Language Based Mapping of Science Assessment Items to Skills

Farah Nadeem and **Mari Ostendorf**
Dept. of Electrical Engineering
University of Washington
{farahn,ostendor}@uw.edu

Abstract

Knowledge of the association between assessment questions and the skills required to solve them is necessary for analysis of student learning. This association, often represented as a Q-matrix, is either hand-labeled by domain experts or learned as latent variables given a large student response data set. As a means of automating the match to formal standards, this paper uses neural text classification methods, leveraging the language in the standards documents to identify online text for a proxy training task. Experiments involve identifying the topic and crosscutting concepts of middle school science questions leveraging multi-task training. Results show that it is possible to automatically build a Q-matrix without student response data and using a modest number of hand-labeled questions.

1 Introduction

In both traditional and online contexts, fine grain diagnostic information can play a crucial role in employing formative assessment to improve student learning outcomes as observed by National Research Council (2012), and for creating scalable systems that provide individualized instruction (Barnes, 2005). A key requirement for this inference is association of each of the assessment tasks, which we will refer to as questions, with attributes, which are the skills (knowledge, concepts and/or strategies) needed to solve the tasks. The association of skills to questions is represented as a Q-matrix (Tatsuoka, 1983).

Hand crafted Q-matrices are created by domain experts who label each assessment task with the required skill(s). While this provides an interpretable matrix for educators, in the sense that the skills are associated with a documented standard or cognitive model, the question annotation process is time consuming and not scalable. When standards change, the old question annotation is no longer useful. The cost of question annotation is a key issue with the domain models in intelligent tutoring systems (ITS), which are created by experts for each subject area and grade level, limiting reusability (Burns et al., 2014).

As an alternative, there has been work on automated discovery of an association of (latent) skills to questions using student response data (Lan et al., 2014; Barnes, 2005; Desmarais, 2010). While these unsupervised automated methods can provide a good fit for the student response data, they are limited by the requirement of a large data set of student scores on a given test, which is not available for individual classroom assessments and hard to obtain for standardized testing. In addition, the latent skills offer limited interpretability for teachers. The results cannot easily be used to identify practice questions to help a student improve in areas of weakness.

It was observed in a report by National Research Council (2001) that fine grained diagnostic models are not widely used due to scalability, reusability and/or interpretability issues, which is still a problem today as stated by National Research Council (2012).

This work aims to develop interpretable and automatic methods for mapping science assessment tasks to underlying skills by using text classification methods that leverage the language in standards documents and teacher training materials. The experiments here use the Framework for K-12 Science education laid out in the framework by National Research Council (2012), but the method is designed to work for any well documented standard and the questions used in this study are not

Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pages 319–326
Copenhagen, Denmark, September 8, 2017.

explicitly designed for that standard.

Specifically, we look at the core disciplinary ideas (topics) and crosscutting concepts described in the standard as the attributes needed to respond to assessment tasks. A multi-task convolutional neural network is designed to jointly label topics and concepts. The greater challenge is in recognizing concepts, for which there is no annotated data available. A key contribution is in the use of standards documentation to automate training annotation and obtain online text for use as a proxy task in an intermediate training stage.

The rest of the paper is organized as follows. Sec. 2 provides a detailed task description, which is followed in Sec. 3 by an overview of prior text classification work that we build on. Experiment details are provided in Sec. 4 with results in Sec. 5. Related work leveraging question text in latent skill learning is discussed in Sec. 6. Findings and open questions are summarized in Sec. 7.

2 Task

From the perspective of formal standards, student learning is measured along specified content areas and concepts. The goal of both classroom teaching and online instruction systems is to ultimately increase proficiency in these areas. This work considers the Framework for K-12 science education presented by National Research Council (2012), and the associated Next Generation Science Standard (NGSS Lead States, 2013). The framework measures student learning along three dimensions: i) disciplinary core areas, ii) crosscutting concepts, and iii) science and engineering practices. For this work, we aim to identify the core content areas (topics) and crosscutting concepts associated with a question. The dimension of science and engineering practices is reflected more in the text of student response, hence we do not consider it here.

