Evaluation Dataset (DT-Grade) and Word Weighting Approach towards Constructed Short Answers Assessment in Tutorial Dialogue Context

Rajendra Banjade, Nabin Maharjan, Nobal B. Niraula, Dipesh Gautam, Borhan Samei, Vasile Rus
Department of Computer Science / Institute for Intelligent Systems
The University of Memphis
Memphis, TN, USA
{rbanjade, nmharjan, nbnraula, dgautam, bsamei, vrus}@memphis.edu

Abstract

Evaluating student answers often requires contextual information, such as previous utterances in conversational tutoring systems. For example, students use coreferences and write elliptical responses, i.e. incomplete but can be interpreted in context. The DT-Grade corpus which we present in this paper consists of short constructed answers extracted from tutorial dialogues between students and an Intelligent Tutoring System and annotated for their correctness in the given context and whether the contextual information was useful. The dataset contains 900 answers (of which about 25% required contextual information to properly interpret them). We also present a baseline system developed to predict the correctness label (such as correct, correct but incomplete) in which weights for the words are assigned based on context.

1 Introduction

Constructed short answers are responses produced by students to questions, e.g. in a test or in the middle of a tutorial dialogue. Such constructed answers are very different form answers to multiple choice questions where students just choose an option from the given list of choices. In this paper, we present a corpus called DT-Grade[1] which contains constructed short answers generated during interaction with a state-of-the-art conversational Intelligent Tutoring System (ITS) called DeepTutor (Rus et al., 2013; Rus et al., 2015). The main instructional task during tutoring was conceptual problem solving in the area of Newtonian physics. The answers in our data set are shorter than 100 words. We annotated the instances, i.e. the student generated responses, for correctness using one of the following labels: correct, correct-but-incomplete, contradictory, or incorrect. The student answers were evaluated with respect to target/ideal answers provided by Physics experts while also considering the context of the student-tutor interaction which consists of the Physics problem description and the dialogue history related to that problem. In fact, during annotation we only limited our context to the immediately preceding tutor question and problem description. This decision was based on previous work by Niraula and colleagues (2014) that showed that most of the referring expressions can be resolved by looking at the past utterance; that is, looking at just the previous utterance could be sufficient for our task as considering the full dialogue context would be computationally very expensive.

Automatic answer assessment systems typically assess student responses by measuring how much of the targeted concept is present in the student answer. To this end, subject matter experts create target (or reference) answers to questions that students will be prompted to answer. Almost always, the student responses depend on the context (at least broadly on the context of a particular domain) but it is more prominent in some situations. Particularly in conversational tutoring systems, the meanings of students' responses often depend on the dialogue context and problem/task description. For example, students frequently use pronouns, such as *they, he, she*, and *it*, in their response to tutors' questions or other prompts.

[1] Available at http://language.memphis.edu/dt-grade

Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 182–187,
San Diego, California, June 16, 2016.

In an analysis of tutorial conversation logs, Niraula et al. (2014) found that 68% of the pronouns used by students were referring to entities in the previous utterances or in the problem description. In addition to anaphora, complex coreferences are also employed by students.

Also, in tutorial dialogues students react often with very short answers which are easily interpreted by human tutors as the dialogue context offers support to fill-in the blanks or untold parts. Such elliptical utterances are common in conversations even when the speakers are instructed to produce more syntactically and semantically complete utterances (Carbonell, 1983). By analyzing 900 student responses given to DeepTutor tutoring systems, we have found that about 25% of the answers require some contextual information to properly interpret them.

Problem description: A car windshield collides with a mosquito, squashing it.
Tutor question: How do the amounts of the force exerted on the windshield by the mosquito and the force exerted on the mosquito by the windshield compare?
Reference answer:
The force exerted by the windshield on the mosquito and the force exerted by the mosquito on the windshield are an action-reaction pair.
Student answers:
A1. *Equal*
A2. *The force of the bug hitting the window is much less than the force that the window exerts on the bug*
A3. *they are equal and opposite in direction*
A4. *equal and opposite*

Table 1: A problem and student answers to the given question.

As illustrated in the Table 1, the student answers may vary greatly. For instance, answer A1 is elliptical. The *"bug"* in A2 is referring to the mosquito and *"they"* in A3 is referring to the amount of forces exerted to each other.

In order to foster research in automatic answer assessment in context (also in general), we have annotated 900 student responses gathered from an experiment with the DeepTutor intelligent tutoring system (Rus et al., 2013). Each response was annotated for: (a) their correctness, (b) whether the contextual information was helpful in understanding the student answer, and (c) whether the student answer contains important extra information. The annotation labels, which are similar to the ones proposed by Dzikovska et al. (2013), were chosen such that there is a balance between the level of specificity and the amount of effort required for the annotation.

