Annotating Educational Questions for Student Response Analysis

Andreea Godea and Rodney Nielsen
University of North Texas
1155 Union Circle, Denton, TX, USA
AndreeaGodea@my.unt.edu
Rodney.Nielsen@unt.edu

Abstract

Questions play an important role in the educational domain, representing the main form of interaction between instructors and students. In this paper, we introduce the first taxonomy and annotated educational corpus of questions that aims to help with the analysis of student responses. The dataset can be employed in approaches that classify questions based on the expected answer types. This can be an important component in applications that require prior knowledge about the desired answer to a given question, such as educational and question answering systems. To demonstrate the applicability and the effectiveness of the data within approaches to classify questions based on expected answer types, we performed extensive experiments on our dataset using a neural network with word embeddings as features. The approach achieved a weighted F_1-score of 0.511, overcoming the baseline by 12%. This demonstrates that our corpus can be effectively integrated in simple approaches that classify questions based on the response type.

Keywords: question dataset, taxonomy, expected answer types, educational environment

1. Introduction

Questions represent natural language sentences that express the information need of the inquirer. The analysis of questions is an important part in educational systems, since questioning is the main form of interaction between instructors and students. In this domain, the automatic classification of questions has multiple potential applications. For instance, it can help in the assessment process, in developing effective teaching strategies or in the analysis of questions asked by the students. Question datasets and taxonomies play a very important part in any automatic approach, being used as a base to identify patterns in annotated data that will be applied further to unseen examples. In the educational field, the existence of approaches using such data will provide important information to instructors regarding their interaction with students and will allow them to adapt their teaching strategies to the classroom's needs.

Various question taxonomies and classification approaches for educational systems have been proposed to help in the analysis of data from this domain. However, the majority of question datasets and taxonomies that consider the expected answer type were designed for question answering (QA) and less for educational systems. This is because question classification is a very important component in QA systems, since the main goal of such systems is to identify the best possible answer among a collection of candidate answers, given a question asked by the user. On the other hand, in the educational domain, the questions have multiple potential applications and can be analyzed from different perspectives. Based on the objective being sought, researchers focused their attention on classifying questions based on their subject (Conner, 1927), the educational objective (Bloom, 1956), the difficulty level (İnce, 2008) or the question goal (Lehnert, 1977). However, we found that this domain lacks datasets and taxonomies that aim to analyze questions with respect to *expected answer types*.

In this paper, we introduce the first taxonomy and annotated corpus that aims to analyze educational questions based on expected answer types. More specific, the taxonomy characterizes the questions based on the type of information that is expected to appear in correct answers. Our analysis of questions provides valuable information regarding the expected answer, which will help in identifying if *the correct answer is expected to provide the solution to a given problem, an equation or a drawing, or if it is expected to be a short or a constructed response*, among others.

The dataset can be integrated as an important component in various systems. For example, it can be employed within question answering systems to automatically identify the types of answers elicited by users' questions. This can help in narrowing the space of candidate answers and ensure more accurate recommendations to users. In addition to this, the corpus can be leveraged in educational systems that aim to analyze questions based on the types of the expected answers. The task has multiple potential applications, such as facilitating the assessment process by comparing student responses with the expected answers or identifying which concepts were understood, misunderstood or omitted by the student. This information can help teachers to draw important conclusions regarding the students' conceptual understanding, and allow them to develop teaching strategies based on the students' needs.

Finally, if incorporated in an educational system, the dataset we introduce can also be used to automatically generate questions depending on the type of answers the instructor wants to elicit. For example, the teacher can choose to ask short/direct questions or elicit constructed responses, drawings, equations or it can ask the student to provide the solution to a given problem.

The main contributions of this paper are: (1) We present the first question taxonomy based on expected answer types for educational applications, comprising 16 categories, (2) We collect a dataset of questions from real middle school science classrooms and construct thorough anno-

tation guidelines based on the analysis of questions, which will both be released for research purposes and (3) We provide evidence that our dataset can be effectively integrated in supervised approaches within educational systems.

In the remainder of this paper, we present an overview of related work, describe our dataset and present additional evidence that our data can be effectively utilized by supervised approaches to determine what to expect in student responses.

