

# Text Categorization for Aligning Educational Standards

**Ozgur Yilmazel, Niranjan Balasubramanian, Sarah Harwell,  
Jennifer A. Bailey, Anne R. Diekema & Elizabeth D. Liddy**

**Center for Natural Language Processing  
School of Information Studies  
Syracuse University  
Syracuse, NY, USA**

[www.cnlp.org](http://www.cnlp.org)

**HICCS  
January 5, 2007**

# Overview

- A specialized case of text categorization with broad generalizability and diverse applications
  - Business requirements analysis & organization
  - Customer Relationship Management – call routing
- Use Natural Language Processing (NLP) and Machine Learning (ML) to align national and 49 state standards that describe in various ways the same / highly similar educational goals
- Promising results on initial implementation in our Standard Alignment Tool (SAT)

# Motivation

- *No Child Left Behind Act of 2001*
  - *“Goal is to ensure that all children have a fair and equal opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging State academic achievement standards.”*
- States are required to conduct annual assessments of each student’s achievement relative to standards at specific grade levels
  - States’ federal funding depends on performing well on them
  - 49 of the 50 states have written their own standards
    - In Math & Science – 64,000 state standard benchmarks
- Some school districts require teachers to document how their daily lesson plan is “aligned to standards”
  - New teachers are planning their curricula based on these

# National Science Digital Library

- Standards-based testing has a significant impact on the National Science Digital Library (NSDL)
  - Funded by NSF since 2000 to build & maintain an online library for educators of science, technology, engineering, and math
  - Contains lesson plans, learning activities, datasets, software, readings, etc. for teachers
- NSDL items are indexed with up to 23 metadata elements that are available for searching, e.g.
  - Title
  - Creator
  - Subject
  - Grade Level
  - National Standard
    - National Council of Teachers of Mathematics
    - National Science Education Standards
- We developed an automatic metadata assignment system that performed equally well as manual assignment

# Sample Automatic Metadata Generation Output

**Title:** Grand Canyon: Flood! - Stream Channel Erosion  
**Grade Levels:** 6, 7, 8  
**GEM Subjects:** Science--Geology  
Mathematics--Geometry  
Mathematics--Measurement  
Science--Process Skills  
**Keywords:** Colorado River (river), Grand Canyon (geography / location),  
Glen Canyon Dam (buildings&structures) channels, conduit,  
controlled release, dam, reservoir, rivers, sediment, streams,  
volume of flow  
**Pedagogy:** Collaborative learning  
Hands on learning  
**Tool For:** Teachers  
**Resource Type:** Lesson Plan  
**Format:** text/HTML  
**Placed Online:** 1998-09-02  
**Name:** PBS Online  
**Role:** onlineProvider  
**Homepage:** <http://www.pbs.org>

# Current Standards Assignment Situation

- Educational repositories manually assign national standards, BUT:
  - An onerous and time-consuming task
  - Hard to achieve consistency across multiple human assigners
  - Standards are constantly being updated
  - Focus groups of teachers said they wanted to search by own state standard
  - No scalable means for extending to state standards
- McREL Compendium of K-12 Standards
  - Produced a synthesis of standards documents from professional subject-area organizations (NCTM, NSES)
- Align to Achieve (A2A)
  - Aligned some of the state and national standards to McREL's Compendium to produce a database of K-12 standards
  - 1 time effort – no longer in operation
- Neither with the capability for dynamic updating

# CNLP's Solution

- We had already developed and tested the Content Assignment Tool (CAT) to support collection providers in assigning their preferred state standard to an item
  - Semi-automatic approach given a resource's URL or file name
  - CAT processes the resource and suggests relevant standards
  - User-in-the-loop has final say in what standards get assigned
  - ML algorithm is run on vetted assignments and CAT can improve standards assignment over time
- Next, developed the automatic Standards Alignment System (SAT) to correlate resources tagged with one state or national standard with every other state standard

# Levels of Information in Standards

Level 1:  
State

Colorado

Level 2:  
Topic

Math

Science

.

.

.

Arts

Literature

Level 3:  
Grade Level

K

0-4

1

2

Level 4:  
Standard

Standard 1

Standard 2

Standard N

Students understand the processes of scientific investigation and design.