NGSS provides content and learning progression descriptions for each dimension. The standard specifies a hierarchy of disciplinary core ideas from physical sciences (PS), life sciences (LS), earth and space sciences (ESS), and engineering, technology and application of sciences (ETS).[1] Our study operates at the middle level of the hierarchy, with 12 topics, focusing on the middle school level. Seven crosscutting concepts are described.[2] Examples of descriptions in the standard are given below.

Topic: ESS3 Earth and human activity - Human activities have altered the biosphere, sometimes damaging it, although changes to ...

Concept: Energy and Matter Tracking energy and matter flows, into, out of, and within systems helps one understand their systems behavior.

The specific task addressed in this work is: given a question, identify the topic and concepts associated with that question. For example, the question:

> *What happens to the sun's energy in the greenhouse effect?*

would be be associated with the topic ESS3 and the concept "Energy and Matter." Topic labeling corresponds to a multi-class decision (which one of 12 topics), and concept labeling involves 7 binary decisions. It is possible for a question to involve none of the concepts in the inventory.

More examples of topic and concept descriptions with sample questions are provided in supplementary materials.

3 Text Classification

Text classification is an established problem, with many different techniques available, including naive Bayes, support vector classifiers, decision trees and k nearest neighbors, which are summarized in (Ikonomakis et al., 2005). For longer documents, a bag-of-words approach is often used, but sequence models can be more useful for classifying sentences or short documents. A variety of neural techniques have been proposed, including (Wiener et al., 1995; Ruiz and Srinivasan, 1998; Nam et al., 2014; Lai et al., 2015). In our study, we build on the convolutional neural network (CNN) presented in (Kim, 2014), which achieves high accuracy for short texts. We briefly describe the model below.

The model takes a sequence of pre-trained word embeddings as input: each word x_i is represented by a k dimensional embedding vector, $x_i \in \mathbf{R}^k$. A sequence of n word embeddings are concatenated to form a $n \times k$ matrix that is input to the network.

The concatenated sequence is convolved with filters that span the entire embedding and h words.

[1] https://www.nextgenscience.org/get-to-know, Appendices E and J

[2] https://www.nextgenscience.org/get-to-know, Appendix G

A $h \times k$ filter w_l is convolved with a concatenated sequence of h words, generating an $n-h+1$ length output sequence, where the ith element of the sequence is given by

$$c_i(l) = f(w_l \circ x_{i:i+h-1} + b_l). \quad (1)$$

where $\circ$ indicates a Hadamard product, and f is a non-linear or piece-wise linear function such as a rectified linear unit (ReLU). Using max-pooling over time results in one feature produced by one filter:

$$\hat{c}(l) = \max\{c_1(l), c_2(l), ..., c_{n-h+1}(l)\} \quad (2)$$

The output from max-pooling for each filter is concatenated into a feature set, resulting in an m dimensional feature vector for m filters.

$$z = [\hat{c}(1), \hat{c}(2), ..., \hat{c}(m)] \quad (3)$$

The output used to predict the label is given by

$$y = g(Yz + b) \quad (4)$$

where g is a non-linear function, e.g. softmax for multi-class problems.

4 Methods

This section first describes the different data sets used in training and testing, and then the modifications to the above CNN for identifying question attributes.

4.1 Data

For our work we consider three sets of resources: science questions, Wikipedia science and mathematics articles, and standard related resources. Middle school science questions are the main training and testing data. Wikipedia articles provide a supplemental source of data for pre-training and for a proxy task for concepts. Standard related resources include descriptions of disciplinary core ideas and crosscutting concepts laid out in the standard, and question templates developed to aid teachers in assessing crosscutting concept proficiency.

The main data consists of 14,985 middle school science questions (Kembhavi et al., 2017), with questions divided into 629 generic science modules. This data represents a generic set of middle school science questions, and is not aligned with the dimensions of NGSS. For our study, the modules were hand-labeled as belonging to one of the 12 topics using NGSS descriptions, and topic labels for questions were determined based on the module label. All questions have module labels. The test data consists of 750 questions (5% of the total data); the rest is for training and validation. Only the test data is hand-labeled with crosscutting concepts.

