Text Categorization for Aligning Educational Standards

Ozgur Yilmazel, Niranjan Balasubramanian, Sarah C. Harwell, Jennifer Bailey, Anne R. Diekema, Elizabeth D. Liddy
Center for Natural Language Processing
School of Information Studies, Syracuse University
{oyilmaz, nbalasub, scharwel, jabail01, diekemar, liddy}@syr.edu

Abstract

Standard alignment (where standards describing similar concepts are correlated) is a necessary task in providing full access to educational resources. Manual alignment is time consuming and expensive. We propose an automatic alignment system, using machine learning techniques utilizing natural language processing. In this paper we discuss our experiments on text categorization for automatic alignment. We explore the role of relevant vocabulary sets in automatic alignment.

1. Introduction

Since passage of the *Leave No Child Behind Act of 2001*[1], there has been an increased emphasis on the design of K-12 curricula around existing and emergent state and national content standards and educational resources. Public educators now seek to demonstrate that classroom activities and curriculum materials will build the competencies embodied by the standards.

The National Science Digital Library (NSDL) is the Nation's online library for education and research in science, technology, engineering, and mathematics (http://www.nsdl.org). Funded by the National Science Foundation (NSF), the online library contains images, video, audio, animations, software, datasets, journal articles and lessons plans. In addition to the regular keyword searching on full-text content, NSDL also provides various types of metadata that are indexed for searches. A number of documents in the NSDL have been assigned national standards as metadata. However, focus groups with teachers have shown that the new generation of teachers prefer to search by state standards (versus national) for educational resources to aid in their teaching [1].Teachers stated that state standards were more relevant to their work than national standards because they are required to teach to state standards as well as document student progress in relation to those standards. However, to avoid assigning state standards for 50 different states, most lesson plan repositories, if they catalog using content standards at all, prefer to use the more general national content standards so as not to limit their work to a single state. Resources assigned with one state's standards are unable to be retrieved or utilized by teachers in other states. The development of an automatic alignment system would enable resources tagged with one state or national standard to be correlated with every other state or national standard.

The Center for Natural Language Processing (CNLP) is in the process of creating a technology for automatically aligning state standards and national standards. While other companies have manually aligned standards (e.g. Align to Achieve [A2A]), this task is onerous and time consuming, and must be constantly updated in order to remain current. Standard alignment is a complicated task—49 states have standards documents in the core subjects (math, English, science and social studies), which are revised every five to seven years [2]. In addition to the state standards, standards taxonomies have been developed by national organizations (e.g. National Council of Teachers of Mathematics and National Science Education Standards), curriculum committees, and local districts. Each set of standards utilizes discrete language, differing grade bands, distinct organizational structures and different levels of specificity in the coverage of a particular standard.

[1] H.R.1, Public Law No: 107-110, Section 1001 Statement of Purpose

1530-1605/07 $20.00 © 2007 IEEE

In the following example, Table 1 shows several benchmark level standards (a benchmark is the lowest and most specific section of the standard and is what the system will match)[2] that are considered equivalent:

Table 1: Equivalent Benchmarks

Source	Benchmark	Grade Band
Washington	recognize that the earth is a spherical planet with a mainly solid interior and a surface composed of landforms, bodies of water, and an atmosphere	Pre K-4
Maryland	SC.2.8.1 Identify different Earth materials and classify them by their physical properties	K-3
Arizona	PO 1. identify basic earth materials	K-K
New Mexico	2. Demonstrate that Earth's materials include solid rocks, soils, liquids, and gases such as those in the atmosphere.	K-4
National Science Educational Standard	Earth materials are solid rocks and soils, water, and the gases of the atmosphere. The varied materials have different physical and chemical properties, which make them useful in different ways, for example, as building materials, as sources of fuel, or for growing the plants we use as food. Earth materials provide many of the resources that humans use.	K-4
Compendix	Knows that Earth materials consist of solid rocks, soils, liquid water, and the gases of the atmosphere	ANY

[2] We utilize a standards' benchmark in the standard alignment process. However we use the term standard and benchmark interchangeably throughout the paper.

A human is able to distinguish the salient characteristics that make these benchmarks equivalent, but for a computer this task is more difficult. Each benchmark is a different length and uses vocabulary that may not be shared in some cases (basic earth materials vs. earth is...composed of landforms, bodies of water and an atmosphere), and overlaps in other cases. In addition, even benchmarks that don't correlate can use similar vocabulary when outlining different concepts. For example this benchmark from Louisiana, "Locate and compare the relative proportions of land and water found on Earth," should not align with the benchmarks above, although an information retrieval system might consider them equivalent based on the overlapping vocabulary.

