A Multi-Tier NL-Knowledge Clustering for Classifying Students' Essays

Umarani Pappuswamy, Dumisizwe Bhembe, Pamela W. Jordan and Kurt VanLehn
Learning Research and Development Center,
3939 0'Hara Street, University of Pittsburgh,
Pittsburgh, PA 15260, USA
umarani@pitt.edu, bhembe@yahoo.com, pjordan@pitt.edu, vanlehn@cs.pitt.edu

Abstract

In this paper, we describe a multi-tier Natural Language (NL) clustering approach to text classification for classifying students' essays for tutoring applications. The main task of the classifier is to map the students' essay statements into target concepts, namely physics principles and misconceptions. A simple `Bag-Of-Words (BOW)' classifier using a naïve-Bayes algorithm was unsatisfactory for our purposes as it frequently misclassified due to the semantic relatedness of the NL descriptions of the target concepts. We describe how we used the NL descriptions to define clusters of concepts that reduce the dimensionality of the data when classifying students' essays. The clustering generated multi-tier tagging schemata (cluster, sub-cluster and class) which led to better classification of the student's essay.

1. Introduction

In this paper, we describe a Natural Language (NL) knowledge-based text classification approach for classifying students' essay statements for tutoring applications. The task of the classifier is to recognize these statements from the essay as principles and misconceptions of qualitative physics. In its simplest form, the Text Classification (TC) problem can be formulated as follows: We are given a set of documents D = $\{d_1, d_2, d_3 \ldots d_n\}$ to be classified and C = $\{c_1, c_2, c_3 \ldots c_n\}$ a predefined set of classes. In the Cartesian product D×C, the values {0, 1} are interpreted as a decision to file a document dj under c_i where 0 means that d_j not relevant to the class defined and 1 means that d_j is relevant to the class defined. The main objective here is to devise a learning algorithm that will be able to accurately classify unseen documents from D (given the training set with the desired annotations in the case of supervised learning).

We found that a simple `Bag-Of-Words (BOW)' approach using a Naïve-Bayes (NB) algorithm to categorize text was unsatisfactory for our purposes as it exhibited many misclassifications because of the relatedness of the concepts themselves and its inability to distinguish principles from misconceptions. Issues related to synonymy, ambiguity, and skewed word distributions interfere with forming classification functions (Lewis 1992) and thus pose a challenge to text classification. Hence, we investigate the performance of the k-nearest neighborhood algorithm coupled with pre-defined clusters of semantically related physics concepts for classifying students' essays. Though there have been many studies on clustering at the word level for language modeling and word co-occurrence (Periera et.al. 1993), very little work has been done on concept clustering for document classification.

We present the results of an empirical study conducted on a corpus of students' essays. The approach has a three-tier tagging schemata (cluster, sub-cluster and class) for each document. Let C and SC refer to the Cluster and Subcluster respectively, and `Class (Cl)' refers to the actual principle or misconception being identified. Thus, C in the original definition now takes the form: C = $\{(C_1, SC_1, Cl_1), \ldots (C_n, SC_n, Cl_n)$[1]$\}$. The new C is derived by an in-depth analysis of NL descriptions of each physics concept of interest. This kind of supervised clustering approach helps us to reduce the dimensionality of the texts and thereby leads to a better classification of the student's essay.

The rest of the paper is organized as follows: Section 2 presents an overview of the previous TC approaches used in the Why2-Atlas project along with some experimental results; Section 3 describes our current approach and its experimental setup in detail, Section 4 presents the evaluation of our new method, and Section 5 provides conclusions and directions for future work.

[1] This is meant for a three-tier classification.

2. An Overview of Previous TC Approaches in the WHY2-Atlas project

The Why2-Atlas system presents students with qualitative physics problems and encourages them to write their answers along with detailed explanations to support their answers (VanLehn et al. 2002). Fig. 1 shows a student explanation from our corpus of human-human computer-mediated tutoring sessions. It illustrates the type of explanation the system strives to elicit from students. It is a form of self-explanation so it has the potential to lead students to construct knowledge (Chi et al. 1994), and to expose deep misconceptions (Slotta et al. 1995).