Level 5:  
Benchmark

Benchmark 1

Benchmark 2

Benchmark n

What students know and are able to do includes communicating about investigations and explanations.

# 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
NSES	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

# Natural Language Processing of Benchmarks

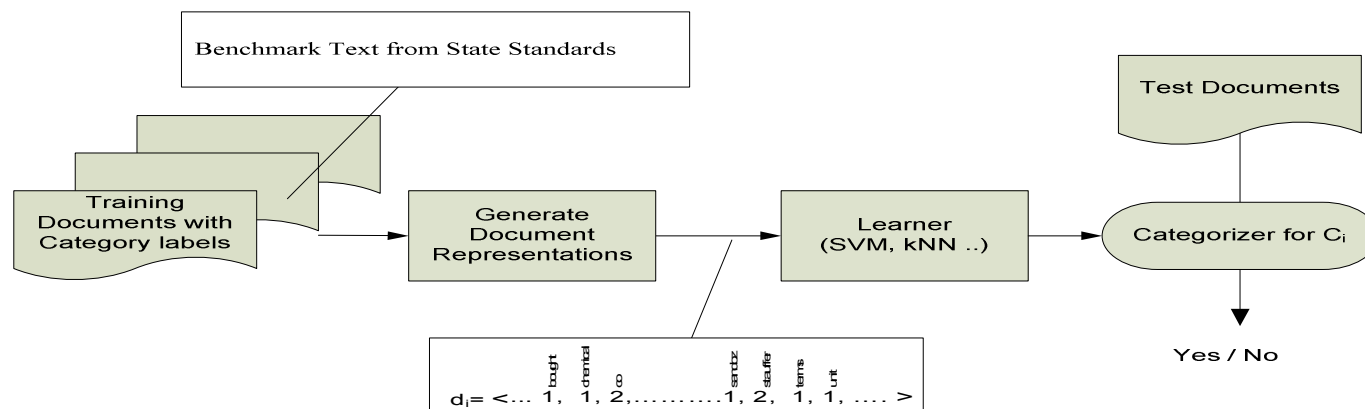
- CNLP's TextTagger
  - Extracts and stems words
  - Does part-of-speech tagging
  - Phrase bracketing
- Represents each benchmark as a vector of extracted terms / phrases
  - Apply classification algorithm to categorize / align vectors
- Alignment is a challenge because humans recognize what makes several benchmarks equivalent, but difficult to do automatically
  - Use same vocabulary when describing different ones
  - Use different vocabulary when describing equivalent ones

# Standard Alignment Tool (SAT)

- Uses a relatively small set of manually-determined alignments between benchmarks to train classifiers for a crosswalk mapping
- Then uses these learned classifiers to align new standards to the crosswalk
  - Thereby enabling alignments between any states and between state and national standards
  - Enables dynamic updating when standards are changed or new ones added
- Mapping is incorporated into NSDL search engine
  - Teachers can search for resources that support their own state standards
    - System will find resources added from anywhere in the US

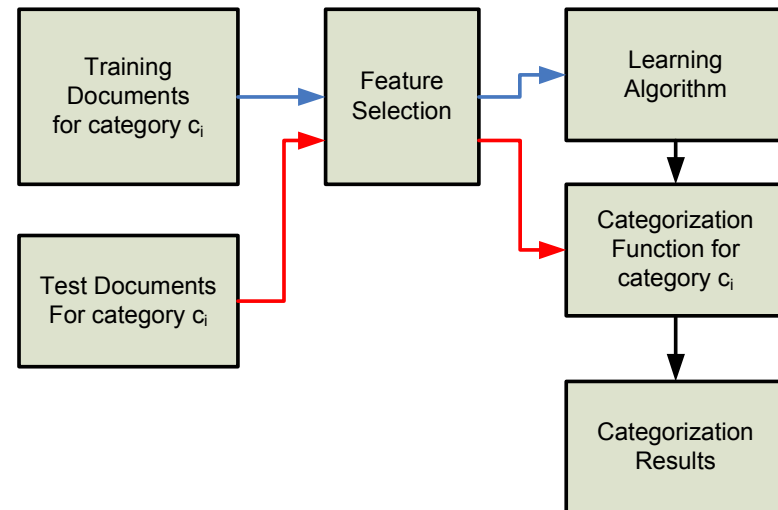
# Text Categorization

- Text Categorization is the task of assigning predefined labels to textual documents
- Algorithm estimates the boolean function  $f$  for category  $c$   
 $\forall d \in \{D_c\} f(d) = \text{true}$  AND  $\forall d \notin \{D_c\} = \text{false}$ 
  - $\{D_c\}$  set of documents in category  $c$  by looking at pre-labeled examples



