Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5370–5382
Marseille, 11–16 May 2020

Multi-class Hierarchical Question Classification for Multiple Choice Science Exams

**Dongfang Xu*, Peter Jansen*, Jaycie Martin†, Zhengnan Xie†, Vikas Yadav*,
Harish Tayyar Madabushi‡, Oyvind Tafjord§ and Peter Clark§**
*School of Information, University of Arizona, Tucson, AZ, USA
†Department of Linguistics, University of Arizona, Tucson, AZ, USA
‡School of Computer Science, University of Birmingham, Birmingham, UK
§Allen Insitute for Artificial Intelligence, Seattle, WA, USA
pajansen@email.arizona.edu

Abstract

Prior work has demonstrated that question classification (QC), recognizing the *problem domain* of a question, can help answer it more accurately. However, developing strong QC algorithms has been hindered by the limited size and complexity of annotated data available. To address this, we present the largest challenge dataset for QC, containing 7,787 science exam questions paired with detailed classification labels from a fine-grained hierarchical taxonomy of 406 problem domains. We then show that a BERT-based model trained on this dataset achieves a large (+0.12 MAP) gain compared with previous methods, while also achieving state-of-the-art performance on benchmark open-domain and biomedical QC datasets. Finally, we show that using this model's predictions of question topic significantly improves the accuracy of a question answering system by +1.7% P@1, with substantial future gains possible as QC performance improves.

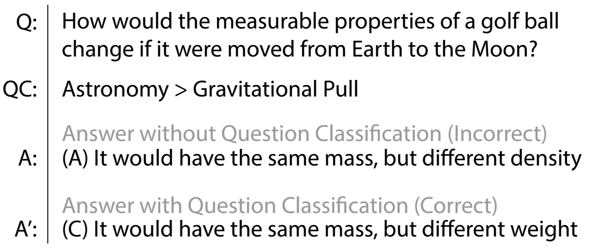
Keywords: question answering, question classification

1. Introduction

Understanding what a question is asking is one of the first steps that humans use to work towards an answer. In the context of question answering, question classification allows automated systems to intelligently target their inference systems to domain-specific solvers capable of addressing specific kinds of questions and problem solving methods with high confidence and answer accuracy (Hovy et al., 2001; Moldovan et al., 2003).

To date, question classification has primarily been studied in the context of open-domain TREC questions (Voorhees and Tice, 2000), with smaller recent datasets available in the biomedical (Roberts et al., 2014; Wasim et al., 2019) and education (Godea and Nielsen, 2018) domains. The open-domain TREC question corpus is a set of 5,952 short factoid questions paired with a taxonomy developed by Li and Roth (2002) that includes 6 coarse answer types (such as *entities*, *locations*, and *numbers*), and 50 fine-grained types (e.g. specific kinds of entities, such as *animals* or *vehicles*). While a wide variety of syntactic, semantic, and other features and classification methods have been applied to this task, culminating in near-perfect classification performance (Madabushi and Lee, 2016), recent work has demonstrated that QC methods developed on TREC questions generally fail to transfer to datasets with more complex questions such as those in the biomedical domain (Roberts et al., 2014), likely due in part to the simplicity and syntactic regularity of the questions, and the ability for simpler term-frequency models to achieve near-ceiling performance (Xia et al., 2018).

In this work we explore question classification in the context of multiple choice science exams. Standardized science exams have been proposed as a challenge task for question answering (Clark, 2015), as most questions contain a variety of challenging inference problems (Clark et al., 2013; Jansen et al., 2016), require detailed scientific and common-sense knowledge to answer and explain the reasoning behind those answers (Jansen et al., 2018), and questions are often embedded in complex examples or other distractors. Question classification taxonomies and annotation are difficult and expensive to generate, and because of the unavailability of this data, to date most models for science questions use one or a small number of generic solvers that perform little or no question decomposition (e.g. Khot et al., 2015; Clark et al., 2016; Khashabi et al., 2016; Khot et al., 2017; Jansen et al., 2017). Our long-term interest is in developing methods that intelligently target their inferences to generate both correct answers and compelling human-readable explanations for the reasoning behind those answers. The lack of targeted solving – using the same methods for inferring answers to spatial questions about planetary motion, chemical questions about photosynthesis, and electrical questions about circuit continuity – is a substantial barrier to increasing performance (see Figure 1).



Q: How would the measurable properties of a golf ball change if it were moved from Earth to the Moon?

QC: Astronomy > Gravitational Pull

Answer without Question Classification (Incorrect)
A: (A) It would have the same mass, but different density

Answer with Question Classification (Correct)
A': (C) It would have the same mass, but different weight

Figure 1: Identifying the detailed problem domain of a question (QC label) can provide an important contextual signal to guide a QA system to the correct answer (A'). Here, knowing the problem domain of *Gravitational Pull* allows the model to recognize that some properties (such as weight) change when objects move between celestial bodies, while others (including density) are unaffected by such a change.

To address this need for developing methods of targetted inference, this work makes the following contributions:

1. We provide a large challenge dataset of question classification labels for 7,787 standardized science exam questions labeled using a hierarchical taxonomy of 406 detailed problem types across 6 levels of granularity. To the best of our knowledge this is the most detailed question classification dataset constructed by nearly an order of magnitude, while also being 30% larger than TREC, and nearly three times the size of the largest biomedical dataset.

2. We empirically demonstrate large performance gains of +0.12 MAP (+13.3% P@1) on science exam question classification using a BERT-based model over five previous state-of-the art methods, while improving performance on two biomedical question datasets by 4-5%. This is the first model to show consistent state-of-the-art performance across multiple question classification datasets.

3. We show predicted question labels significantly improve a strong QA model by +1.7% P@1, where ceiling performance with perfect classification can reach +10.0% P@1. We also show that the error distribution of question classification matters when coupled with multiple choice QA models, and that controlling for correlations between classification labels and incorrect answer candidates can increase performance.

2. Related work

Question classification typically makes use of a combination of syntactic, semantic, surface, and embedding methods. Syntactic patterns (Li and Roth, 2006; Silva et al., 2011; Patrick and Li, 2012; Mishra et al., 2013) and syntactic dependencies (Roberts et al., 2014) have been shown to improve performance, while syntactically or semantically important words are often expanding using Wordnet hypernyms or Unified Medical Language System categories (for the medical domain) to help mitigate sparsity (Huang et al., 2008; Yu and Cao, 2008; Van-Tu and Anh-Cuong, 2016). Keyword identification helps identify specific terms useful for classification (Liu et al., 2011; Roberts et al., 2014; Khashabi et al., 2017). Similarly, named entity recognizers (Li and Roth, 2002; Neves and Kraus, 2016) or lists of semantically related words (Li and Roth, 2002; Van-Tu and Anh-Cuong, 2016) can also be used to establish broad topics or entity categories and mitigate sparsity, as can word embeddings (Kim, 2014; Lei et al., 2018). Here, we empirically demonstrate many of these existing methods do not transfer to the science domain.