2. Related Work

Over time, researchers have proposed various datasets and taxonomies for question classification based on their interests. One widely used way to classify questions is based on the expected answer types. Although multiple applications can benefit from analyzing questions based on this criterion, the majority of datasets and taxonomies were designed for question answering systems. However, this type of classification can have multiple potential applications in educational systems as well, from facilitating student assessment to identifying the students' knowledge gaps in order to initiate classrooms discussions. Since educational systems currently lack such data, we present the first question dataset and taxonomy based on expected answer types which can help with the analysis of student responses.

Question classification is an important component in QA systems, which use question datasets and taxonomies to learn patterns in question-answer pairs. In the QA area, the most well known datasets were developed within the Text Retrieval Conference (TREC) QA tracks, which published large amounts of data each year to support competitive research toward developing systems capable of answering open-domain, closed-class questions. Starting with TREC-8 (Voorhees and Tice, 1999), new subsets of questions were included each year, extracted from different sources (e.g., Encarta, Excite, MSNSearch, AskJeeves logs).

Subsets of the TREC datasets have been used by various researchers in their approaches. For example, Hovy et al. (2001) proposed the USC dataset, containing question-answer pairs from TREC-8, TREC-9 and $answers.com$. They created a question taxonomy that reflects the user's intention, such as *veracity (yes-no, true-false), entity (e.g., agent, quantity, location)* and *narative (e.g., history, evaluation, cause-effect)*. Later, Li and Roth (2002) introduced the UIUC dataset based on the USC dataset and TREC-10. They proposed a two layer taxonomy containing 6 coarse classes (abbreviation, entity, description, human, location and numeric value) and 50 fine classes. This hierarchical taxonomy allows the classification of questions at various degrees of granularity and allows more flexibility than the flat one proposed by Hovy et al. (2001).

More recently, Yang et al. (2015) proposed a dataset for open-domain question answering, named WIKIQA. The dataset contains questions collected from Bing query logs and each question is associated with a Wikipedia page assumed to be the topic of the question. The candidate answers for a given question are considered to be all the sentences in the summary paragraph of the Wikipedia page. In contrast with the TREC datasets, WIKIQA is more challenging because it includes questions with no correct answers. The questions were labeled in a similar manner with previous works, based on the following answer types: *location, human, numeric, abbreviation, entity, description.*

In this paper, we introduce the first dataset with questions from the educational domain, annotated with expected answer types. In contrast with previous datasets with questions primarily used for question answering systems, we propose the first dataset that can facilitate the analysis of student responses in the educational environment.

3. The Corpus

The corpus presented in this paper contains questions from the educational environment. Specifically, we collected science questions asked by teachers in real middle school classrooms. The teachers entered their questions on a web-enabled device and presented them to the students in order to initiate discussions and identify potential gaps in the conceptual understanding. After a thorough analysis of the questions in our data, we propose a novel taxonomy containing 16 categories of questions based on expected answer types. Specifically, the taxonomy was created considering previously proposed schemes and the particularities of questions in our data, with the focus on the features that would facilitate the identification of what is expected from a correct student response. Since our questions can contain one or more sentences and each sentence can elicit different information, we created our taxonomy based on each unique question sentence in our data.

In this section, we present our taxonomy, the inter-class correlations, the annotation process and data distribution.

3.1. Question Taxonomy

We propose a taxonomy with 16 question categories based on expected answer types, as follows:

Clarification – the question elicits a response, but only clarifies, details or paraphrases information already requested in preceding sentences. Example: "Describe mammals. *Discuss the brain, dermis, and child rearing.*"

SubjectiveConcept – the question asks for feelings or opinions rather than facts. There is no wrong answer assuming the response is honest and on task and that any background information or supporting claims are accurate. Example: "*Describe your favorite activity or lab.*"

Select1 – the question provides two or more possible answers and only one of them is correct. The option labels must be included in the text of the question. Example: "*Which would you expect to have a higher density, hot water or cold water?*"

SelectN – the question provides a list of possible answers and elicits the selection of all responses that apply. The option labels must be included in the text of the question. Example: "*Which of the following sources of energy are considered clean?* Coal, Solar, Wind, Oil, Gas, Nuclear."