> Question: Suppose you are in a free-falling elevator and you hold your keys motionless right in front of your face and then let go. What will happen to them? Explain.
>
> Explanation (Essay): Free-fall means without gravity. The keys should stay right in front of your face since no force is acting on the keys to move them.

Fig. 1. An actual problem and student explanation.

In the above example, there is a clear statement of misconception `Freefall means without gravity'. Unless we evaluate the answers that students type in, we would not be able to help them reconstruct their knowledge. There are a variety of ways in which a student essay can be evaluated or graded. For instance, Autotutor (Graesser et al. 2000) uses Latent Semantic Analysis to analyze student essays. AutoTutor "comprehends" the student input by segmenting the contributions into speech acts and matching the student's speech acts to the expectations, which would be physics principles in this case. If the expectations are covered in the student's essay, the essay is considered to be `good'.

Why2-Atlas uses a similar method. Using a list of `Principles and Misconceptions', it looks for the presence of such Principles and Misconceptions in the student essay. If the student's essay contains all of the Principles and none of the Misconceptions, then it is considered to be a reasonably good essay and we allow the student to move on to the next problem. Thus, it is important to classify the students' essay statements into Principles and Misconceptions in order to help the student to reconstruct his/her knowledge.

Several attempts have been made by the Why2-Atlas project to analyze students' essays in the past. In all our TC experiments, assuming class "A" as the class of interest and "Not A" as a conjunction of all other classes, there are four possible outcomes when detecting a class "A" as shown in Table 1.

	Predicted `A'	**Predicted `Not A'**
Actual `A'	True Positives (TP)	False Negatives (FN)
Acutal `Not A'	False Positives (FP)	True Negatives (TN)

Table 1: Possible Outcomes when classifying Class `A'

From Table 1, we define precision, recall, accuracy and standard error as follows:

Precision = TP/ (TP+FP)
Recall = TP/ (TP+FN)

Accuracy = (Number of TP predictions + Number of TN predictions) / Number of total instances. Standard error of the prediction is computed over the accuracy.

The goal of the first experiment was to identify "correct answer aspects" in a student's essay. Rosé et al. (2003) defined 6 classes of `keypoints' that refer to "Correct Answer Aspects" present in the student essay and `nothing' otherwise to classify essay strings; average precision and recall for the Pumpkin problem[2] was 81% and 73% respectively.

Another attempt was made in the project for identifying only 'Principles' for five problems. The results are shown in the first five rows of Table 2. As the number of classes increased, the accuracy declined. This was mainly because the classes share many words and that they could easily serve as a classification feature for more than one class.

Set/ Subset[3]	**No. of classes**	**No.of examples**	**Accuracy (%)**	**Std. Error**
Pumpkin	17	465	50.9	1.38
Packet	14	355	55.5	1.99
Keys	20	529	48.5	1.62
Sun	8	216	60.6	1.42
Truck	8	273	65.2	0.93
Global[4]	***38***	***586***	***50.9***	***0.02***

Table 2. Performance of naïve-Bayes (NB) classifier

The question then arose as to whether principle training could be generated across problems. Furthermore, as this approach did not include training for misconceptions, the

[2] "Pumpkin" was one of the 10 training problems given to the students in the tutoring session on kinematics.
[3] Subset includes data for specific problems "Pumpkin", "Keys", "Packet", "Sun" and "Truck", names of kinematics qualitative problems presented to the students. Fig 1 presents the "keys" problem.
[4] This included data from all 5 problems.

classifier grouped all such instances as `nothing' (false negatives) or put them under different `wrong' classes (false positives) neither of which was desirable for our purposes. Because these problems share many principles and misconceptions, we tried to combine the examples from the subsets (in Table 2) into one. We included training examples for misconceptions as well and tested this new dataset using the naïve-Bayes algorithm and the results are shown in the last row of Table 2. Due to the similarity of the words present in the list of principles and misconceptions, there were still many misclassifications. Pappuswamy et al. (2005) present an analysis of a small sample that shows clearly that the complexity of the problem lies in the nature of the natural language used to describe the physics concepts. As the naïve-Bayes algorithm ignored the relationships between significant words that did not co-occur in the document, we investigated the performance of various other classifiers on this issue and decided to use the k-nearest neighborhood algorithm along with our new clustering technique (see section 3.3.2 for details).