The highest performing question classification systems tend to make use of customized rule-based pattern matching (Lally et al., 2012; Madabushi and Lee, 2016), or a combination of rule-based and machine learning approaches (Silva et al., 2011), at the expense of increased model construction time. A recent emphasis on learned methods has shown a large set of CNN (Lei et al., 2018) and LSTM (Xia et al., 2018) variants achieve similar accuracy on TREC question classification, with these models exhibiting at best small gains over simple term frequency models. These recent developments echo the observations of Roberts et al. (2014), who showed that existing methods beyond term frequency models failed to generalize to medical domain questions. Here we show that strong performance across multiple datasets is possible using a single learned model.

Measure	**TREC** Open	**GARD** Medical	**ARC** Science
Average per question:			
Words	9.1	10.3	20.5
Sentences	1.0	1.0	1.7
Clausal Dependencies	0.2	0.6	0.8
Prep. Dependencies	0.9	1.1	2.7
Total Questions	5,952	2,936	7,787
Question Categories	6 or 50	13	9 to 406

Table 1: Summary statistics comparing the surface and syntactic complexity of the TREC, GARD, and ARC datasets. ARC questions are complex, syntactically-diverse, and paired with a detailed classification scheme developed in this work.

Due to the cost involved in their construction, question classification datasets and classification taxonomies tend to be small, which can create methodological challenges. Roberts et al. (2014) generated the next-largest dataset from TREC, containing 2,936 consumer health questions classified into 13 question categories. More recently, Wasim et al. (2019) generated a small corpus of 780 biomedical domain questions organized into 88 categories. In the education domain, Godea et al. (2018) collected a set of 1,155 classroom questions and organized these into 16 categories. To enable a detailed study of science domain question classification, here we construct a large-scale challenge dataset that exceeds the size and classification specificity of other datasets, in many cases by nearly an order of magnitude.

3. Questions and Classification Taxonomy

Questions: We make use of the 7,787 science exam questions of the Aristo Reasoning Challenge (ARC) corpus (Clark et al., 2018), which contains standardized 3^{rd} to 9^{th} grade science questions from 12 US states from the past decade. Each question is a 4-choice multiple choice question. Summary statistics comparing the complexity of ARC and TREC questions are shown in Table 1.

Taxonomy: Starting with the syllabus for the NY Regents exam, we identified 9 coarse question categories *(Astronomy, Earth Science, Energy, Forces, Life Science, Matter, Safety, Scientific Method, Other)*, then through a data-driven analysis of 3 exam study guides and the 3,370 training questions, expanded the taxonomy to include 462 fine-grained categories across 6 hierarchical levels of granularity. The taxonomy is designed to allow categorizing questions into broad curriculum topics at its coarsest level, while labels at full specificity separate questions into narrow problem domains suitable for targetted inference methods. Because of its size, a subset of the classification taxonomy is shown in Table 2, with the full taxonomy and class definitions included in the supplementary material.

Prop.	Category
8.0%	**Astronomy / Celestial Events**
2.7%	Planetary/Stellar Features
2.0%	Natural Cycles and Patterns
0.7%	Planetary/Stellar Distances
0.4%	Orbits
22.4%	**Earth Science**
8.4%	Human Impacts on the Earth
6.8%	Weather
4.5%	Geology
2.4%	Outer Structure (Atmosphere/Hydrosphere)
1.3%	Inner Structure (Crust/Mantle/Core)
7.4%	**Energy**
1.6%	Properties of Light
1.5%	Converting Energy
0.9%	Electricity
0.9%	Sound Energy
0.4%	Potential/Kinetic Energy
3.5%	**Forces**
0.8%	Gravity
0.7%	Friction
0.5%	Speed/Velocity
0.4%	Mechanical Energy
0.3%	Newton's Laws
34.7%	**Life Science**
16.3%	Life Functions
13.6%	Features and their Functions
5.7%	Cellular Biology
5.3%	Animal Features and Functions
3.1%	Plant Features and Functions
1.2%	Photosynthesis
0.7%	Reproduction/Pollination
0.1%	Seed Dispersal
0.4%	Leaves
0.3%	Roots
1.2%	Environmental Effects on Development
0.8%	Responses to Environment Changes
0.8%	Basic Life Functions
6.3%	Interdependence/Food Chains
4.7%	Reproduction
3.3%	Adaptations and the Environment
1.4%	Continuity of Life/Life Cycle
17.0%	**Matter**
5.0%	Chemistry
2.2%	Measurement
2.4%	Changes of State
2.5%	Properties of Materials
1.8%	Physical vs Chemical Changes
1.4%	Mixtures
1.1%	**Safety**
0.7%	Safety Procedures
0.4%	Safety Equipment
7.6%	**Scientific Method**
5.8%	Components of Inference
0.9%	Graphing Data
0.6%	Scientific Models
3.3%	**Other**
1.6%	History of Science

Table 2: A subset (approximately 10%) of our question classification taxonomy for science exams, with top-level categories in bold. The full taxonomy contains 462 categories, with 406 of these having non-zero counts in the ARC corpus. *"Prop."* represents the proportion of questions in ARC belonging to a given category. One branch of the taxonomy *(Life Science → ... → Seed Dispersal)* has been expanded to full depth.

Annotation: Because of the complexity of the questions, it is possible for one question to bridge multiple categories – for example, a wind power generation question may span both *renewable energy* and *energy conversion*. We allow up to 2 labels per question, and found that 16% of questions required multiple labels. Each question was independently annotated by two annotators, with the lead annotator a domain expert in standardized exams. Annotators first independently annotated the entire question set, then questions without complete agreement were discussed until resolution. Before resolution, interannotator agreement (Cohen's Kappa) was $\kappa = 0.58$ at the finest level of granularity, and $\kappa = 0.85$ when considering only the coarsest 9 categories. This is considered moderate to strong agreement (McHugh, 2012). Based on the results of our error analysis (see Section 4.3.), we estimate the overall accuracy of the question classification labels after resolution to be approximately 96%. While the full taxonomy contains 462 fine-grained categories derived from both standardized questions, study guides, and exam syllabi, we observed only 406 of these categories are tested in the ARC question set.

4. Question Classification Models

4.1. Question Classification on Science Exams

We identified 5 common models in previous work primarily intended for learned classifiers rather than hand-crafted rules. We adapt these models to a multi-label hierarchical classification task by training a series of one-vs-all binary classifiers (Tsoumakas and Katakis, 2007), one for each label in the taxonomy. With the exception of the CNN and BERT models, following previous work (e.g. Silva et al., 2011; Roberts et al., 2014; Xia et al., 2018) we make use of an SVM classifier using the LIBSvM framework (Chang and Lin, 2011) with a linear kernel. Models are trained and evaluated from coarse to fine levels of taxonomic specificity. At each level of taxonomic evaluation, a set of non-overlapping confidence scores for each binary classifier are generated and sorted to produce a list of ranked label predictions. We evaluate these ranks using Mean Average Precision (see Manning et al., 2008). ARC questions are evaluated using the standard 3,370 questions for training, 869 for development, and 3,548 for testing, as in Clark et al. (2018).