In order to obtain concept labels for the training data, we used question templates that have been developed for each of the seven concepts,[3] which were designed to aid teachers in implementing these concepts into their own assessments. For example, one of the templates for the **Patterns** concept is:

> *What patterns do you observe in the data presented above in the chart?*

We pick keywords from each of the templates (e.g. "patterns", "presented", "observe", "data" and "chart" in the above question), and search for questions in the training data that contain at least two keywords associated with a concept. This results in labels for 890 questions, approximately 6% of the training set, of which 44 questions are assigned multiple concept labels. Keyword matching returned few matches since the questions have not been developed to test for crosscutting concepts specifically. Twenty percent of the results from the keyword search were randomly sampled and hand checked to ensure they matched the assigned concepts, and found to be correct.

The distribution of topics and counts of concepts in the question training and test sets are shown in tables 1 and 2, respectively. Both tables indicate that the class distributions are not balanced. This is expected, since the topics and concepts are designed for all grades (K-12), and some skills are more applicable to high school science curriculum.

Since only 6% of the training data has concept labels, we use external data to create an additional proxy training task for concepts. Using phrases from the concept descriptions from NGSS and the STEM Teaching Tools templates, we hand selected 122 Wikipedia science articles associated with the seven concepts, with 8 of these articles spanning multiple concepts. Sample article titles include:

[3] STEM teaching tools 2014-2017: http://stemteachingtools.org/

Topic	Train	Test
Matter and its interactions	12.0%	10.7%
From molecules to organisms	30.2%	31.2%
Engineering design	2.0%	2.4%
Heredity	1.1%	0.9%
Earth's place in the universe	13.7%	13.1%
Motion and stability	5.0%	4.3%
Earth and human activity	7.1%	6.7%
Waves and energy transfer	5.6%	5.6%
Earths systems	17.0%	18.6%
Energy	2.4%	2.0%
Ecosystems	3.1%	3.6%
Evolution	0.8%	0.9%

Table 1: Topic distribution in train and test sets

Concept	Train	Test
Stability & change (S&C)	46	5
Scale, proportion & quantity (SP&Q)	150	77
Patterns (P)	166	93
Structure & function (S&F)	24	9
Systems & system models (Sys)	88	27
Cause & effect (C&E)	97	41
Energy & matter (E&M)	365	116

Table 2: Concept counts in train and test sets

Patterns: Patterns in nature, Taxonomy, Correlation and dependence

System and system models: System, Systems modeling, System design

The articles are split into smaller "documents" based on linebreaks, yielding a set of 16,892 documents with concept labels. These documents correspond to paragraphs, section heads, related links and references, so they are not a good match to the style of a science question but they provide training examples of important keywords and phrases.

In addition to these sources, we use a pool of 40,000 general Wikipedia science and mathematics articles for pre-training word embeddings.

In summary, four data sources are used in training: questions labeled with both topic and concept D_{TC}, questions labeled with topic D_T (a superset of D_{TC}), concept-labeled Wikipedia paragraphs D_C, and unlabeled Wikipedia articles D_W, as shown in table 3.

Data	Number of Samples
D_{TC}	890 questions
D_T	14,235 questions
D_C	16,892 documents
D_W	40,000 articles
Test set	750 questions with topic and concept labels

Table 3: Data

4.2 Multi-task Topic-Concept Classification

Our use of the CNN for text classification involves multiple outputs:

$$y_t = g_t(Y_t z + b_t) \quad (5)$$

$$y_c = g_c(Y_c z + b_c) \quad (6)$$

where z is the output of the max-pooling layer, $Y_t \in \mathbf{R}^{m \times n_t}$, $b_t \in \mathbf{R}^{n_t}$, $Y_c \in \mathbf{R}^{m \times n_c}$, and $b_c \in \mathbf{R}^{n_c}$. For topics, g_t is a softmax layer, giving the per-class probabilities y_{ti}, from which the topic is chosen according to $argmax_i(y_{ti})$. For detecting multi-label concepts, g_c is a sigmoid, which outputs the concept probability without assuming that concepts are mutually exclusive. The labels are decided by thresholding the sigmoid output, $c_i = \{\mathbf{1}_{y_{ci}(k) > thr}\}^{n_c}$ for $k = 1, \ldots, n_c$. As noted earlier, there are 12 topics ($n_t = 12$) and 7 concepts ($n_c = 7$). The CNN for multi-task training is shown in figure 1.