3. Experimental Design

In this section, we describe our clustering experiment, the datasets used in the experiment, the clustering technique and the document modeling procedure.

3.1 Dataset

All of the datasets used in this work are extracted from the WHY-Essay[5] corpus which contains 1954 sentences from essays. A list of Principles and Misconceptions that corresponds to physics concepts of interest in the WHY2-Atlas project is used as the set of classes to be assigned to these essay strings. There are 50 such principles (with IDs ranging from P1 to P50) and 53 misconceptions (M1 to M53). A sample of the Principles (P) and Misconceptions (M) is presented in Table 3.

ID	NL description of the P or M
P6	When gravity is the only force acting on an object, it is in freefall
P5	All objects in freefall have the same acceleration.
P44	If A exerts a force on B, B exerts a force on A of the same magnitude and opposite direction.
M53	A body in freefall does not experience force of gravity
P26	Two objects with the same velocity have the same displacement at all times

Table 3. Examples of principles and misconceptions

[5] The WHY-Essay corpus consists of students' essay statements mostly from Spring and Fall 2002 experiments of human-human tutoring sessions.

The training and test data are representative samples of responses to physics problems drawn from the same corpus. The data was tagged for both principles and misconceptions. We used the 2/3 and 1/3 split of training and test data for this experiment. We carried out many classification trials and the performance on `old data' was used to guide data-cleaning and to revise the relations between classes that are to be identified/predicted.

3.2. Creation of Clusters

The Principles and Misconceptions used for tagging the essay segments have similar topics (e.g. gravity-freefall and gravitational force, second law etc) and therefore share many words. The classification task is typically hard because of a lack of unique terms and thus increases the feature dimensionality of these documents. Thus, it is highly desirable to reduce this space to improve the classification accuracy. The standard approach used for this kind of task is to extract a `feature subset' of single words through some kind of scoring measures (for example, using `Info-gain'). The basic idea here is to assign a score to each feature (assigned to each word that occurred in the document), sort these scores, and select a pre-defined number of the best features to form the solution feature subset (as in Latent Semantic Indexing approaches). In contrast to this standard approach, we use a method to reduce the feature dimensionality by grouping "similar" words belonging to specific concept descriptions into a smaller number of `word-clusters' and viewing these features to create concept-clusters. Thus, we reduce the number of features from `hundreds' to `tens'. Though there have been many studies (for example, Hotho et al. (2003)) that use word-clusters to improve the accuracy of unsupervised document classification, there are very few studies that have used this kind of indirect `supervised' clustering techniques for text classification. Baker and McCallum (1998) showed that word-clustering reduced the feature dimensionality with a small change in classification performance. Slonim and Tishby (2001) use an information-bottleneck method to find word-clusters that preserve the information about the document categories and use these clusters as features for classification. They claim that their method showed 18% improvement over the performance of using words directly (given a small training set). Our work is unique in that it uses a multi-tier NL-knowledge-based word-clustering method to label each student essay statement. We endorse the same claims as the other two works, that clustering even when done on concept descriptions instead of directly on the data improves the classification performance significantly.