N-grams, POS, Hierarchical features: A baseline bag-of-words model incorporating both tagged and untagged unigrams and bigams. We also implement the hierarchical classification feature of Li and Roth (Li and Roth, 2002), where for a given question, the output of the classifier at coarser levels of granularity serves as input to the classifier at the current level of granularity.

Dependencies: Bigrams of Stanford dependencies (De Marneffe and Manning, 2008). For each word, we create one unlabeled bigram for each outgoing link from that word to its dependency (Patrick and Li, 2012; Roberts et al., 2014).

Question Expansion with Hypernyms: We perform hypernym expansion (Huang et al., 2008; Silva et al., 2011; Roberts et al., 2014) by including WordNet hypernyms (Fellbaum, 1998) for the root dependency word, and words on

		ARC Science Exams						
Adapted From	Model	L1	L2	L3	L4	L5	L6	Gain (L6)
	Unigram Model	0.885	0.714	0.602	0.535	0.503	0.490	
Li and Roth (2002)	Uni+Bi+POS+Hier *(UBPH)*	0.903	0.759	0.644	0.582	0.549	0.535	Baseline
Van-tu et al. (2016)	UBPH+WordNet Expansion	0.901	0.755	0.645	0.582	0.552	0.535	–
Roberts et al. (2014)	UBPH+Dependencies	0.906	0.760	0.645	0.583	0.549	0.536	–
He et al. (2015)	MP-CNN	0.908	0.757	0.654	0.597	0.563	0.532	–
Khashabi et al. (2017)	UBPH+Essential Terms	0.913	0.774	0.666	0.607	0.575	0.564	+0.03*
This Work	BERT-QC	**0.942**	**0.841**	**0.745**	**0.684**	**0.664**	**0.654**	**+0.12***
Number of Categories		*9*	*88*	*243*	*335*	*379*	*406*	

Table 3: Results of the empirical evaluation on each of the question classification models on the ARC science question dataset, broken down by classification granularity (*coarse (L1)* to *fine (L6)*). Performance reflects mean average precision (MAP), where a duplicate table showing P@1 is included in the appendix. The best model at each level of granularity is shown in bold. * signifies that a given model is significantly better than baseline performance at full granularity ($p < 0.01$).

		TREC	
Model	**Desc.**	**Coarse**	**Fine**
Learned Models			
Li and Roth (2002)	SNoW	91.0%	84.2%
Kim (2014)	CNN	93.6%	–
Xia et al. (2018)	TF-IDF	94.8%	–
Van-tu et al. (2016)	SVM	95.2%	91.6%
Xia et al. (2018)	LSTM	95.8%	–
Lei et al. (2018)	RR-CNN	95.8%	–
This Work	BERT-QC	96.2%	**92.0%**
Xia et al. (2018)	Att-LSTM	**98.0%**	–
Rule Based Models			
da Silva et al. (2011)	Rules	95.0%	90.8%
Madabushi et al. (2016)	Rules	–	97.2%
Number of Categories		*6*	*50*

Table 4: Performance of BERT-QC on the TREC-6 (6 coarse categories) and TREC-50 (50 fine-grained categories) question classification task, in context with recent learned or rule-based models. Performance is reported as classification accuracy. Bold values represent top reported learned model performance. BERT-QC achieves performance similar to or exceeding the top reported non-rule-based models.

its direct outgoing links. WordNet sense is identified using Lesk word-sense disambiguation (Lesk, 1986), using question text for context. We implement the heuristic of Van-tu et al. (2016), where more distant hypernyms receive less weight.

Essential Terms: Though not previously reported for QC, we make use of unigrams of keywords extracted using the Science Exam Essential Term Extractor of Khashabi et al. (2017). For each keyword, we create one binary unigram feature.

CNN: Kim (2014) demonstrated near state-of-the-art performance on a number of sentence classification tasks (including TREC question classification) by using pre-trained word embeddings (Mikolov et al., 2013) as feature extractors in a CNN model. Lei et al. (2018) showed that 10 CNN variants perform within +/-2% of Kim's (2014) model on TREC QC. We report performance of our best CNN model based on the MP-CNN architecture[1] of Rao et al. (Rao et al., 2016), which works to establish the similarity between question text and the definition text of the question classes. We adapt the MP-CNN model, which uses a "Siamese" structure (He et al., 2015), to create separate representations for both the question and the question class. The model then makes use of a triple ranking loss function to minimize the distance between the representations of questions and the correct class while simultaneously maximising the distance between questions and incorrect classes. We optimize the network using the method of Tu (2018).

BERT-QC (This work): We make use of BERT (Devlin et al., 2018), a language model using bidirectional encoder representations from transformers, in a sentence-classification configuration. As the original settings of BERT do not support multi-label classification scenarios, and training a series of 406 binary classifiers would be computationally expensive, we use the duplication method of Tsoumakas et al. (2007) where we enumerate multi-label questions as multiple single-label instances during training by duplicating question text, and assigning each instance one of the multiple labels. Evaluation follows the standard procedure where we generate a list of ranked class predictions based on class probabilities, and use this to calculate Mean Average Precision (MAP) and Precision@1 (P@1). As shown in Table 3, this BERT-QC model achieves our best question classification performance, significantly exceeding baseline performance on ARC by 0.12 MAP and 13.3% P@1.[2]

4.2. Comparison with Benchmark Datasets

Roberts et al. (2014) observed that, apart from term frequency methods, question classification methods developed on one dataset generally do not exhibit strong transfer performance to other datasets. While BERT-QC achieves large gains over existing methods on the ARC dataset, here we demonstrate that BERT-QC also matches state-of-the-art performance on TREC (Li and Roth, 2002), while surpassing state-of-the-art performance on the GARD corpus of

[1] `https://github.com/castorini/Castor`

[2] Question classification performance evaluated using Precision@1 is reported in Table 11 (see `Appendix`)

Model	Desc.	Accuracy
Learned Models		
Roberts et al. (2014)	Bag of Words	76.9%
Roberts et al. (2014)	CQT2/SVM	80.4%
This Work	BERT-QC	**84.9%**

Table 5: Performance of BERT-QC on the GARD consumer health question dataset, which contains 2,937 questions labeled with 13 medical question classification categories. Following Roberts et al. (2014), this dataset was evaluated using 5-fold crossvalidation.

consumer health questions (Roberts et al., 2014) and MLBioMedLAT corpus of biomedical questions (Wasim et al., 2019). As such, BERT-QC is the first model to achieve strong performance across more than one question classification dataset.

4.2.1. TREC Question Classification

TREC question classification[3] is divided into separate coarse and fine-grained tasks centered around inferring the expected answer types of short open-domain factoid questions. TREC-6 includes 6 coarse question classes (*abbreviation, entity, description, human, location, numeric*), while TREC-50 expands these into 50 more fine-grained types.

TREC question classification methods can be divided into those that learn the question classification task, and those that make use of either hand-crafted or semi-automated syntactic or semantic extraction rules to infer question classes. To date, the best reported accuracy for learned methods is 98.0% by Xia et al. (2018) for TREC-6, and 91.6% by Van-tu et al. (Van-Tu and Anh-Cuong, 2016) for TREC-50[4]. Madabushi et al. (2016) achieve the highest to-date performance on TREC-50 at 97.2%, using rules that leverage the strong syntactic regularities in the short TREC factoid questions.