The training loss function is multi-class cross-entropy for the topic output and binary cross-entropy for each of the concept outputs. Multi-task training uses a sum of topic and concept loss. 10% of the training data was used for validation at a time, using ten fold cross validation. This was used to tune drop out, set number of filters and filter sizes. The entire labeled data set was then used for training.

Training was done with both pre-trained and randomly initialized word embeddings. We used 128 dimensional word embeddings, and a vocabulary size of $75,000$. The filter lengths used were $[1, 3, 4, 5]$, with 64 filters for each size. Drop out of 50% was used for regularization. For concept classification, a threshold of 0.2 was set for positive label detection. This was empirically chosen on the training data.

Experiments are conducted to compare: i) random initialization vs. pre-training, ii) independent vs. multi-task training, and iii) different methods

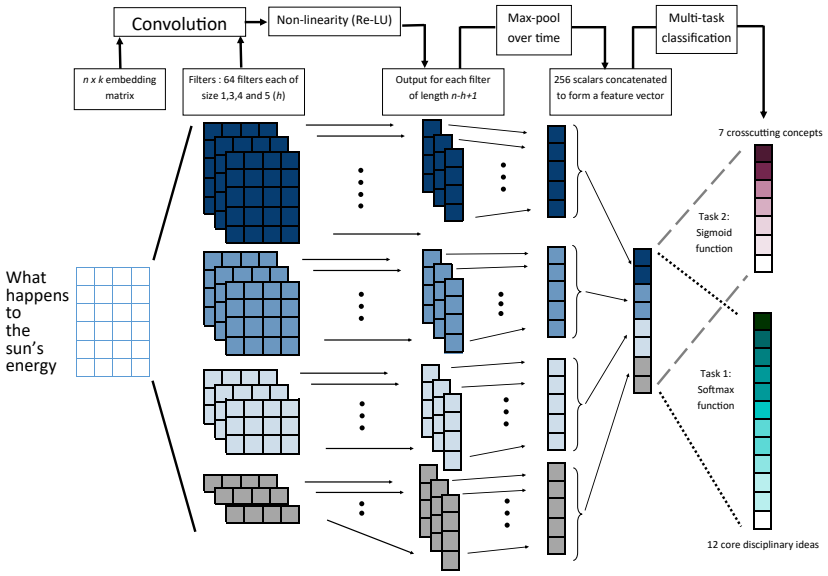

Figure 1: CNN for Multi-task Classification

of using the labeled training sets. Pre-training is based on D_W for both multi-task and independent classifiers.

For independent classifier training, D_T is used for topics, and D_{TC} is used for concepts. In addition, we explored a two-stage approach to training the concept model, using D_C in a first pass of training, followed by D_{TC} for fine tuning ($C \rightarrow TC$).

For multi-task training, three alternatives are explored:

- Stage 1: run single task training (with D_T). Stage 2: run multi-task training with D_{TC}. Do not use D_C. ($T \rightarrow TC$)
- Stage 1: alternate between batches of single task training (with D_T and D_C), starting with D_T. Stage 2: run multi-task training with D_{TC}. ($T/C \rightarrow TC$)
- Alternate between batches of the different labeled sets, starting with D_T and ending with D_{TC}. ($T/C/TC$)

All multi-task models are pre-trained using D_W.

5 Results and Discussion

Results for the different training schemes are shown in table 4. The first four rows correspond to systems with topic and concept classifiers trained separately, and the last three involve multi-task training. The first row indicates baseline performance using n-gram features in an SVM. Comparing the next two rows in the table shows that pre-training word embeddings with the unlabeled Wikipedia articles benefits both topic and concept classifiers, so it was used in all subsequent experiments with multi-task training. The fourth row uses the two-stage concept training, which slightly hurts performance. All the different options for multi-task training (rows 5-7) improve over learning independent classifiers (row 3 for the case with pre-training). Unlike the independent training case, the proxy concepts represented by the Wikipedia article paragraphs benefit both topic and concept labeling when used in multi-task training.