TrueFalse – the question requires a positive (e.g., true, yes) or negative response (e.g., false, no). Example: "*Does energy have anything to do with physical movement?*"

List – the question elicits a list of items. Example: "*Explain at least two differences between longitudinal and transverse waves.*"

Data	Sents	Clarif	SubjC	Sel1	SelN	T/F	List	MultiP	ShrtAns	OthCR	Proc	Eq	Soln	Draw	CntxtS	AnsVry	Order
Train	1085	25	21	24	0	29	145	102	272	633	62	14	21	6	252	207	112
Test	569	17	13	20	1	15	75	52	151	313	38	8	16	4	125	99	58
Total	1654	42	34	44	1	44	220	154	423	946	100	22	37	10	377	306	170

Table 1: Number of Instances per Class.

Multi-Part – the question asks for a number of various items, differentiated in text. Example: "*What is mass and how do we measure it?*"
VeryShortAnswer – the question elicits an extremely short phrase or a single word. Example: "*Where was Helium discovered?*"
OtherConstructedResponse – the question seeks a constructed response that can have up to several sentences. The answer should contain at least a verb phrase. Example: "*Compare and contrast microwaves with gamma waves.*"
ProcessProcedure – the question requests the process by which something happens (e.g., a natural/involuntary process of change) or the procedure for accomplishing a task. Example: "*How did Marie Currie discover radioactivity?*"
Equation – the question elicits an equation/formula. Example: "*Write the equation for calculating an object's speed.*"
Solution – the question asks the student to solve a computational or mathematical problem. Example: "*What is the volume of 103 g of water?*"
Drawing – the question asks for a drawing. Example: "*Sketch the atomic structure for nitrogen and boron using the Bohr Model.*"
ContextSensitive – the question is referring to a picture, video, image, or previously-conducted lab, etc. The answer should refer to material that is not explicitly included in the text of the question and that is not based on the general subject matter the course is teaching. Example: "*Summarize what you learned yesterday by using the simulation.*"
AnswersWillVary – the question has more than one correct answer. Hence, the student responses can vary. Example: "*Think of a chemical reaction you are familiar with and list the reactants and products.*"
Ordered – the question expects a specific sequence within the response. Example: "*Name the planets in order from the closest to the sun to the furthest from the sun.*"

3.2. Inter-class Correlations

From the description of our taxonomy, it can be noticed that only a subset of types are in general mutually exclusive – a sentence cannot be considered both a *VeryShortAnswer* and an *OtherConstructedResponse* unless we deal with a *Multi-Part* question. It can be observed that several question types are ***not*** mutually exclusive. For example, the question "*Name one important lab safety procedure and explain why it is important.*" will be considered *Multi-Part, VeryShortAnswer and OtherConstructedResponse*, since it elicits two different parts – a name (a short answer) and an explanation (a constructed response). In fact, we identified several class pairs that are highly correlated in our data.

- A question annotated as a *SubjectiveConcept* is always also considered an *AnswersWillVary* ("*What interested you most about elements in periodic table?*"), but the reverse does not always apply ("*What scientific practice could you reflect upon for the Atoms?*"). Based on the context, *SubjectiveConcept* questions are often considered *OtherConstructedResponse*, but could be annotated as *VeryShortAnswer*, among others.
- A *List* is very frequently also annotated as *AnswersWillVary* ("*Describe some of the unique qualities of the water molecule.*"), but there are also exceptions ("*Name **the three** parts of an atom.*").
- A *ProcessProcedure* question is very often considered an *OtherConstructedResponse* and *Ordered*, since it implies the description of a sequence of steps.
- A *Solution* is in general assumed to also be a *VeryShortAnswer*, unless it asks the students to show their work (where it is *OtherConstructedResponse*).
- A *Multi-Part* question is expected to also be *Ordered*.