We compare the performance of BERT-QC with recently reported performance on this dataset in Table 4. BERT-QC achieves state-of-the-art performance on fine-grained classification (TREC-50) for a learned model at 92.0% accuracy, and near state-of-the-art performance on coarse classification (TREC-6) at 96.2% accuracy.[5]

4.2.2. Medical Question Classification

Because of the challenges with collecting biomedical questions, the datasets and classification taxonomies tend to be small, and rule-based methods often achieve strong results (e.g. Sarrouti et al., 2015). Roberts et al. (2014) created the largest biomedical question classification dataset to date, annotating 2,937 consumer health questions drawn from the Genetic and Rare Diseases (GARD) question database with 13 question types, such as *anatomy, disease cause, diagnosis, disease management,* and *prognoses*. Roberts et al. (2014) found these questions largely resistant to learning-based methods developed for TREC questions. Their best model (CPT2), shown in Table 5, makes use of stemming and lists of semantically related words and cue phrases to achieve 80.4% accuracy. BERT-QC reaches 84.9% accuracy on this dataset, an increase of +4.5% over the best previous model. We also compare performance on the recently released MLBioMedLAT dataset (Wasim et al., 2019), a multi-label biomedical question classification dataset with 780 questions labeled using 88 classification types drawn from 133 Unified Medical Language System (UMLS) categories. Table 6 shows BERT-QC exceeds their best model, focus-driven semantic features (FDSF), by +0.05 Micro-F1 and +3% accuracy.

Model	Desc.	μF1	Accuracy
Learned Models			
Wasim et al. (2019)	SSVM	0.42	0.37
Wasim et al. (2019)	LPLR	0.47	0.42
Wasim et al. (2019)	FDSF	0.50	0.45
This Work	BERT-QC	**0.55**	**0.48**

Table 6: Performance of BERT-QC on the MLBioMedLAT biomedical question dataset, which contains 780 questions labeled with 88 medical question classification categories. Following Wasim et al. (2019), this dataset was evaluated using 10-fold crossvalidation. Micro-F1 (μF1) and Accuracy follow Wasim et al.'s definitions for multi-label tasks.

4.3. Error Analysis

We performed an error analysis on 50 ARC questions where the BERT-QC system did not predict the correct label, with a summary of major error categories listed in Table 7.

Associative Errors: In 35% of cases, predicted labels were nearly correct, differing from the correct label only by the finest-grained (leaf) element of the hierarchical label (for example, predicting *Matter* $\rightarrow$ *Changes of State* $\rightarrow$ *Boiling* instead of *Matter* $\rightarrow$ *Changes of State* $\rightarrow$ *Freezing*). The bulk of the remaining errors were due to questions containing highly correlated words with a different class, or classes themselves being highly correlated. For example, a specific question about *Weather Models* discusses "environments" changing over "millions of years", where discussions of environments and long time periods tend to be associated with questions about *Locations of Fossils*. Similarly, a question containing the word "evaporation" could be primarily focused on either *Changes of State* or the *Water Cycle* (cloud generation), and must rely on knowledge from the entire question text to determine the correct problem domain. We believe these associative errors are addressable technical challenges that could ultimately lead to increased performance in subsequent models.

Errors specific to the multiple-choice domain: We observed that using both question and all multiple choice answer text produced large gains in question classification performance – for example, BERT-QC performance increases from 0.516 (question only) to 0.654 (question and all four answer candidates), an increase of 0.138 MAP. Our error

[3] `http://cogcomp.org/Data/QA/QC/`

[4] Model performance is occasionally reported only on TREC-6 rather than the more challenging TREC-50, making direct comparisons between some algorithms difficult.

[5] Xia et al. (2018) also report QC performance on MS Marco (Nguyen et al., 2016), a million-question dataset using 5 of the TREC-6 labels. We believe this to be in error as MS Marco QC labels are automatically generated. Still, for purposes of comparison, BERT-QC reaches 96.2% accuracy, an increase of +3% over Xia et al. (2018)'s best model.

Proportion	Error Type
46%	Question contains words correlated with incorrect class
35%	Predicted class is nearly correct, and distance 1 from gold class (different leaf node selected in taxonomy)
25%	Predicted class is highly correlated with an incorrect multiple choice answer
18%	Predicted class and gold class are on different aspects of similar topics/otherwise correlated
10%	Annotation: Gold label appears incorrect, predicted label is good.
8%	Annotation: Predicted label is good, but not in gold list.
8%	Correctly predicting the gold label may require knowing the correct answer to the question.

Table 7: BERT-QC Error Analysis: Classes of errors for 50 randomly selected questions from the development set where BERT-QC did not predict the correct label. These errors reflect the BERT-QC model trained and evaluated with terms from both the question and all multiple choice answer candidates. Questions can occupy more than one error category, and as such proportions do not sum to 100%.

analysis observed that while this substantially increases QC performance, it changes the *distribution of errors* made by the system. Specifically, 25% of errors become highly correlated with an incorrect answer candidate, which (we show in Section 5.) can reduce the performance of QA solvers.[6]

5. Question Answering with QC Labels

Because of the challenges of errorful label predictions correlating with incorrect answers, it is difficult to determine the ultimate benefit a QA model might receive from reporting QC performance in isolation. Coupling QA and QC systems can often be laborious – either a large number of independent solvers targeted to specific question types must be constructed (e.g. Minsky, 1986), or an existing single model must be able to productively incorporate question classification information. Here we demostrate the latter – that a BERT QA model is able to incorporate question classification information through query expansion.

BERT (Devlin et al., 2018) recently demonstrated state-of-the-art performance on benchmark question answering datasets such as SQUaD (Rajpurkar et al., 2016), and near human-level performance on SWAG (Zellers et al., 2018). Similarly, Pan et al. (2019) demonstrated that BERT achieves the highest accuracy on the most challenging subset of ARC science questions. We make use of a BERT QA model using the same QA paradigm described by Pan et al. (2019), where QA is modeled as a next-sentence prediction task that predicts the likelihood of a given multiple choice answer candidate following the question text. We evaluate the question text and the text of each multiple choice answer candidate separately, where the answer candidate with the highest probablity is selected as the predicted answer for a given question. Performance is evaluated using Precision@1 (Manning et al., 2008). Additional model details and hyperparameters are included in the *Appendix*.

Original Question Text
What happens to water molecules during the boiling process?
Expanded Text (QC Label)
Matter Changes of State Boiling What happens to water molecules during the boiling process?

Table 8: An example of the query expansion technique for question classification labels, where the definition text for the QC label is appended to the question. Here, the gold label for this question is "MAT_COS_BOILING" *(Matter → Changes of State → Boiling)*.