To create an automatic mapping between state and various national standards, the tool that we are developing, the Standard Alignment Tool (SAT), uses a relatively small set of manually-determined alignments between benchmarks to learn classifiers for a crosswalk and uses these learned classifiers to align new standard to the crosswalk, thus allowing alignments between any state. This mapping from state standards to national standards can be incorporated into the search capabilities of educational resource repositories so that teachers can search for resources using either their home state standards or national standards. The assignment of content standards will be enhanced by the utilization of the crosswalk because a cataloger can assign standards using their state standards or a national set of standards, while being assured that the learning resource will still be searchable and 'findable' using other state standards through use of the crosswalk. As a result, educational resources from anywhere in the country can easily be shared once this translation between state standards is facilitated.

We have modeled the alignment problem as a multi-label text categorization problem. Using the Mid-Continent Research for Education and Learning (McRel) benchmarks [3], the A2A+McRel Compendix provides alignment of different state benchmarks. Each McRel benchmark is treated as a category in a multi-label text categorization task. The different state benchmarks that are aligned to a McRel benchmark are used as our training data. This research and preliminary results are discussed within the body of the paper.

2. Relevant Work

Alignment systems have been approached through two methods, manual and automatic. Manual alignment systems are created by standards experts and compare learning objectives between states, or between states and national standards. Starting in the mid 1980s organizations have used databases to make this process more efficient, [2]. Prior work has been done both in the for-profit and non-profit sector. In the profit sector, MediaSeek created an early prototype of an "intermediary-based" correlation system in which a set of statements describing the K-12 educational process becomes the core of a relational database, which serves as an intermediary to various sets of standards. Plato Learning Inc. [4] has built on this design to provide users with standards aligned to instructional resources and district curricula. Publishers of educational content have developed distinct taxonomies that they then treat as assets, limiting use to their clients. In the non-profit sector, Martha Cyr from Worcester Polytechnic Institute created a matrix for mappings between Massachusetts State mathematics and science standards to McRel benchmarks, National Council of Teachers of Mathematics standards (NCTM), the National Science Education Standards (NCES), and AAAS benchmarks. This matrix was then expanded to three additional states: Colorado, North Carolina and Oklahoma. This manually aligned set of state to national standards is being used in the TeachEngineering digital library (www.teachengineering.org). This project proved to be extremely time consuming, even after limiting the mapping to a single state, a single area (math) and subsets of grade levels. Align to Achieve [5] obtained funding to provide alignment services. They have manually aligned all 49 states using an intermediary based on McRel's standards, entitled the A2A+McRel Compendix. This was available via the Web and through licensing agreements, but has recently had cease doing business due to lack of funding. We are using this Compendix to train and evaluate SAT. At this writing, alignment services are hampered by the lack of uniformity in approach and various interpretations of the correct standard match for alignments (that is, coders are inconsistent in interpreting a correct alignment).

At present only one other system attempts to use natural language processing to correlate standards. AlignPro, a product of SmartPro [6], aligns state, national and district standards through natural language technology. It does not use an intermediary but aligns the standards based on descriptions of content and instructional objectives which are then used as the basis for a ranking of documents by concept. The system is sold to publishers and educational departments. CNLP originally received a grant from NSF to create a system for the National Science Digital Library to automatically assign standards to educational resources. The Content Assignment Tool (CAT) [7] streamlines the assignment of standards to educational resources by automatically retrieving and ranking standards for human evaluation. CAT analyzes the content of a resource through natural language processing and then employs search algorithms to make suggestions of relevant standards. Over time, machine learning improves CAT's accuracy. While CAT tags resources with individual state standards, SAT is designed to allow access to resources already tagged with a limited set of standards. Ultimately we hope that both programs will be used in conjunction to provide access to NSDL's large repository of educational resources.

3. Text Categorization for Standards Alignment

Text categorization assigns predefined labels to text documents. Automatic text categorization systems are utilized for different types of tasks, for example, electronic documents categorization, controlled vocabulary assignment, document filtering, metadata generation and word sense disambiguation [8]. Machine learning based text categorization systems achieve high levels of accuracy suitable for automated systems.