The precision, recall and F1 scores for crosscutting concepts are shown in table 5. As expected, the best performance is observed for the class that dominates the training data. The topic confusion matrix also shows that topics which are well represented in the training data tend to be more reliably identified.

In order to ensure that the independent CNN classifiers provided a strong baseline, we also ran experiments with other approaches using the same training data. Specifically, we implemented an SVM with n-gram features ($n = 1, 2, 3$) and a k-nearest neighbor classifier using a vector space representation of questions based on latent semantic analysis (LSA). Two independent SVMs were trained, one using D_T for topic classification, and one using D_{TC} for concept classification. Performance on topic classification was slightly worse than the CNN result, but results for concept recog-

Training	Initialize	Topic	Concept
SVM	-	73.4	14.2
Separate T, TC	Random	81.3	34.7
Separate T, TC	Pre-train	84.1	38.2
$C \rightarrow TC$	Pre-train	-	35.8
$T \rightarrow TC$	Pre-train	84.5	41.2
$T/C/TC$	Pre-train	84.5	44.4
$T/C \rightarrow TC$	Pre-train	86.2	57.7

Table 4: Classifier Performance: Topic (accuracy) and concept (macro F1 score)

Concept	F1 Score	Precision	Recall
S&C	54.54	50.00	60.00
SP&Q	55.62	51.08	61.03
P	60.96	60.63	61.29
S&F	66.67	66.67	66.67
Sys	43.47	52.63	37.03
C&E	50.57	47.82	61.03
E&M	72.16	60.00	90.51

Table 5: Per-Concept Classification Performance

nition with the best macro F1 being 14.2 with an SVM. The LSA-based model gave much worse results when trained using D_{TC} and tested for concept accuracy; presumably topic factors dominate this unsupervised representation.

An additional factor that impacts performance for concept labeling is ignoring the data in figures accompanying the questions. Generally, for crosscutting concepts, some information is presented graphically, which we are not using in the current work. Hand annotating 200 questions from the test set, we find that roughly a quarter of the questions have associated images. Not all crosscutting concepts are impacted by the presence of an accompanying image. The categories that do worse are scale proportion and quantity, where questions are accompanied by graphs, energy and matter flows, with questions related to water, carbon and nitrogen cycles, and system and system models, which have associated block diagrams. It would be possible to achieve higher accuracy by combining information from the text and features from associated figures, since using text alone is not always enough to identify the correct concept. Consider the following question: *"Which gas is represented by letter F?"* Without the accompanying figure that depicts the carbon cycle, it is not possible to identify the underlying concept of matter and energy flow.

6 Related Work

As noted earlier, automated discovery of latent skills to question mapping provide a good fit to student response data, but the skills are abstract and cannot be easily used by teachers. In (Barnes, 2005; Lan et al., 2014), this problem is addressed by hand-labeling questions with topics and associating the latent concepts learned the different topics that are most frequently represented in the corresponding data. Interpretation of the latent factors is in terms of these topic combinations. This requires hand labeling of training data. The results may generalize to other data, but this was not evaluated. In related work, (Lan et al., 2013) uses multi-objective optimization to learn both skill-to-item and student-to-skill proficiency mappings, as well finding a list of keywords associated with each estimated skill. While both solutions add to the interpretability of the model, the skills are not aligned with formal standards or cognitive models.

Non-negative matrix factorization is used by (Desmarais, 2010) to associate questions with skills using student response data. The data sets consist of 4 subject (mathematics, biology, world history and French). The number of latent skills are 4, the hypothesis is that matrix factorization should separate student proficiency in the four subjects. This work does not provide fine grained proficiency within individual subjects. The model achieves 72% accuracy on all four subjects, and 96% on only mathematics and French, which are the most separable. Results on a set of trivia questions are also reported, where latent skill to topic matching achieves an accuracy of 35% for 4 topics.