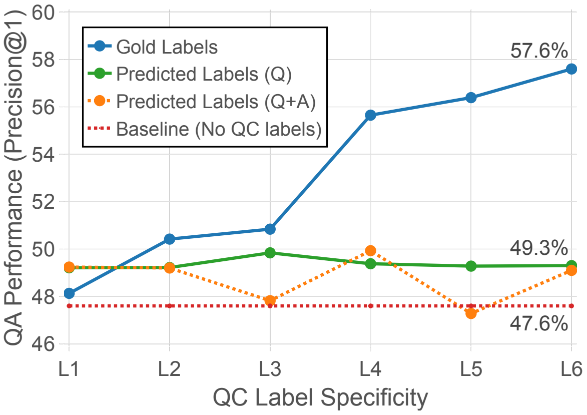


Figure 2: Question answering performance (the proportion of questions answered correctly) for models that include question classification labels using query expansion, compared to a no-label baseline model. While BERT-QC trained using question and answer text achieves higher QC performance, it leads to unstable QA performance due to its errors being highly correlated with incorrect answers. Predicted labels using BERT-QC (question text only) show a significant increase of +1.7% P@1 at L6 ($p < 0.01$). Models with gold labels show the ceiling performance of this approach with perfect question classification performance. Each point represents the average of 10 runs.

We incorporate QC information into the QA process by implementing a variant of a query expansion model (Qiu and Frei, 1993). Specifically, for a given $\{question,\ QC\_label\}$ pair, we expand the question text by concatenating the definition text of the question classification label to the start of the question. We use of the top predicted question classification label for each question. Because QC labels are hierarchical, we append the label definition text for each level of the label $L_1 ... L_n$. An exampe of this process is shown in Table 8.

Figure 2 shows QA peformance using predicted labels from the BERT-QC model, compared to a baseline model that does not contain question classification information. As predicted by the error analysis, while a model trained with question and answer candidate text performs better at QC than a model using question text alone, a large proportion of the incorrect predictions become associated with a negative answer candidate, reducing overall QA performance,

[6] When a model is trained using only question text (instead of both question and answer candidate text), the distribution of these highly-correlated errors changes to the following: 17% chose the correct label, 17% chose the same label, and 66% chose a different label not correlated with an incorrect answer candidate.

and highlighting the importance of evaluating QC and QA models together. When using BERT-QC trained on question text alone, at the finest level of specificity (L6) where overall question classification accuracy is 57.8% P@1, question classification significantly improves QA performance by +1.7% P@1 ($p < 0.01$). Using gold labels shows ceiling QA performance can reach +10.0% P@1 over baseline, demonstrating that as question classification model performance improves, substantial future gains are possible. An analysis of expected gains for a given level of QC performance is included in the *Appendix*, showing approximately linear gains in QA performance above baseline for QC systems able to achieve over 40% classification accuracy. Below this level, the decreased performance from noise induced by incorrect labels surpasses gains from correct labels.

5.1. Automating Error Analyses with QC

Detailed error analyses for question answering systems are typically labor intensive, often requiring hours or days to perform manually. As a result error analyses are typically completed infrequently, in spite of their utility to key decisions in the algortithm or knowledge construction process. Here we show having access to detailed question classification labels specifying fine-grained problem domains provides a mechanism to automatically generate error analyses in seconds instead of days.

To illustrate the utility of this approach, Table 9 shows the performance of the BERT QA+QC model broken down by specific question classes. This allows automatically identifying a given model's strengths – for example, here questions about *Human Health*, *Material Properties*, and *Earth's Inner Core* are well addressed by the BERT-QA model, and achieve well above the average QA performance of 49%. Similarly, areas of deficit include *Changes of State*, *Reproduction*, and *Food Chain Processes* questions, which see below-average QA performance. The lowest performing class, *Safety Procedures*, demonstrates that while this model has strong performance in many areas of scientific reasoning, it is *worse than chance* at answering questions about safety, and would be inappropriate to deploy for safety-critical tasks.

While this analysis is shown at an intermediate (L2) level of specificity for space, more detailed analyses are possible. For example, overall QA performance on *Scientific Inference* questions is near average (47%), but increasing granularity to L3 we observe that questions addressing *Experiment Design* or *Making Inferences* – challenging questions even for humans – perform poorly (33% and 20%) when answered by the QA system. This allows a system designer to intelligently target problem-specific knowledge resources and inference methods to address deficits in specific areas.

Question Category	**QA Accuracy**	**N**
Strong Performance		
Life - Human Health	73%	11
Forces - Friction	71%	7
Energy - Sound	71%	7
Energy - Light	70%	10
Matter - Material Properties	68%	22
Matter - Object Properties	66%	9
Forces - Gravity	66%	9
Science - Scientific Models	66%	9
Earth - Inner Core	64%	33
Astronomy - Natural Cycles	64%	11
Energy - Device Use	63%	8
Energy - Waves	63%	8
Earth - Weather	62%	74
Energy - Conversion	62%	13
Energy - Thermal	60%	10
Above Average Performance		
Earth - Geology	58%	38
Science - Graphs	56%	9
Life - Environmental Adaptations	53%	32
Matter - Phys./Chemical Changes	53%	17
Astronomy - Features	52%	27
Astronomy - Tides	50%	6
Approximately Average Performance		
Earth - Human Impacts	51%	84
Matter - Chemistry	50%	46
Other - History of Science	50%	10
Life - Features and Functions	49%	176
Below Average Performance		
Science - Scientific Inference	47%	58
Life - Food Chains	44%	54
Astronomy - Celestial Distances	44%	9
Life - Reproduction	41%	41
Energy - Electrical	40%	5
Life - Classification	38%	13
Matter - Changes of State	29%	21
Below Chance Performance		
Earth - Outer Core	17%	12
Safety - Safety Procedures	7%	14

Table 9: Analysis of question answering performance on specific question classes on the BERT-QA model (L6). Question classes in this table are at the L2 level of specificity. Performance is reported on the development set, where N represents the total number of questions within a given question class.

6. Conclusion

Question classification can enable targetting question answering models, but is challenging to implement with high performance without using rule-based methods. In this work we generate the most fine-grained challenge dataset for question classification, using complex and syntactically diverse questions, and show gains of up to 12% are possible with our question classification model across datasets in open, science, and medical domains. This model is the first demonstration of a question classification model achieving state-of-the-art results across benchmark datasets in open, science, and medical domains. We further demonstrate attending to question type can significantly improve question answering performance, with large gains possible as quesion classification performance improves. Our error analysis suggests that developing high-precision methods of question classification independent of their recall can offer the opportunity to incrementally make use of the benefits of question classification without suffering the consequences of classification errors on QA performance.

7. Resources

Our Appendix and supplementary material (available at `http://www.cognitiveai.org/explanationbank/`) includes data, code, experiment details, and negative results.

8. Acknowledgements

The authors wish to thank Elizabeth Wainwright and Stephen Marmorstein for piloting an earlier version of the question classification annotation. We thank the Allen Insitute for Artificial Intelligence and National Science Founation (NSF 1815948 to PJ) for funding this work.

9. Bibliographical References

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27.

Clark, P., Harrison, P., and Balasubramanian, N. (2013). A study of the knowledge base requirements for passing an elementary science test. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC'13, pages 37–42.