Automatic text categorization is treated as a supervised learning task. The goal of this task is to estimate a Boolean function to determine whether a given document belongs to the category or not by looking at pre-labeled examples. In applications where more than one category exists, a binary categorization system is created for each category label, and unseen documents are processed through each system. Using a one-versus-all approach, the system utilizes the category examples as positive instances and treats the combination of examples from all other categories as negative cases. An alternative approach to one-versus-all is to create a pair wise binary classifier for each

category label, which then leads to $C\binom{n}{2}$ binary classifiers.

The problem of aligning multiple state standards to the McRel standards has been approached by CNLP as a multi-label text categorization problem. The A2A+McRel Compendix contains manual alignments made by experts in the field of educational standards. Thus, we felt this "gold standard" would be appropriate for training a multi-label classifier. Given the expected size of the label set (in the hundreds) implementing a pair-wise multi-label categorization system would be computationally expensive; therefore a one-versus-all implementation was chosen. For our system, we determined that each McRel benchmark would be treated as a separate category. State standards aligned to the given McRel standard were used as positive examples and all other state standards were used as negative examples.

State standards are ordered hierarchically, in order to organize different levels and types of information embedded in the standard. Within one standard the hierarchy tree can carry as many as nine levels (and as few as one). The divisions can include grade level, topic, and a descending hierarchical view of a standard. In the following example, this particular standard has five levels of information:

- Level 1: Colorado
- Level 2: Science
- Level 3: 0-4
- Level 4: Standard 1: Students understand the processes of scientific investigation and design, conduct, communicate about, and evaluate such investigations.
- Level 5 (Benchmark): In grades K-4, what students know and are able to do includes communicating about investigations and explanations.

In addition, McRel has assigned a relevant vocabulary set to each benchmark. According to McRel "benchmark vocabulary includes terms and phrases that appear in the benchmarks, as well as terms and phrases that do not, but that capture ideas within benchmarks" [8]. In the example above, the relevant vocabulary set is: *scientific method, scientific investigation, explanation of data.* As discussed later, we have found that almost all the terms and phrases encompassed by the vocabulary set appear within the collection as a whole, but are not evenly distributed among equivalent standards.

When aligning standards we utilized three types of text content:

- Benchmark text
- Hierarchical text: the text of all the levels from the path to the root
- Relevant vocabulary assigned by McRel.

We processed the text through CNLP's TextTagger, a rule-based natural language information extraction engine, and extracted tokens, stemmed, performed part-of-speech assignment and bracketed phrases. We used some of this information for our experiments and we got mixed results for verb substitution and number substitution (experiments where certain parts of speech were substituted for the literal word or number). While these results were intriguing, they weren't stable enough to base our research on. We hope to explore some of the more sophisticated NLP options in further experiments.

We used the Machine Learning Toolkit (MLToolkit), a flexible framework to support text categorization experiments with various document representations. MLToolkit manages the flow of text categorization experiments from document representation to feature selection, categorization and analysis. It was originally implemented for Yilmazel's doctoral work and details can be found in [10].

The MLToolkit supports various classifiers, including Support Vector Machines (SVM) [11], Naïve Bayes, k-Nearest Neighbor and Decision Trees. For this work we have used the SVM implementation extended from LibSVM [12] in the MLToolkit. SVM performs better empirically on standard text categorization datasets and successfully handles large dimensional learning problems [13, 14]. Classifying new documents using SVM is a quicker process when compared to other machine learning algorithms [15]. Since we know that there are hundreds of McRel benchmarks and a corresponding number of categories, the two features of SVMs listed above make it a suitable algorithm for SAT.

The results of each binary classifier were evaluated by using the standard information retrieval metrics precision, recall and F-measure, as commonly done in the text categorization field. Most of the evaluation metrics are adapted from the Information Retrieval field and can be calculated from a confusion matrix as shown in Figure 1.

		Correct Class	
		C	Not C
Classifier output	C	TP	FP
	Not C	FN	TN

Figure 1. Confusion matrix.