7 Conclusion

In summary, this work provides a method for identifying skills required to solve specific science questions based on the text of the question, where skills associated with documented standards are characterized with a relatively small amount of manual annotation. We use state-of-the-art text classification methods that are made more effective by: i) leveraging standards documentation to harvest and automatically annotate training data, and ii) applying multi-task learning to jointly classify both topics and concepts. The best case mod-

els achieve 86% topic accuracy and 57.7 concept F1 score. Compared to current data driven models that are unsupervised and do not provide an explicit connection to standards, our approach is interpretable. In addition, it does not require student response data and can be used with any question set. Compared to frameworks that require human experts to align questions with attributes, our approach is scalable to large question sets. It enables teachers to leverage a variety of assessment materials and provide more individualized feedback to students. While the experiments described here are based on NGSS documentation, the methods are general and can be used with any well-documented standard or cognitive model.

The ability to automatically build a Q-matrix is promising for student learning evaluation and statistical models for online systems, particularly intelligent tutoring systems. Aligning the Q-matrix to elements of learning outcomes specified in standards gives the ability to automatically adapt existing material to new standards and curricula without extensive input from domain experts, improving reusability of tutoring system material. It can also provide new tools for educators analyzing learning across larger populations. In particular, the concept annotation work can help educators study learning progression along crosscutting concepts, which is largely undocumented at this point as stated in the report by National Research Council (2012). It can also provide a complementary tool that may be useful for interpreting unsupervised analyses based on large student response data sets. For example, it may be interesting to look for factors that are predictive of question difficulty based on classifier predictions or confidence of different skills.

Whether the level of accuracy is sufficient for downstream tasks is an open question, since goodness of fit is generally evaluated using student response data, which is not used in the current work. However, there are multiple opportunities for improvement, particularly for concept classification. For example, semantic similarity can be leveraged in using question templates to select articles associated with concepts, and the data could be filtered to exclude sections that are not well matched to questions. Semi-supervised training could increase the number of actual questions used in training. In addition, the use of information in tables and figures represents an important direction for future work. Neural classifiers are well suited to integrating features from different modalities, and we expect that significant gains may be possible with this approach.

References

Tiffany Barnes. 2005. The q-matrix method: Mining student response data for knowledge. In *American Association for Artificial Intelligence 2005 Educational Data Mining Workshop*. pages 1–8.

Hugh Burns, Carol A Luckhardt, James W Parlett, and Carol L Redfield. 2014. *Intelligent tutoring systems: Evolutions in design*. Psychology Press.

Michel Desmarais. 2010. Conditions for effectively deriving a q-matrix from data with non-negative matrix factorization. In *Educational Data Mining 2011*.

M Ikonomakis, S Kotsiantis, and V Tampakas. 2005. Text classification using machine learning techniques. *WSEAS transactions on computers* 4(8):966–974.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* .

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *AAAI*. volume 333, pages 2267–2273.

Andrew S Lan, Christoph Studer, Andrew E Waters, and Richard G Baraniuk. 2013. Joint topic modeling and factor analysis of textual information and graded response data. *arXiv preprint arXiv:1305.1956* .

Andrew S Lan, Andrew E Waters, Christoph Studer, and Richard G Baraniuk. 2014. Sparse factor analysis for learning and content analytics. *Journal of Machine Learning Research* 15(1):1959–2008.

Jinseok Nam, Jungi Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. 2014. Large-scale multi-label text classification - revisiting neural networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pages 437–452.

National Research Council. 2001. *Knowing what students know: The science and design of educational assessment*. National Academies Press.

National Research Council. 2012. *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

NGSS Lead States. 2013. Next generation science standards: For states, by states http://www.nextgenscience.org/.

Miguel E Ruiz and Padmini Srinivasan. 1998. Automatic text categorization using neural networks. In *Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research*. pages 59–72.

Kikumi K Tatsuoka. 1983. Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of educational measurement* 20(4):345–354.

Erik Wiener, Jan O Pedersen, Andreas S Weigend, et al. 1995. A neural network approach to topic spotting. In *Proceedings of SDAIR-95, 4th annual symposium on document analysis and information retrieval*. Las Vegas, NV, volume 317, page 332.