# Text Categorization for Standards Alignment

- Applied one-vs-all classification at the benchmark level
- Tested different representations of benchmarks
- Used LibSVM algorithm from the MLToolkit
  - Support Vector Machine algorithms have proven successful in our other classification tasks



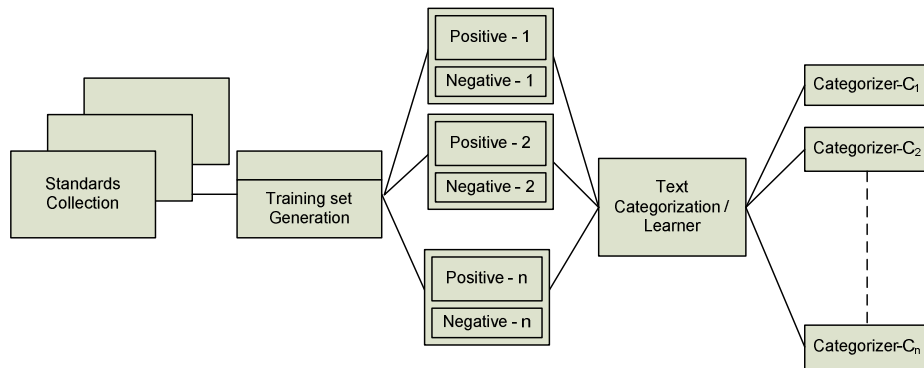
# Document Representations for Experiments

- **Benchmark** – Text of the benchmarks alone
- **Benchmark + Parent** – Both text of the benchmark and the parent categories
- **Vocabulary** – Relevant vocabulary set alone – manually assigned by McRel
- **Benchmark + Vocabulary** – Both text of the benchmarks and relevant vocabulary set for the benchmarks
- **Benchmark + Parent + Vocabulary** – Combination of the text of the benchmarks, the parent categories' text, and the relevant vocabulary set for the benchmarks

# Experiment 1

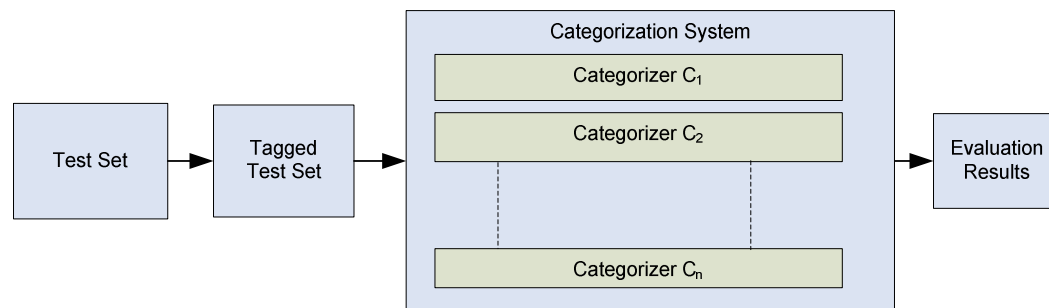
- Used 611 Benchmarks from A2A+McRel Compendix
- Unbalanced datasets with relatively small number of positive examples
- Used benchmarks with more than 30 positive examples
- 133 categories in the end
- Split the dataset: 75% training  
25% testing

# Experimental Setup



Trained categorizers for each of the 133 categories

Evaluated each categorizer with the test set



# Text Categorization Evaluation Metrics

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

		Correct	
		C	$\neg C$
Assigned	C	TP	FP
	$\neg C$	FN	TN

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

# Results – Experiment 1

Representation	Feature Vector Size	Precision	Recall	F-Measure
<b>Benchmark</b>	<b>3710</b>	<b>58.29</b>	<b>42.32</b>	<b>47.06</b>
<b>Benchmark + Parent</b>	<b>3800</b>	<b>58.00</b>	<b>43.27</b>	<b>47.55</b>
<b>Vocabulary</b>	<b>647</b>	<b>92.95</b>	<b>84.46</b>	<b>87.77</b>
<b>Benchmark+ Vocabulary</b>	<b>3741</b>	<b>85.42</b>	<b>79.43</b>	<b>81.76</b>
<b>Benchmark+ Parent+ Vocabulary</b>	<b>3829</b>	<b>81.49</b>	<b>76.69</b>	<b>78.34</b>