3.2.1 The Multi-Tier Clustering method

Determining the `similarity' of words in the physics concept descriptions is a difficult task. Given the list of the principles and misconceptions used for tagging the students' essay strings, we examined the semantics of the NL descriptions of each principle and misconception and extracted those words (word clusters) that seemed to best describe a particular concept and put them together. Fig. 2 illustrates this idea. The upper levels (cluster and sub-cluster) describe the topic of discussion and the lower level describes the specific principle or misconception. The + sign in each node means the presence of that particular `word(s)' in a concept description. For example, from the trees in Fig 2, we can see that **+freefall** and **+only force of gravity** describe Principle `P6' while **+freefall** and

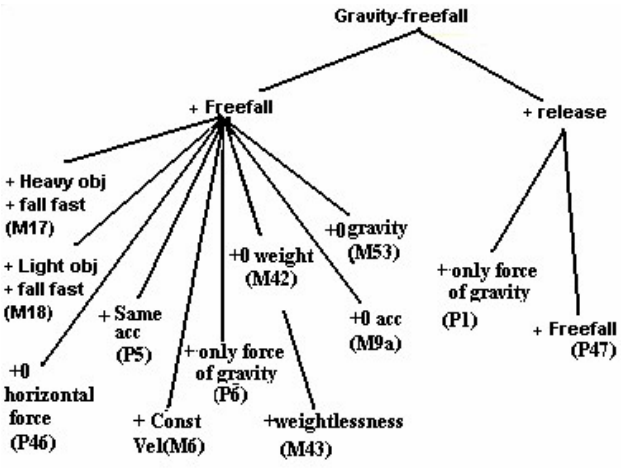

Fig. 2: Chart showing the features related to the cluster `Gravity-Freefall'

+0gravity describe a Misconception `M53' (See Table 3 for the actual NL text description of `P6 and M53'). Thus, words in the lower level that are shared across concepts migrate into an upper tier. The top-most level was created using the concepts described at the middle level. We created ten such clusters based on the prominent keywords for the concept descriptions (see Table 4 for specifics). Absence of sub-clusters (in 4 clusters) means that their features were dissimilar in the group. The number of tiers needed depends on the domain knowledge and the task at hand. A three-tier clustering was sufficient for our purposes. The original corpus was augmented with this information so that the training data took the form:

C = {(clustername, subclustername, class)

as exemplified below:

1. Freefall acceleration is the same for all objects and the keys the person and the elevator are all accelerating downwards with the same acceleration. {*gravity-freefall, freefall-prin, P5* }.
2. If the two forces were equal they would cancel each other out because they are in opposite directions. {*3rdLaw, act-react, M35*}

Cluster	Sub-cluster	Classes	
		P	M
Gravity-freefall	Freefall	3	7
	Release	2	0
Gravitational-force	-	3	11
Secondlaw	Netforce	3	1
	Force	2	8
Thirdlaw	One-object	1	0
	2obj	4	1
	Act-react	0	5
Kinematics and vectors	Force	2	1
	Zero-netforce	4	0
One-object-second-third-law	Lightobj	0	4
	heavyobj	0	2
	objhit	0	1
Two-objects-motion	Samevel	7	2
	cons.vel-over-t	1	0
	jointmotion	3	0
Acceleration-velocity-displacement	-	4	1
Weight-mass	-	0	4
General	-	5	5

Table4. The three-tier clusters of principles and misconceptions

In addition, there was also a `nothing' class. The student statements that were neither a `P' nor a `M' are in this class.

3.3 Document Modeling

Our main interest is to prove that the `BOW approach with clusters' outperforms the `BOW approach without clusters' on students' essay strings. Additionally, we are concerned with how this comparison is affected by the size and the nature of the training set.

We used the bag-of-words representation to index our documents with binary weights (1 denoting presence and 0 absence of the term in the document). A document for us is a whole proposition and not a general topic (commonly used in most BOW approaches to classify web pages).

We investigated the performance of BOW with clusters using the following algorithms: SVM, kNN, Decision Tree (DT) and NB. We randomly chose 163 instances belonging to three different physics concepts (Class a: Freefall, Class b: Gravitational Force and Class c: Release) and built models. From Table 5, it can be seen that both kNN and DT perform equally well and better than the other two learners. However, kNN took less time to predict the classes.