3.3. Data Annotation

The dataset presented in this paper comprises 1155 questions, with 1654 sentences in total. Each question is composed of one or more sentences, with the largest question having 18 sentences (only six of which elicited a response). Since our questions can contain multiple sentences and each sentence can elicit different information, we created the taxonomy based on the question sentences in our data. Hence, in the annotation process, each question was first split into sentences. Then, each sentence was independently annotated with class types by two graduate students and adjudicated by a third. Sentences were tagged with one or more labels, since the categories are not mutually exclusive and each sentence can elicit multiple response types.
An analysis of our data revealed that the maximum number of labels attached to a sentence is 6. This applies to the italicized sentence from the following question: "What was his apparatus of choice? *Draw and label its components.*", which is annotated as: (1) *Multi-Part* (it elicits two different things – a drawing and a list of components), (2) *List* (the answer should contain a list), (3) *VeryShortAnswer* (the list should contain short answers), (4) *Drawing* (the sentence elicits a drawing), (5) *ContextSensitive* (it refers to context sensitive material – "his apparatus of choice") and (6) *AnswersWillVary* (the list length was not specified).
The inter-annotator agreement over each class is Kappa = 0.75, which is a substantial agreement according to Landis and Koch (1977). Further, we analyzed the independent labels of the first two annotators and we observed that most disagreements were in labeling the *ContextSensitive* class. More specifically, 24% of disagreements were for the *ContextSensitive* class, followed by *VeryShortAnswer* and *OtherConstructedResponse* with 18% each. On the other hand, the annotators agreed in all cases when labeling *Drawing*

Clarification	SubjCon	Select1	T/F	List	MultiP	ShrtAns	OthConRrsp	Process	Equation	Solution	CntxtSens	AnsWillVry	Order
0.000	0.273	0.125	0.100	0.483	0.364	0.51	0.75	0.154	0.571	0.522	0.50	0.472	0.174

Table 2: Test Set F_1-score per Class.

and *Ordered* questions. This can be explained by the fact that these classes have clearer patterns in data and can be easily separated from the other class types. However, identifying if a question elicits context sensitive information or a short versus a longer response appears to be more subjective, based on each annotator's interpretation.

3.4. Data Distribution

To assess the applicability of our dataset, we also tested it within a supervised approach. For this purpose, we split the *questions* in our dataset into two separate sets – 66% for *train* and 34% for *test*. We performed the split at the question level instead of sentence level, to ensure that all the information comprised in a sentence is located in a single subset of data - either train or test. This is an important aspect in the classification approach, since sometimes the labels are preserved between sentences within the same question. For instance, if a sentence comprises *ContextSensitive* information, the next sentences within the same question will also have this label if they refer to the same concepts. Similarly, a *Clarification* sentence always refers to previous sentences within the question, since its goal is to detail or paraphrase what was previously elicited. We provide the data distribution based on adjudicated labels in Table 1.

The distribution of question types in the data reveals that the dominant class is *OtherConstructedResponse* with 946 instances, followed by *VeryShortAnswer* with 423 instances. This implies that these instructors generally focused on asking deeper questions in this dataset. This is an important observation regarding our data, since it was shown that involving deep questioning during tutoring can improve knowledge learning (Chi et al., 1994). On the other hand, the least frequent question types in the data collected during tutoring are *SelectN, Drawing* and *Equation*. These question types require direct answers – the selection of all options that apply from the question text, a drawing or an equation.

4. Methodology for Question Classification

We propose a supervised approach in order to validate our data and demonstrate that it can be effectively learned. More specifically, we trained separate artificial neural networks using the one-vs.-all strategy for each class and used pre-trained word embeddings (Pennington et al., 2014) to classify questions based on their expected answer types. To this end, we split the training data into two subsets – 66% for training and 34% for validation, and used the *validation* set to run experiments in order to tune the parameters for each class.

We experimented with various word embeddings dimensions (50, 100 or 200) for each class type and finally set the dimension to 100 for the *ContextSensitive, Equation, List, Multi-Part, ProcessProcedure, Select1* and *SubjectiveConcept* classes and 200 for the remaining classes. The word embeddings corresponding to a question were combined into a single feature vector by computing their normalized sum, as in the following equation:

$$v(Q, C) = || \sum_{w \in Q} glove(w_C)||$$

where $v(Q, C)$ represents the features for a given question-class pairing, w iterates over all the words in the question, and the function $glove(w_C)$ retrieves the Glove word embedding for w with the dimension specific to class C.

As a result of tuning the network's parameters on the validation data, the number of iterations was set to 3000, the learning rate had values between 0.01 and 0.1, and the number of hidden layers was set to 2, with the number of nodes ranging from 3 to 10 in a layer. The results obtained for our classes using these parameters are reported in Table 2. As it can be observed, we did not include the *Drawing* and *SelectN* classes, because they have fewer than 10 examples in the entire dataset (see Table 1 for the distribution).