# Impact of Relevant Vocabulary

- Best results obtained when using relevant vocabulary
  - Balances disparities between similar, yet differently worded benchmarks
    - Manually assigned to similar state standards by McRel

State	Benchmark	Vocabulary
Kansas	Observe, compare and sort earth materials	Earth materials, solid rock, soil
North Carolina	Compare the components of soil samples from different places	Earth materials, soil

- But as new state standards are added, relevant vocabulary will **not** be assigned
  - Nor when standard is updated, will it be changed
- Needed to obtain better experimental results that do **not** utilize manually-assigned vocabulary

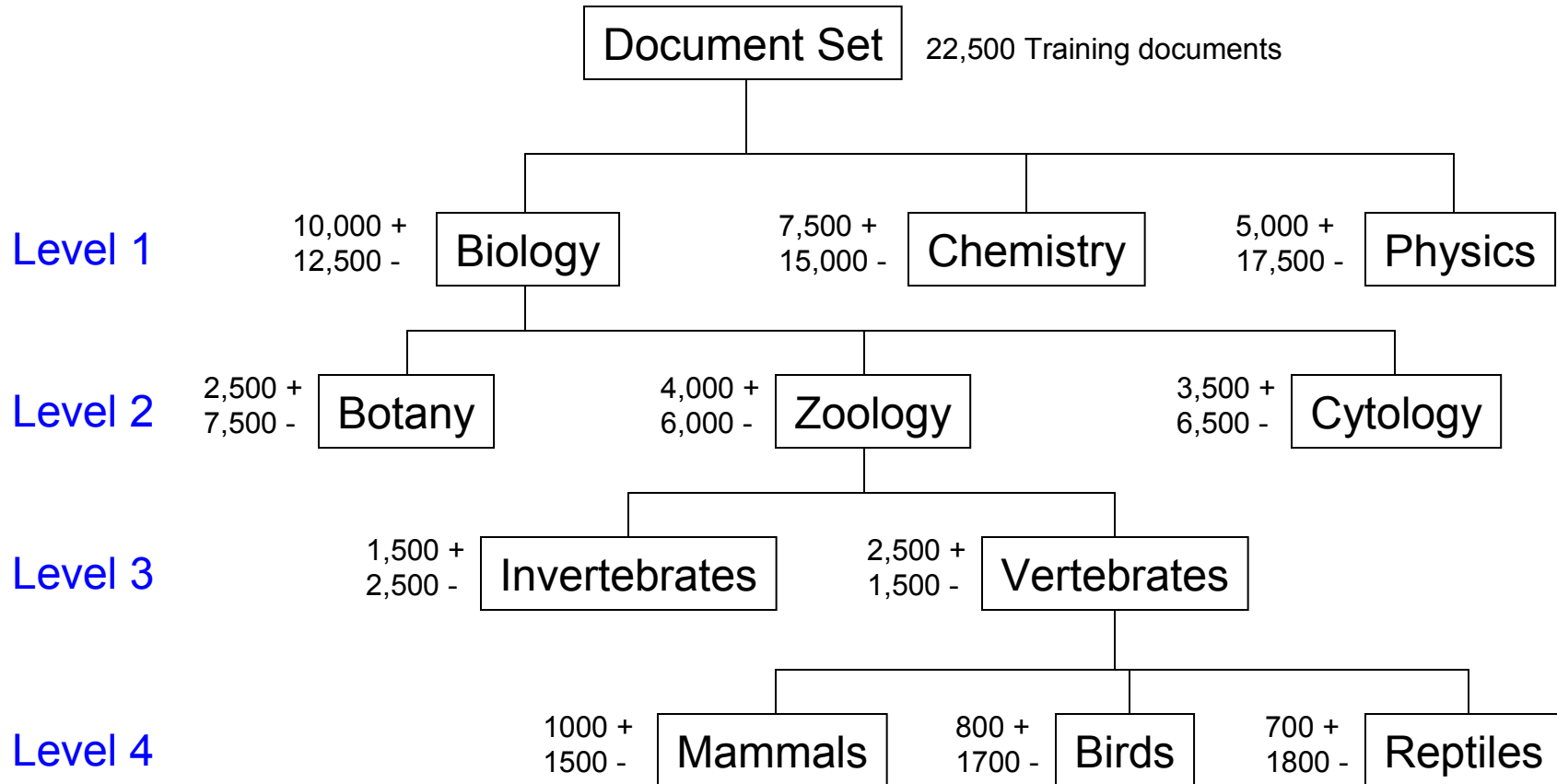
# The Quest for Better Results – Experiment 2