Classifier	Precision (%)	Recall (%)	Accuracy (%)	Time Taken (secs)
SVM	76	77	75.46	1.19
kNN	80	82	79.14	0.03
DT	77	80	79.14	1.19
NB	77	75	76.07	0.19

Table 5. Summary of the classifiers' performance[6]

Fig. 3 shows that there were more misclassifications of class `a' into class `c' in the models kNN and DT but less in the reverse order (i.e. class `c' into class `a'). Based on these initial results, we chose kNN algorithm as the one best-suited for our purposes.

a	b	c
59	1	21
0	21	4
10	4	43

SVM

a	b	c
57	3	21
0	22	3
4	3	50

kNN

a	b	c
59	6	16
2	20	3
5	2	50

DT

a	b	c
67	2	12
3	19	3
17	2	38

NB

Fig.3: Confusion matrices of the classifiers

Furthermore, to get better insight, we also examined the NL descriptions related to classes `a' and `c' and found that they share many words which in turn could be merged into a single concept `Gravity-Freefall'. The class `b', on the other hand, denotes a separate concept `Gravitational force'. This kind of "similarity" across concepts inspired us to build the multi-tier clusters (described in section 3.2) for the classification purposes.

3.4. Building Multi-tier Data Models

Based on the above experiments, we chose the k-Nearest Neighborhood (kNN) algorithm[7] for the classification task. The specifications and design of the classifier are explained in our previous paper (Pappuswamy et al. 2005).

Here we step through an example of how the data models are built. We use an automatic model-maker that can generate models for the multi-tier system given `cluster-maps'. A cluster-map lists the clusters and its members (the tags Ps and Ms). For example, the cluster-map for "gravity-freefall" is:

gravity-freefall:
- prin-drop-only-grav
- prin-release-freefall
- prin-only-gravity-implies-freefall
- prin-freefall-same-accel
- prin-freefall-0-horizontal-force
- misc-freefall-const-vel
- misc-freefall-zero-acc
- misc-freefall-zero-weight
- misc-weightlessness-freefall-zero-weight

After creating all the desired clusters for the top-tier, we define cluster-maps for each sub-cluster described in Table 4. The entire dataset is divided into the specified number of sub-clusters which is further bifurcated into two (one for Ps and another for Ms). For example, the cluster "gravity-freefall" points to a sub-cluster map that contains "freefall' and "release" as its members, where "freefall" is:

Freefall:
- prin-only-gravity-implies-freefall
- prin-freefall-same-accel
- prin-freefall-0-horizontal-force
- misc-freefall-const-vel
- misc-freefall-zero-acc
- misc-freefall-zero-weight
- misc-weightlessness-freefall-zero-weight

The next step is to create a training file (e.g. Fig. 4) for each cluster. Additionally, we build a model for each cluster to recognize the parent of each P or M.

gravity-freefall.train.out `F' since the object is falling and has only gravitational force on it.
gravity-freefall.train.out `F' ok so freefall means that everything is moving under the force of gravity
gravity-freefall.train.out `R' because after the cable snaps the only force acting on the keys is gravity

Fig 4: Excerpt for cluster `gravity-freefall'; F= SC1, R= SC2

[6] Average precision and recall is computed across classes for each model.

[7] We used the kNN algorithm from the RAINBOW software devised by McCallum (1996). McCallum, Andrew Kachites. "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering." http://www.cs.cmu.edu/~mccallum/bow

The sub-cluster and class models are created in the same fashion. Thus, we have classification outputs at each level for the cluster, sub-cluster and class tags respectively. At runtime for previously unseen student statements, the output of a level is used to select a model in the next level.

4. Evaluation

The metrics used to measure the performance of the three-tiered learner are: accuracy, standard error, and precision (as defined in Section 2) and the results are reported in Table 6. In our current context, if a document D is related to a C, it will be considered to be a `TP', with a value of `1'. If a document D is not related to C, it will have a value of `0' and can either be marked as `nothing' which constitutes the `FN' for us or it can be misclassified (as some other C) which means that it is a `FP'. For example, if a student string `Freefall means without gravity, is correctly classified as misconception statement (M53), it is a TP. On the other hand, if it is categorized as `nothing' then it is a `FN' and if it is misclassified as anything else then it is `FP'.