Our results show that the best performing class is *OtherConstructedResponse* with F_1-score = 0.75, followed by *Equation, Solution, VeryShortAnswer* and *ContextSensitive*, each achieving an F_1-score higher than 0.5. In case of *OtherConstructedResponse* and *VeryShortAnswer* classes, the results can be explained by the large number of examples in the training set, which helped in identifying patterns for these question types. Although *Equation* and *Solution* have less training examples, these classes possess clearer patterns in the data. On the other hand, the worst performing class is *Clarification*, for which the simple classifier (strawman) was not able to capture patterns. This is because the classifier employs only the sentence's context and does not take into account what was requested in previous sentences. A *Clarification* can be easier identified if the information from previous sentences is taken into account, since the goal of this class type is to clarify, detail or paraphrase other sentence.

It can be seen that *TrueFalse* and *Select1* are among the worst performing classes, although they intuitively follow specific patterns. We analyzed our data and found that there are no specific keywords in the question sentences associated with these types of categories. For *TrueFalse*, our questions do not explicitly contain terms such as *true* or *false*, but generally start with auxiliary verbs and require yes/no responses – *"Does energy have anything to do with physical movement?"*. We experimented with adding a new binary feature to the word embeddings' feature vector, which checks if the question starts with an auxiliary verb. The performance reached an F_1-score = 0.65, with an increase of 55% compared with using only word embeddings.

With respect to Select1, we found that the corresponding question sentences do not contain keywords such as *"select"* and there are cases when the options given are part of other question sentences. However, we found that frequent words associated with *Select1* within a sentence are the wh-word *"which"* and the conjunction *"or"* – *"Which would you expect to have a higher density, hot water or cold wa-*

ter?". We experimented with adding two binary features to the word embeddings to check if the sentence contains any of these words and the performance reached an F_1-score = 0.26, doubling the performance obtained by using only word embeddings. These experiments show that our data can be learned by simple approaches and the performance can be improved by leveraging more of the question types' patterns.

The overall performance using word embeddings was computed in terms of weighted F_1-score over all classes at 0.511. As a baseline, we considered the *Majority Class* approach (all instances were labeled *OtherConstructedResponse*), which achieved a weighted F_1-score = 0.39. As can be observed, our approach surpassed the baseline by 12%. This suggests that our dataset can be effectively learned by a more complex approach by leveraging features specific to each question type, in addition to word embeddings.

5. Conclusion

In this paper, we presented an educational corpus collected from middle school science classrooms. The data contains questions asked by instructors during tutoring, which are annotated based on expected answer types. The goal of the proposed corpus is to enable the classification of instructors' questions in order to help with the analysis of student responses. This has the potential to improve the assessment process by facilitating the interpretation, comparison and contrasting of student responses. This will, in turn, provide the instructor with better formative assessment regarding the concepts understood, misunderstood or omitted by students in their answers. The corpus can also be leveraged in the question answering domain, where systems need information about the expected answer to a user's question in order to make accurate recommendations of answers.

The quality of the annotation is attested by high inter-rater reliability, K=0.75 (substantial agreement). We also tested our data on a supervised approach employing artificial neural networks and word embeddings, achieving a weighted F_1-score of 0.51, outperforming the baseline by 12%. This demonstrates that our annotations of a question's expected answer type are learnable, even by a relatively simple approach. The dataset will be released to the research community to improve and extend these findings.

6. Acknowledgements

This research was supported by the Institute of Education Sciences, U.S. Department of Education, Grant R305A120808 to University of North Texas. Opinions expressed are those of the authors.

7. Language Resource References

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals: Cognitive Domain*. Longman.

Chi, M. T., Leeuw, N., Chiu, M.-H., and LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive science*, 18(3):439–477.

Conner, M. (1927). What a reference librarian should know. *Library Journal*, 52(8):415–418.

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.

İnce, İ. F. (2008). *Intellegent question classification for e-learning environments by data mining techniques*. Ph.D. thesis, Institute of Science.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Lehnert, W. G. (1977). The process of question answering. Technical report, DTIC Document.

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.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Voorhees, E. M. and Tice, D. M. (1999). The trec-8 question answering track evaluation. In *TREC*, volume 1999, page 82.

Yang, Y., Yih, W.-t., and Meek, C. (2015). Wikiqa: A challenge dataset for open-domain question answering. In *EMNLP*, pages 2013–2018.