We also developed a baseline system using semantic similarity approach with word weighting scheme utilizing contextual information.

2 Related Work

Nielsen et al. (2008) described a representation for reference answers, breaking them into detailed facets and annotating their relationships to the learners answer at finer level. They annotated a corpus (called SCIENTSBANK corpus) containing student answers to assessment questions in 15 different science domains. Sukkarieh and Bolge (2010) introduced an ETS-built test suite towards establishing a benchmark. In the dataset, each target answer is divided into a set of main points (called content) and recommended rubric for assigning score points.

Mohler and Mihalcea (2009) published a collection of short student answers and grades for a course in Computer Science. Most recently, a Semantic Evaluation (SemEval) shared task called Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge was organized (Dzikovska et al., 2013) to promote and streamline research in this area. The corpus used in the shared task consists of two distinct subsets: BEETLE data, based on transcripts of students interacting with BEETLE II tutorial dialogue system (Dzikovska et al., 2010), and SCIENTSBANK data. Student answers, accompanied with their corresponding questions and reference answers are labeled using five different categories. Basu et al. (2013) created a dataset called Powergrading-1.0 which contains responses from hundreds of Mechanical Turk workers to each of 20 questions from the 100 questions published by the USCIS as preparation for the citizenship test.

Our work differs in several important ways from previous work. Our dataset is annotated paying special attention to context. In addition to the tutor question, we have provided the problem description

as well which provides a greater amount of contextual information and we have explicitly marked whether the contextual information was important to properly interpret/annotate the answer. Furthermore, we have annotated whether the student answer contains important extra information. This information is also very useful in building and evaluating natural language tools for automatic answer assessment.

3 Data Collection and Annotation

Data Collection: We created the DT-Grade dataset by extracting student answers from logged tutorial interactions between 40 junior level college students and the DeepTutor system (Rus et al., 2013). During the interactions, each student solved 9 conceptual physics problems and the interactions were in the form of purely natural language dialogues, i.e., with no mathematical expressions and special symbols. Each problem contained multiple questions including gap-fill questions and short constructed answer questions. As we focused on creating constructed answer assessment dataset with sentential input, we filtered out other types of questions and corresponding student answers. We randomly picked 900 answers for the annotation.

Annotation: The annotation was conducted by a group of graduate students and researchers who were first trained before being asked to annotate the data. The annotators had access to an annotation manual for their reference. Each annotation example (see Figure 1) contained the following information: (a) problem description (describes the scenario or context), (b) tutor question, (c) student answer in its natural form (i.e., without correcting spelling errors and grammatical errors), (d) list of reference answers for the question. The annotators were asked to read the problem and question to understand the context and to assess the correctness of the student answer with respect to reference answers. Each of the answers has been assigned one of the following labels.

Correct: Answer is fully correct in the context. Extra information, if any, in the answer is not contradicting with the answer.

Correct-but-incomplete: Whatever the student provided is correct but something is missing, i.e. it is not complete. If the answer contains some incorrect part also, the answer is treated as incorrect.

Contradictory: Answer is opposite or is very contrasting to the reference answer. For example, "equal", "less", and "greater" are contradictory to each other. However, Newton's first law and Newton's second law are not treated as contradictory since there are many commonalities between these two laws despite their names.

Incorrect: Incorrect in general, i.e. none of the above three judgments is applicable. Contradictory answers can be included in the incorrect set if we want to find all kinds of incorrect answers.

```
<Instance ID="386">
<MetaInfo StudentID ="DTSU017"
TaskID="LP03_PR09.bLK.sh"
DataSource="DeepTutorSummer2014"/>
<ProblemDescription>A car windshield collides with a
mosquito, squashing it.</ProblemDescription>
<Question>How does Newton's third law apply to this
situation?</Question>
<Answer>both objects exert the same amount of force on each
other</Answer>
<Annotation
Label="correct(0)|correct_but_incomplete(0)|contradictory(0)|i
ncorrect(0)">
<AdditionalAnnotation ContextRequired="0|1"
ExtraInfoInAnswer="0|1"/>
<Comments Watch="0|1"> </Comments>
</Annotation>
<ReferenceAnswers>
1: The action is the windshield squashing the mosquito, and
the equal and opposite reaction is the mosquito hitting the
windshield.
</ReferenceAnswers>
</Instance>
```

Figure 1: An annotation example.