Clark, P., Etzioni, O., Khot, T., Sabharwal, A., Tafjord, O., Turney, P. D., and Khashabi, D. (2016). Combining retrieval, statistics, and inference to answer elementary science questions. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 2580–2586.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Clark, P. (2015). Elementary school science and math tests as a driver for AI: take the aristo challenge! In Blai Bonet et al., editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 4019–4021. AAAI Press.

De Marneffe, M.-C. and Manning, C. D. (2008). The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

Godea, A. and Nielsen, R. (2018). Annotating educational questions for student response analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

He, H., Gimpel, K., and Lin, J. (2015). Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1576–1586.

Hovy, E., Gerber, L., Hermjakob, U., Lin, C.-Y., and Ravichandran, D. (2001). Toward semantics-based answer pinpointing. In *Proceedings of the first international conference on Human language technology research*, pages 1–7. Association for Computational Linguistics.

Huang, Z., Thint, M., and Qin, Z. (2008). Question classification using head words and their hypernyms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 927–936. Association for Computational Linguistics.

Jansen, P., Surdeanu, M., and Clark, P. (2014). Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jansen, P., Balasubramanian, N., Surdeanu, M., and Clark, P. (2016). What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2956–2965, Osaka, Japan, December.

Jansen, P., Sharp, R., Surdeanu, M., and Clark, P. (2017). Framing qa as building and ranking intersentence answer justifications. *Computational Linguistics*.

Jansen, P., Wainwright, E., Marmorstein, S., and Morrison, C. (2018). Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Khashabi, D., Khot, T., Sabharwal, A., Clark, P., Etzioni, O., and Roth, D. (2016). Question answering via integer programming over semi-structured knowledge. In *Proceedings of the International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 1145–1152.

Khashabi, D., Khot, T., Sabharwal, A., and Roth, D. (2017). Learning what is essential in questions. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 80–89.

Khot, T., Balasubramanian, N., Gribkoff, E., Sabharwal, A., Clark, P., and Etzioni, O. (2015). Exploring markov logic networks for question answering. In *EMNLP*.

Khot, T., Sabharwal, A., and Clark, P. (2017). Answering complex questions using open information extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 311–316.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1746–1751.

Lally, A., Prager, J. M., McCord, M. C., Boguraev, B. K., Patwardhan, S., Fan, J., Fodor, P., and Chu-Carroll, J. (2012). Question analysis: How watson reads a clue. *IBM Journal of Research and Development*, 56(3.4):2–1.

Lei, T., Shi, Z., Liu, D., Yang, L., and Zhu, F. (2018). A novel cnn-based method for question classification in intelligent question answering. In *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, page 54. ACM.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM.

Li, X. and Roth, D. (2002). Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Li, X. and Roth, D. (2006). Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(3):229–249.

Liu, F., Antieau, L. D., and Yu, H. (2011). Toward automated consumer question answering: Automatically separating consumer questions from professional questions in the healthcare domain. *Journal of biomedical informatics*, 44(6):1032–1038.

Madabushi, H. T. and Lee, M. (2016). High accuracy rule-based question classification using question syntax and semantics. In *COLING*, pages 1220–1230.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Minsky, M. (1986). *The Society of Mind*. Simon & Schuster, Inc., New York, NY, USA.

Mishra, M., Mishra, V. K., and Sharma, H. (2013). Question classification using semantic, syntactic and lexical features. *International Journal of Web & Semantic Technology*, 4(3):39.

Moldovan, D., Paşca, M., Harabagiu, S., and Surdeanu, M. (2003). Performance issues and error analysis in an open-domain question answering system. *ACM Trans. Inf. Syst.*, 21(2):133–154, April.

Neves, M. and Kraus, M. (2016). Biomedlat corpus: Annotation of the lexical answer type for biomedical questions. *OKBQA 2016*, page 49.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Pan, X., Sun, K., Yu, D., Ji, H., and Yu, D. (2019). Improving question answering with external knowledge. *arXiv preprint arXiv:1902.00993*.

Patrick, J. and Li, M. (2012). An ontology for clinical questions about the contents of patient notes. *Journal of Biomedical Informatics*, 45(2):292–306.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Qiu, Y. and Frei, H.-P. (1993). Concept based query expansion. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 160–169. ACM.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Rao, J., He, H., and Lin, J. (2016). Noise-contrastive estimation for answer selection with deep neural networks. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1913–1916. ACM.

Roberts, K., Kilicoglu, H., Fiszman, M., and Demner-Fushman, D. (2014). Automatically classifying question types for consumer health questions. In *AMIA Annual Symposium Proceedings*, volume 2014, page 1018. American Medical Informatics Association.

Sarrouti, M., Lachkar, A., and Ouatik, S. E. A. (2015). Biomedical question types classification using syntactic and rule based approach. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, volume 1, pages 265–272. IEEE.

Silva, J., Coheur, L., Mendes, A. C., and Wichert, A. (2011). From symbolic to sub-symbolic information in question classification. *Artificial Intelligence Review*, 35(2):137–154.

Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Tu, Z. (2018). An experimental analysis of multi-perspective convolutional neural networks. *University of Waterloo Master's Thesis*.

Van-Tu, N. and Anh-Cuong, L. (2016). Improving question classification by feature extraction and selection. *Indian Journal of Science and Technology*, 9(17).

Voorhees, E. M. and Tice, D. M. (2000). Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207. ACM.

Wasim, M., Asim, M. N., Khan, M. U. G., and Mahmood, W. (2019). Multi-label biomedical question classification for lexical answer type prediction. *Journal of biomedical informatics*, page 103143.

Wu, W., Li, H., Wang, H., and Zhu, K. Q. (2012). Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 481–492. ACM.

Xia, W., Zhu, W., Liao, B., Chen, M., Cai, L., and Huang, L. (2018). Novel architecture for long short-term memory used in question classification. *Neurocomputing*, 299:20–31.

Yu, H. and Cao, Y.-g. (2008). Automatically extracting information needs from ad hoc clinical questions. In *AMIA annual symposium proceedings*, volume 2008, page 96. American Medical Informatics Association.

Zellers, R., Bisk, Y., Schwartz, R., and Choi, Y. (2018). Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.

10. Appendix

10.1. Annotation

Classification Taxonomy: The full classification taxonomy is included in separate files, both coupled with definitions, and as a graphical visualization.

Annotation Procedure: Primary annotation took place over approximately 8 weeks. Annotators were instructed to provide up to 2 labels from the full classification taxonomy (462 labels) that were appropriate for each question, and to provide the most specific label available in the taxonomy for a given question. Of the 462 labels in the classification taxonomy, the ARC questions had non-zero counts in 406 question types. Rarely, questions were encountered by annotators that did not clearly fit into a label at the end of the taxonomy, and in these cases the annotators were instructed to choose a more generic label higher up the taxonomy that was appropriate. This occurred when the production taxonomy failed to have specific categories for infrequent questions testing knowledge that is not a standard part of the science curriculum. For example, the question:

Which material is the best natural resource to use for making water-resistant shoes? (A) cotton (B) leather (C) plastic (D) wool

tests a student's knowledge of the water resistance of different materials. Because this is not a standard part of the curriculum, and wasn't identified as a common topic in the training questions, the annotators tag this question as belonging to *Matter → Properties of Materials*, rather than a more specific category.