A Confusion matrix provides counts of different outcomes from an evaluation system. True Positive (TP) represents the number of documents the system correctly labeled as positive, True Negative, represents the number of documents the system correctly labeled as negative. False Positive (FP) and False Negative (FN) are the number of documents the system incorrectly labeled positive or negative respectively. Text categorization tasks are unlike normal machine learning problems in two respects: examples can be given multiple category labels, requiring that separate binary classifiers be trained for each; and the positive examples of each class are usually in a very small minority. These two characteristics combined mean that a plain accuracy statistic is not adequate to evaluate performance. To deal with the unbalanced nature of the classes, precision and recall are used instead of accuracy. Precision is the proportion of examples labeled positive by the system that were truly positive.

$$\Pr ecision = \frac{TP}{TP + FP}$$

Recall is the proportion of truly positive examples labeled positive by the system.

$$\mathrm{Re}call = \frac{TP}{TP + FN}$$

$$F - measure = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \qquad \beta = 1 \qquad F_1 = \frac{2PR}{P + R}$$

β defines the weighting of Precision vs Recall. β=1 gives equal weight to precision and recall. Using f-measure will give us a single number to compare the performance of different systems. The results from individual binary classifiers are combined by using the macro average method, where precision and recall numbers for each category is calculated and averaged. The following section describes our experiments by using the SAT system.

4. Experiments

We used 611 McRel benchmarks from the A2A+McRel Compendix for the topic Science. All of these McRel benchmarks had one or more state benchmarks aligned to them. Approximately half of the McRel benchmarks had very few training instances. In general multi-label text categorization problems suffer from non-uniform distribution; positive class usually has fewer examples then the negative class. For certain category labels there might be very few examples in the positive set, which makes it harder for the machine learning algorithms to generalize from the limited set of examples. There has been work done on improving the performance of categorization systems for imbalanced datasets, by doing positive and negative feature selections [17] and sampling of negative examples. Since we are reporting our preliminary results on the SAT, for the initial set of experiments we decided to enforce a minimum training size constraint on the categories we test. We ignored categories that had fewer than 30 training examples, and our final category pool had 133 categories. In our experiments we divided the benchmarks in each category into training and testing. After repeated runs, we settled on a 3/1 split, with 75% of the benchmarks utilized for training and 25% for testing.

A binary classifier was built for each category by employing 75% of a category's benchmarks for positive training while the training benchmarks from the remaining categories served as negative training examples. This binary classifier was then tested in the same manner, 25% of the standards belonging to this category comprised the positive testing set and the testing standards from the remaining categories were treated as the negative test set.

We report classification results here for the 133 categories classified, using the different combinations of the three text content types. The various combinations are defined below and Table 2 shows the results we obtained for the content type combinations.

- Benchmark: the text of the benchmarks alone.
- Benchmark + Hierarchy: The combination of the text of the benchmark and the hierarchical text.
- Vocabulary: The relevant vocabulary set alone.
- Benchmark + Vocabulary: The combination of the text of the benchmarks and relevant vocabulary set for the benchmarks
- Benchmark + Hierarchy + Vocabulary: The combination of the text of the benchmarks, the hierarchical text and the relevant vocabulary set for the benchmarks.

As evident from Table 2, the text of the benchmarks alone and the combination of the benchmark content yielded an acceptable classification performance for a multi-label classification problem with 133 categories. However, for an automated alignment system we thought higher levels of accuracy were necessary. In the experiments where the relevant vocabulary set was the only type of text utilized we obtained extremely high precision and recall.

Table 2: Classification Results for various content types

Content Type	Precision	Recall	F-Measure
Benchmark	58.29	42.32	47.06
Benchmark + Hierarchical	58.00	43.27	47.55
Vocabulary	**92.95**	**84.46**	**87.77**
Benchmark+ Vocabulary	85.42	79.43	81.76
Benchmark+ Hiearchical+ Vocabulary	81.49	76.69	78.34

We believe that the relevant vocabulary set serves as an intermediary for bridging the variations between the standards' texts. Although the relevant vocabulary set is mostly a subset of the standards' text vocabulary[3], the relevant vocabulary set represents key concepts embodied in the benchmarks that may not be consistently represented in the text of every benchmark.

For example, the McRel benchmark :

"Knows that short-term *weather* conditions (e.g., temperature, rain, snow) can change daily, and *weather* patterns change over the seasons"

is represented by these vocabulary words:

A: daily *weather conditions*,
B: *seasonal weather pattern*,
C: *temperature,*
D: snow.

A matching benchmark from Colorado:

"recognizing how our daily activities are affected by the *weather* (for example, types of clothing, travel plans, recreational activity)"

has this vocabulary assigned to it:

A: *weather conditions,*
B: daily weather pattern,
C: *seasonal weather pattern,*
D: *temperature.*

Table 3: Feature vectors

Content Type	Feature Vector Size
Benchmark	3710
Benchmark + Hierarchical	3800
Vocabulary	**647**
Benchmark + Vocabulary	3741
Benchmark + Hierarchical + Vocabulary	3829

In the example above, the first benchmark and the second benchmark share the common word "weather" but use different terms to talk about a similar concept. The relevant vocabulary set provides a link between the two standards by providing three additional shared concepts: "weather conditions", "seasonal weather pattern", and 'temperature".

