benchmark, or leaf, level. Using the A2A database as input to our algorithm, our approach utilizes the hierarchical structure present in the standards to get a more accurate result. Details of this approach and an evaluation are presented in the sections below.

3. TEXT CATEGORIZATION

3.1 Hierarchical Text Categorization
Text Categorization is the task of assigning predefined labels to textual documents. We used the LibSVM algorithm from the MLToolkit [7], as Support Vector Machine algorithms have proven successful in our other classification tasks. Originally we used a one-vs-all classification at the benchmark level which created very small positive training sets and large negative training sets. For example, the hypothetical document set in Figure 1 contains 30,000 documents. 75% of those documents (22,500) are used for training. Out of these only 30 are positive examples, leaving 22,470 negative training examples for that benchmark.

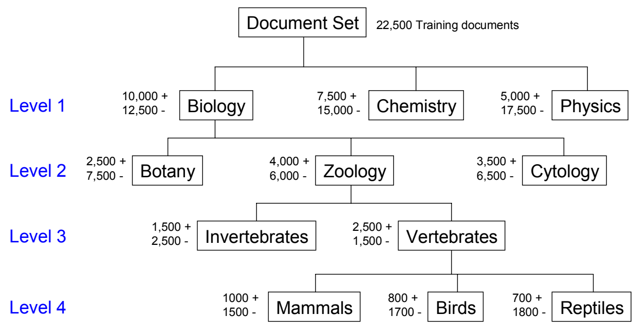

Figure 1. Utilizing the hierarchical structure in reducing negative training examples.

Using the hierarchical nature of the standards effectively reduces the size of the negative training set for each category [2,5]. For example, in the same training set above, at Level 1: 10,000 documents are identified as members of the Biology category – these are the positive training documents. The remaining 12,500 documents in the document set do not belong to Biology and are therefore negative training documents. At Level 2: there are 4,000 positive training documents for Zoology. The remaining 6,000 documents in the Biology document set do not belong to Zoology and are therefore negative training documents, and so forth until the last node has been reached. At each descending level, the pool of training documents from which a classifier is built is smaller and better defined.

3.2 Building the classifiers

In the hierarchical categorization system, we build binary categorizers for each node. A new document will be run through all the categorizers at level 1 and follow the children of nodes which produced a positive result. The document follows the children of positive nodes until it reaches a benchmark (leaf) node. All the labels that were positive at the benchmark level are assigned to the document.

4. EXPERIMENTS

To evaluate our approach we used the A2A database entries of manually aligned state standards for Science and Math. The dataset was divided 80/20 for training and testing purposes. The split was done over state standards, so that we used 27 state standards for training our system, and 20 state standards for testing. To enable us to do cross validations with the training set, we also restricted the experiments to the benchmarks with at least 30 training and 10 testing examples, which left us 354 benchmarks (leaf nodes).

Each benchmark is later processed with natural language processing-based information extraction software, and the resultant NLP tagging, as well as the original benchmark text, is used to create document representation vectors for each benchmark. We have experimented with various document representations, such as using part of speech tag information to filter out terms or using phrases instead of individual tokens etc. The experiments reported on in this poster use the representation where every term from the benchmark was stemmed and numeric concepts and stop words were filtered.

We trained a classifier for each node in the hierarchy as described in the previous section. Each test document was run through the hierarchical categorization system and any benchmark (leaf) node labels that were positive were assigned to the document. We have calculated the confusion matrix for each benchmark and calculated Precision, Recall and F-measure. Overall scores are reported with macro averages [6].

Table 1 Alignment Results for Math and Science Standards

	Math	Science
Precision	72.89	60.55
Recall	99.23	99.68
F-measure	78.22	66.26

As seen in Table 1, using hierarchical categorization to accomplish standards alignment produced good results in terms of both precision and recall on both Math and Science Standards. Also, the real world use of our alignment system is as a suggestion system. That is, for a given standard this system provides to the user a list of possible standards to which the incoming standard might align. In order to measure the success of our system, we evaluated our accuracy at 5 for the same test documents and achieved 99.42% and 99.61% for Math and Science respectively.

5. CONCLUSIONS

Automatic standards alignment achieves a better result using a relatively small set of training examples compared to one-vs-all categorization technique. We are currently planning experiments using NLP to detect semantic similarity in differently worded, but conceptually similar standards, and finding an automatic means to expand source benchmarks with additional vocabulary to aid in finding its equivalents. We expect these efforts will further improve precision.

6. ACKNOWLEDGMENTS

This paper is based upon work supported by the National Science Foundation under Grant No. 0435339.

7. REFERENCES

[1] Devaul, H. and Kelly, K. Searching by Educational Standards in DLESE: What does it mean and what do users want? Poster at NSDL Annual Mtg, Washington, D.C. 2003.

[2] Koller, D. and Sahami, M. Hierarchically classifying documents using very few words. 14th International Conference on Machine Learning, Nashville, TN, 1997.

[3] McREL. The McREL home page. Available: http://www.mcrel.org/topics/Standards/. Accessed 5 February 2007. Last updated: unknown.

[4] Ratvitch, D. National Standards in American Education: A Citizen's Guide. Brookings Institution Press, 1995, 242p.

[5] Ruiz, M. E. Combining machine learning and hierarchical structures for text categorization. PhD. Thesis. December 2001.

[6] Yang, Y. An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval, Vol. 1, 69-90, 1999.

[7] Yilmazel, O. Empirical Selection of NLP-Driven Document Representations for Text Categorization. Doctoral Dissertation. Syracuse University, 2006. 103p.