Questions from the training, development, and test sets were randomly shuffled to counterbalance any learning effects during the annotation procedure, but were presented in grade order (3^{rd} to 9^{th} grade) to reduce context switching (a given grade level tends to use a similar subset of the taxonomy – for example, 3^{rd} grade questions generally do not address *Chemical Equations* or *Newtons* 1^{st} *Law of Motion*).

Interannotator Agreement: To increase quality and consistency, each annotator annotated the entire dataset of 7,787 questions. Two annotators were used, with the lead annotator possessing previous professional domain expertise. Annotation proceeded in a two-stage process, where in stage 1 annotators completed their annotation independently, and in stage 2 each of the questions where the annotators did not have complete agreement were manually resolved by the annotators, resulting in high-quality classification annotation.

Because each question can have up to two labels, we treat each label for a given question as a separate evaluation of interannotator agreement. That is, for questions where both annotators labeled each question as having 1 or 2 labels, we treat this as 1 or 2 separate evaluations of interannotator agreement. For cases where one annotator labeled as question as having 1 label, and the other annotator labeled that same question as having 2 labels, we conservatively treat this as two separate interannotator agreements where one annotator failed to specify the second label and had zero agreement on that unspecified label.

Classification Level	# of Classes	Interannotator Agreement (κ)
L1 (Coarsest)	9	0.85
L2	88	0.71
L3	243	0.64
L4	335	0.60
L5	379	0.58
L6 (Finest)	406	0.58

Table 10: Interannotator Agreement at L6 (the native level the annotation was completed at), as well as agreement for truncated levels of the heirarchy from coarse to fine classification.

Though the classification procedure was fine-grained compared to other question classification taxonomies, containing an unusually large number of classes (406), overall raw interannotator agreement before resolution was high (Cohen's κ = 0.58). When labels are truncated to a maximum taxonomy depth of N, raw interannotator increases to κ = 0.85 at the coarsest (9 class) level (see Table 10). This is considered moderate to strong agreement (see McHugh (2012) for a discussion of the interpretation of the Kappa statistic). Based on the results of an error analysis on the question classification system (see Section 10.3.2.), we estimate that the overall accuracy of the question classification labels after resolution is approximately 96% .

Annotators disagreed on 3441 (44.2%) of questions. Primary sources of disagreement before resolution included each annotator choosing a single category for questions requiring multiple labels (e.g. annotator 1 assigning a label of X, and annotator 2 assigning a label of Y, when the gold label was multilabel X, Y), which was observed in 18% of disagreements. Similarly, we observed annotators choosing similar labels but at different levels of specificity in the taxonomy (e.g. annotator 1 assigning a label of *Matter → Changes of State → Boiling*, where annotator 2 assigned *Matter → Changes of State*), which occurred in 12% of disagreements before resolution.

10.2. Question Classification

10.2.1. Precision@1

Because of space limitations the question classification results are reported in Table 3 only using Mean Average Precision (MAP). We also include Precision@1 (P@1), the overall accuracy of the highest-ranked prediction for each question classification model, in Table 11.

10.2.2. Negative Results

CNN: We implemented the CNN sentence classifier of Kim (2014), which demonstrated near state-of-the-art performance on a number of sentence classification tasks (including TREC question classification) by using pre-trained word embeddings (Mikolov et al., 2013) as feature extractors in a CNN model. We adapted the original CNN non-static model to multi-label classification by changing the fully connected softmax layer to sigmoid layer to produce a sigmoid output for each label simultaneously. We followed the same parameter settings reported by Kim et al. except the learning rate, which was tuned based on the development set. Pilot experiments did not show a performance improvement over the baseline model.

		ARC Science Exams						
Adapted From	Model	L1	L2	L3	L4	L5	L6	Gain (L6)
	Unigram Model	82.2	62.1	51.8	44.2	40.8	39.6	
Li and Roth (2002)	Uni+Bi+POS+Hier *(UBPH)*	84.2	67.6	56.6	49.4	46.5	44.5	Baseline
Van-tu et al. (2016)	UBPH+WordNet Expansion	84.1	67.1	56.4	49.3	46.4	44.7	+0.2
Roberts et al. (2014)	UBPH+Dependencies	84.7	68.0	56.5	49.2	45.6	44.8	+0.3
He et al. (2015)	MP-CNN	84.8	66.3	56.3	50.7	46.6	43.5	–
Khashabi et al. (2017)	UBPH+Essential Terms	85.9	69.4	58.7	51.9	48.4	48.0	+3.5*
This Work	BERT-QC	**90.2**	**78.2**	**67.6**	**60.6**	**58.9**	**57.8**	**+13.3***
Number of Categories		*9*	*88*	*243*	*335*	*379*	*406*	

Table 11: Performance of each question classification model, expressed in Precision@1 (P@1). * signifies a given model is significantly different from the baseline model ($p < 0.01$).

Label Definitions: Question terms can be mapped to categories using manual heuristics (e.g. Silva et al., 2011). To mitigate sparsity and limit heuristic use, here we generated a feature comparing the cosine similarity of composite embedding vectors (e.g. Jansen et al., 2014) representing question text and category definition text, using pretrained GloVe embeddings (Pennington et al., 2014). Pilot experiments showed that performance did not significantly improve.

Question Expansion with Hypernyms (Probase Version): One of the challenges of hypernym expansion (e.g. Huang et al., 2008; Silva et al., 2011; Roberts et al., 2014) is determining a heuristic for the termination depth of hypernym expansion, as in Van-tu et al. (2016). Because science exam questions are often grounded in specific examples (e.g. a car rolling down a hill coming to a stop due to friction), we hypothesized that knowing certain categories of entities can be important for identifying specific question types – for example, observing that a question contains a kind of *animal* may be suggestive of a *Life Science* question, where similarly *vehicles* or *materials* present in questions may suggest questions about *Forces* or *Matter*, respectively. The challenge with WordNet is that key hypernyms can be at very different depths from query terms – for example, *"cat"* is distance 10 away from `living thing`, *"car"* is distance 4 away from `vehicle`, and *"copper"* is distance 2 away from `material`. Choosing a static threshold (or decaying threshold, as in Van-tu et al. (2016)) will inheriently reduce recall and limit the utility of this method of query expansion.

To address this, we piloted a hypernym expansion experiment using the Probase taxonomy (Wu et al., 2012), a collection of 20.7M `is-a` pairs mined from the web, in place of WordNet. Because the taxonomic pairs in Probase come from use in naturalistic settings, links tend to jump levels in the WordNet taxonomy and be expressed in common forms. For example, $cat \rightarrow animal$, $car \rightarrow vehicle$, and $copper \rightarrow material$, are each distance 1 in the Probase taxonomy, and high-frequency (i.e. high-confidence) taxonomic pairs.

Similar to query expansion using WordNet Hypernyms, our pilot experiments did not observe a benefit to using Probase hypernyms over the baseline model. An error analysis suggested that the large number of noisy and out-of-context links present in Probase may have reduced performance, and in response we constructed a filtered list of 710 key hypernym categories manually filtered from a list of hypernyms seeded using high-frequency words from an in-house corpus of 250 in-domain science textbooks. We also did not observe a benefit to question classification over the baseline model when expanding only to this manually curated list of key hypernyms.