Model		Precision (%)	Recall (%)	Acc (%)	Std. Error
Three-tier clustering	Cluster (one level)	80.9	92.1	78.0	0.016
	Subcluster (two levels)	74.3	88.7	74.5	0.020
	Classes (three levels)	62.6	90.7	64.2	0.185
Without clustering (using NB)		68.6	83.4	50.9	0.020

Table 6. Performance measures for the three-tier model

The above results show that the three-tier clustering indeed helped to improve the performance of the classification. Ambiguity (or noise) among classes was significantly reduced as the documents were forced to traverse the whole path (cluster → sub-clusters → classes). Our model significantly outperformed the bow-only approach using the NB classifier with an improvement of 27.02%, 23.51% and 13.17% in the classification accuracy for the levels 1, 2 and 3 respectively.

5. Conclusions

This paper discussed a multi-tier clustering approach for classifying data pertaining to students' essays about qualitative physics problems in a tutoring system. We showed that the integration of concepts and sub-concepts into the feature representation improves classification result. This is very important for evaluating essays because it gives additional information about the topics of discussion to the Why2-Atlas system. We intend to investigate semi-automatic methods to extract the now hand-crafted features at the sub-cluster level with the goal of further reducing misclassifications. We also conjecture that expansion of the training corpus to include more examples for the concepts will further improve the clustering results.

Acknowledgements

This work was funded by grant N00014-00-1-0600 from ONR Cognitive Science and NSF grant 0325054.

References

Baker L.D. and A.K.McCallum. 1998. Distributional Clustering of Words for Text Classification. In *ACM SIGIR 98*.

Chi M, N de Leeuw, M.Chiu, and C.LaVancher. 1994. Eliciting self-explanations improves understanding. *Cognitive Science*, 18:439–477.

Graesser A.C, P.Wiemer-Hastings, K..Wiemer-Hastings, D. Harter, N. Person, and the Tutoring Research Group. 2000. Using Latent Semantic Analysis to Evaluate the Contributions of Students in AUTOTUTOR. *Interactive Learning Environments* 8:129–148.

Hotho A, S. Staab and G. Stumme. 2003. Text Clustering Based on Background Knowledge. Institute of Applied Informatics and Formal Description Methods AIFB, *Technical Report No. 425*.

Lewis D. 1992. An evaluation of phrasal and clustered representations on a text categorization task. *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval.* Copenhagen, Denmark . pages: 37 – 50.

Pappuswamy U, D.Bhembe, P.Jordan and K.VanLehn. 2005. A Supervised Clustering Method for Text Classification, Lecture Notes in Computer Science, Volume 3406, Jan 2005, Pages 704 – 714.

Periera F, Nishby, and L.Lee. 1993. Distributional clustering of English words. In *31st Annual Meeting of the ACL,* Pages: 183-190.

Slonim N. and N.Tishby. 2001. The power of word clusters for text classification. In *Proceedings of ECIR-01, 23rd European Colloquium on Information Retrieval Research*.

Slotta,J, M.Chi ,and E.Joram. 1995. Assessing students' misclassifications of physics concepts: An ontological basis for conceptual change. *Cognition and Instruction*, 13(3):373–400.

Rosé C.P, A. Roque, D.Bhembe, K.VanLehn.2003. A Hybrid Text Classification Approach for Analysis of Student Essays, *Proceedings of the HLT-NAACL 03 Workshop on Educational Applications of NLP*.

VanLehn K, P. Jordan, C.P.Rosé, D.Bhembe, M.Bottner, A. Gaydos, M.Makatchev, U.Pappuswamy, M.Ringenberg, A.Roque, S Siler, and R Srivastava. 2002. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proceedings of Intelligent Tutoring Systems Conference*, volume 2363 of *LNCS*, pages: 158–167. Springer.