- The one-vs-all classification at the benchmark level created very small positive training sets and large negative training sets
  - Example: A data set contains 30,000 documents
    - 75% of those documents (22,500) are used for training
    - Only 30 are positive examples
    - 22,470 training documents are negative examples for that benchmark
- Without the relevant vocabulary, the disparity between similar benchmarks increases as the size of the negative training set increases

# Hierarchical Categorization

- Utilized the hierarchical nature of the Standards to reduce the size of the negative training set for each category
- If the training set contains 22,500 documents
  - 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: the Biology category can be divided into three categories: Botany, Zoology, and Cytology.
    - Of the 10,000 documents that belong to Biology, 4,000 are positively identified as members of the Zoology category.
      - These are positive training documents for Zoology.
    - Remaining 6,000 documents in the Biology document set do not belong to Zoology and are therefore negative training documents

# Hierarchical Categorization



# Hierarchical Categorization

- At each descending level, pool of training documents from which a model is built is smaller & better defined
- Test documents are used to evaluate models at each level. For example:
  - A test document is evaluated against the Biology, Chemistry, and Physics models. It evaluates as true positive for Biology and true negative for Chemistry and Physics.
  - Since the document tested positive for Biology, it is then evaluated against the Botany, Zoology, and Cytology models. It evaluates as true positive for Zoology and true negative for Botany and Cytology.
  - The document tested positive for Zoology, so it is then evaluated against the Invertebrate and Vertebrate models.
  - This process continues until the last node has been reached.

# Results - Experiment 2

Representation	Categorization Type	Precision	Recall	F-Measure
Benchmark+ Parent+ Vocabulary	Simple	65.445	56.633	60.059
Benchmark+ Vocabulary	Simple	76.632	67.23	70.925
Benchmark + Parent	Simple	33.039	19.957	23.839
<b>Benchmark + Parent</b>	<b>Hierarchical</b>	<b>59.535</b>	<b>96.125</b>	<b>65.491</b>

- Results obtained using a full set of state standards from A2A for Science
- Minimum of 30 documents per category with at least 10 documents for testing
  - Resulted in 230 categories
- Again, using relevant vocabulary produces the best results
  - Not an ideal situation as result is not realistic
- BUT, a significant improvement for benchmark and parent text when using hierarchical categorization vs. simple categorization

# Conclusions

- Automatic standards alignment is feasible with hierarchical categorization algorithm
  - Achieve better results using a relatively small set of training examples compared to one-vs-all categorization technique
- Will do experiments using NLP to detect semantic similarity in differently worded, but conceptually similar standards
  - An automatic means replacing the manually-assigned relevant vocabulary in experiment 1
- A specialized case of text categorization with broad generalizability and diverse applications
  - Business requirements analysis & organization
  - CRM – call routing

# Application – Business Requirements Alignment

## CNLP ModSpec

### Actions

Home  
Suggest  
Indexing  
Project  
View Tree  
Genre Management  
Co-Occurrence  
Clean

### View Tree

- All Projects
  - ...
  - ...
  - ...
  - Stakeholders, Program/Project Teams (1)
    - Business Challenges (5)
  - Business Goals (6)
  - Business Value (9)
    - ModSpec (10)
    - Customers (11)
  - Project Management Requirements (12)
  - Business Requirements (22)
    - Business Non-Functional Requirements (23)
      - Marketing and Sales (24)
      - Requirements Quality Metrics (28)
      - Product(s) (31)
      - Sales (35)
      - Standards (36)

# Thanks!

- Questions?