10.2.3. Additional Positive Results

Topic words: We made use of the 77 TREC word lists of Li and Roth (2002), containing a total of 3,257 terms, as well as an in-house set of 144 word lists on general and elementary science topics mined from the web, such as *ANIMALS*, *VEGETABLES*, and *VEHICLES*, containing a total of 29,059 words. To mitigate sparsity, features take the form of counts for a specific topic – detecting the words *turtle* and *giraffe* in a question would provide a count of 2 for the *ANIMAL* feature. This provides a light form of domain-specific entity and action (e.g. types of *changes*) recognition. Pilot experiments showed that this wordlist feature did add a modest performance benefit of approximately 2% to question classification accuracy. Taken together with our results on hypernym expansion, this suggests that manually curated wordlists can show modest benefits for question classification performance, but at the expense of substantial effort in authoring or collecting these extensive wordlists.

10.2.4. Additional BERT-QC Model Details

Hyperparameters: For each layer of the class label hierarchy, we tune the hyperparameters based on the development set. We use the pretrained BERT-Base (uncased) checkpoint. We use the following hyperparameters: maximum sequence length = 256, batch size = 16, learning rates: 2e-5 (L1), 5e-5 (L2-L6), epochs: 5 (L1), 25 (L2-L6).

Statistics: We use non-parametric bootstrap resampling to compare the baseline (Li and Roth (2002) model) to all experimental models to determine significance, using 10,000 bootstrap resamples.

10.3. Question Answering with QC Labels

Hyperparameters: Pilot experiments on both pre-trained BERT-Base and BERT-Large checkpoints showed similar performance benefits at the finest levels of question classification granularity (L6), but the BERT-Large model demonstrated higher overall baseline performance, and larger incremental benefits at lower levels of QC granularity, so we evaluated using that model. We lightly tuned hyperparameters on the development set surrounding those reported by Devlin et al. (2018), and ultimately settled on parameters similar to their original work, tempered by technical limitations in running the BERT-Large model on available hardware: maximum sequence length = 128, batch size = 16, learning rate: 1e-5. We report performance as the average of 10 runs for each datapoint. The number of epochs were tuned on each run on the development set (to a maximum of 8 epochs), where most models converged to maximum performance within 5 epochs.

Preference for uncorrelated errors in multiple choice question classification: We primarily report QA performance using BERT-QC trained using text from only the multiple choice questions and not their answer candidates. While this model achieved lower overall QC performance compared to the model trained with both question and multiple choice answer candidate text, it achieved slightly higher performance in the QA+QC setting. Our error analysis in Section 4.3. shows that though models trained on both question and answer text can achieve higher QC performance, when they make QC errors, the errors tend to be highly correlated with an incorrect answer candidate, which can substantially reduce QA performance. This is an important result for question classification in the context of multiple choice exams.In the context of multiple choice exams, correlated noise can substantially reduce QA performance, meaning the kinds of errors that a model makes are important, and evaluating QC performance in context with QA models that make use of those QC systems is critical.

Related to this result, we provide an analysis of the noise sensitivity of the QA+QC model for different levels of question classification prediction accuracy. Here, we perturb gold question labels by randomly selecting a proportion of questions (between 5% and 40%) and randomly assigning that question a different label. Figure 3 shows that this uncorrelated noise provides roughly linear decreases in performance, and still shows moderate gains at 60% accuracy (40% noise) with uncorrelated noise. This suggests that when making errors, making random errors (that are not correlated to incorrect multiple choice answers) is preferential.

Training with predicted labels: We observed small gains when training the BERT-QA model with predicted QC labels. We generate predicted labels for the training set using 5-fold crossvalidation over only the training questions.

Statistics: We use non-parametric bootstrap resampling to compare baseline (no label) and experimental (QC labeled) runs of the QA+QC experiment. Because the BERT-QA model produces different performance values across successive runs, we perform 10 runs of each condition. We then compute pairwise p-values for each of the 10 no label and QC labeled runs (generating 100 comparisons), then use Fisher's method to combine these into a final statistic.

10.3.1. Interpretation of non-linear question answering gains between levels

Question classification paired with question answering shows statistically significant gains of +1.7% P@1 at L6 using predicted labels, and a ceiling gain of up to +10% P@1 using gold labels. The QA performance graph in Figure 2 contains two deviations from the expectation of linear gains with increasing specificity, at L1 and L3. *Region at* $L2 \rightarrow L3$: On gold labels, L3 provides small gains over L2, where as L4 provides large gains over L3. We hypothesize that this is because approximately 57% of question labels belong to the *Earth Science* or *Life Science* categories which have much more depth than breadth in the standardized science curriculum, and as such these categories are primarily differentiated from broad topics into detailed problem types at levels L4 through L6. Most other curriculum categories have more breadth than depth, and show strong (but not necessarily full) differentiation at L2. *Region at* $L1$: Predicted performance at L1 is higher than gold performance at L1. We hypothesize this is because we train using predicted rather than gold labels, which provides a boost in performance. Training on gold labels and testing on predicted labels substantially reduces the difference between gold and predicted performance.

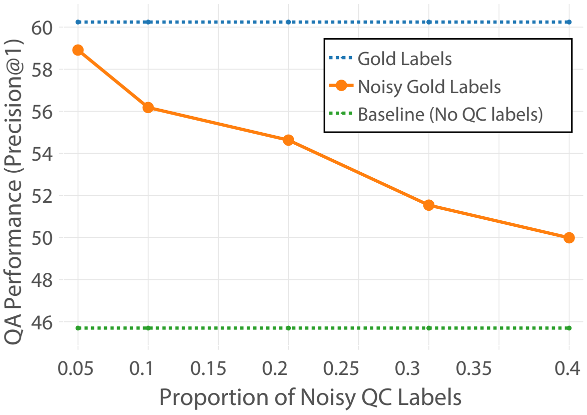


Figure 3: Analysis of noisy question classification labels on overall QA performance. Here, the X axis represents the proportion of gold QA labels that have been randomly switched to another of the 406 possible labels at the finest level of granularity in the classification taxonomy (L6). QA performance decreases approximately linearly as the proportion of noisy QC labels increases. Each point represents the average of 20 experimental runs, with different questions and random labels for each run. QA performance reported is on the development set. Note that due to the runtime associated with this analysis, the results reported are using the BERT-Base model.

10.3.2. Overall annotation accuracy

Though initial raw interannotator agreement was measured at $kappa = 0.58$, to maximize the quality of the annotation the annotators performed a second pass where all disagreements were manually resolved. Table 11 shows question classification performance of the BERT-QC model at 57.8% P@1, meaning 42.2% of the predicted labels were different than the gold labels. The question classification error analysis in Table 7 found that of these 42.2% of errorful predictions, 10% of errors (4.2% of total labels) were caused by the gold labels being incorrect. This allows us to estimate that the overall quality of the annotation (the proportion of questions that have a correct human authored label) is approximately 96%.