It is interesting to note that using the hierarchical information (which contains higher-level concepts similar to the relevant vocabulary set) in addition to the benchmark text or the relevant vocabulary set did not result in any significant improvement in accuracy. The hierarchical text is added to multiple benchmarks which belong to other categories. For example, the hierarchical information that would have been added to the previous example about weather, "Understands atmospheric processes and the water cycle" is added to 20 other benchmarks in the McRel Compendix. Thus addition of

[3] Only 31 additional terms were added to the feature vector when the vocabulary and the benchmark were combined, and only 29 distinct terms were added when the hierarchy was included (see Table 3).

hierarchical text introduces confusion and lowers the classification performance.

5. Conclusions and Future Work

Our experiments using text categorization show that automatic standards alignment is feasible with acceptable performance provided we have a particular type of training data. When we used only the standards text in our experiments, the classifier performance varied widely for different categories even with similar training set size. In this closed domain (science) of application, there is high overlap between the texts of unrelated standards, which affects the generalizing capability of the learning algorithm. Using a set of controlled vocabulary to represent standards provided more discriminative features to the learning algorithm and led to better performance for all 133 categories.

The McRel vocabulary terms clearly had an impact on the system's ability to categorize national and state standards correctly. While all standards in our test set have vocabulary assigned to them, this is not true for new and updated standards that we will eventually categorize. To remedy this problem we are currently investigating the possibility of automatically assigning vocabulary to standards that do not have the relevant vocabulary set. We envision this as a preprocessing step before categorization takes place. CNLP has experience in automating processes that have traditionally been manual. Our work on automatic metadata extraction [16] revealed that users were satisfied with the quality of our automatically generated metadata. We expect that the automatic assignment of controlled vocabulary to the benchmarks would generate similar, satisfactory results.

While the experiments described in this paper apply to science standards we are currently applying similar ML algorithms for math.

We are also investigating incorporating additional information into the feature vectors such as terms from lesson plans and other resources that have been associated with a standard. The text in standards is relatively short and additional terms may help in assigning it to the correct category.

6. References

[1] H.Devaul, K. Kelly, "Searching by Educational Standards in DLESE: What does it mean and what do users want?" http://nsdl.comm.nsdl.org/meeting/poster_docs/2003/752_dlese_edStandards.pdf, 2004.

[2] M. Jay and D. Longdon, "Death, Taxes and Correlations: A Primer on the State of Correlation in the K-12 Education Marketplace", *Upgrade*, SIIA, Oct/Nov, 2003, pp. 20-21.

[3] McRel, http://www.mcrel.org/, 2006.

[4] Plato Learning, Inc., http://www.plato.com/, 2006.

[5] Align to Achieve, http://www.aligntoachieve.org/, 2006.

[6] SmartPro, Inc., http://www.smartpro3.com/, 2006.

[7] A. Diekema, "CASAA: Content Alignment Tool," CNLP,http://www.cnlp.org/documents/casaa-web/casaa.html, 2006.

[8] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys, vol. 34, pp. 1-47, 2002.

[9] J. Kendall. "Vocabulary", McRel. http://www.mcrel.org/compendium/kskillsintro.asp#voc, 2006.

[10] O. Yilmazel, "Empirical Selection of NLP-Driven Document Representations For Text Categorization," in Engineering and Computer Science. Syracuse: Syracuse University, 2006, pp. 103.

[11] V. N. Vapnik, The Nature of Statistical Learning Theory. New York: Springer, 1995.

[12] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[13] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization.," presented at 7th ACM International Conference on Information and Knowledge Management, Bethesda,US, 1998.

[14] T. Joachims, Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms: Kluwer Academic Publishers, 2002.

[15] X. Wu, "Support Vector Machines For Text Categorization," in Graduate School, vol. Ph.D. Buffalo,NY: State University of New York at Buffalo, 2004

[16] O. Yilmazel, C. M. Finneran, and E. D. Liddy, "Metaextract: an NLP system to automatically assign

metadata," in Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries. Tuscon, AZ, USA: ACM Press, 2004.

[17] Z. Zheng, X. Wu and R. Srihari, "Feature Selection for Text Categorization on Imbalanced Data," *ACM KDD Explorations Newsletter*, 6(1), June 2004: 80-89.