As shown in Figure 1, annotators were asked to assign one of the mutually exclusive labels - correct, correct-but-incomplete, contradictory, or incorrect. Also, annotators were told to mark whether contextual information was really important to fully understand a student answer. For instance, the student answer in the Figure 1 contains the phrase *"both forces"* which is referring to the force of windshield and the force of mosquito in problem description. Therefore, contextual information is useful to fully understand what both forces the student is referring to. As shown in Table 1 (in Section 1), a student answer could be an elliptical sentence (i.e., does not contain complete information on its own). In such

Parameter	Value
All	900
Correct	365 (40.55%)
Correct but incomplete	209 (23.22%)
Contradictory	84 (9.33%)
Incorrect	242 (26.88%)
Requiring context	223 (24.77%)
Containing extra info	102 (11.33%)

Table 2: Summary of DT-Grade dataset.

cases, annotators were asked to judge the student response based on the available contextual information and reference answers and nothing more; that is, they were explicitly told not to use their own science knowledge to fill-in the missing parts.

If a student response contained extra information (i.e., more information than in the reference/ideal answer provided by experts), we asked annotators to ignore the extra parts unless it expressed a misconception. However, we told annotator to indicate whether the student answer contains some additional important information such as a detailed explanation of their answer. The annotators were encouraged to write comments and asked to set the 'watch' flag whenever they felt a particular student response was special/different. Such 'to watch' instances were considered for further discussions with the entire team to either improve the annotation guidelines or to gain more insights regarding the student assessment task.

The dataset was divided equally among 6 annotators who then annotated independently. In order to reach a good level of inter-annotator agreement in annotation, 30 examples were randomly picked from each annotation subset and reviewed by a supervisor, i.e. one of the creators of the annotation guidelines. The agreements (in terms of Cohen's kappa) in assigning correctness label, identifying whether the context was useful, and identifying whether the student answer contained extra information were 0.891, 0.78, and 0.82 respectively. In another words, there were significant agreements in all components of the annotation. The main disagreement was on how to use the contextual information. The disagreements were discussed among the annotators team and the annotations were revised in few cases.

The Dataset: We have annotated 900 answers. Table 2 offers summary statistics about the dataset. The 40.55% of total answers are correct whereas 59.45% are less than perfect. We can see that approximately 1 in every 4 answers required contextual information to properly evaluate them.

4 Alignment Based Similarity and Word Weighting Approach

Approach: Once the dataset was finalized we wanted to get a sense of its difficulty level. We developed a semantic similarity approach in order to assess the correctness of student answers. Specifically, we applied optimal word alignment based method (Banjade et al., 2015; Rus and Lintean, 2012) to calculate the similarity between student answer and the reference answer and then used that score to predict the correctness label using a classifier. In fact, the alignment based systems have been the top performing systems in semantic evaluation challenges on semantic textual similarity (Han et al., 2013; Agirre et al., 2014; Sultan et al., 2015; Agirre et al., 2015).

The challenge is to address the linguistic phenomena such as ellipsis and coreferences. An approach can be to use off-the-shelf tools, such as coreference resolution tool included in Stanford CoreNLP Toolkit (Manning et al., 2014). However, we believe that such NLP tools that are developed and evaluated in standard dataset potentially introduce errors in the NLP pipeline where the input texts, such as question answering data, are different from literary style or standard written texts.

As an alternative approach, we assigned a weight for each word based on the context: we gave a low weight to words in the student answer that were also found in the previous utterance, e.g. the tutoring systems question, and more weight to new content. This approach gives less weight to answers that simply repeat the content of the tutors question and more weight to the answers that add the new, asked-for information; as a special case, the approach provides more weight to concise answers (see A1 and A2 in Table 1). The same word can have different weight based on the context. Also, it partially addresses the impact of coreferences in answer grading because the same answer with and without coreferences will

be more likely to get comparable scores. The reference answers are usually self contained, i.e. without using coreferring expressions and only those student answers which are also self-contained and similar to reference answer will get higher score. On the other hand, answers using coreferences (such as: they, it) will get lower score unless they are resolved and the student answer becomes similar to reference answer. Giving lower weights to the words, if present in the student answer, for which student could use coreferrences makes these two types of answers somewhat equivalent.

Finally, the similarity score was calculated as:

$$sim(A,R) = 2 * \frac{\sum_{(a,r)\in OA} w_a * w_r * sim(a,r)}{\sum_{a\in A} w_a + \sum_{r\in R} w_r}$$

Where A/R refers to student/reference answer and a/r is a token in it. The $sim(a,r)$ referes to the similarity score between a and r calculated using word2vec model (Mikolov et al., 2013). OA is optimal alignment of words between A and R obtained using Hungarian algorithm as described in Banjade et al. (2015). The $0 \leq w_a \leq 1$ and $0 \leq w_r \leq 1$ refer to weight of the word in A and R respectively.

Experiments and Results: In order to avoid noisy alignments, the word-to-word similarity score below 0.4 was set to 0.0 (empirically set). The $sim(A,R)$ was then used with Multinomial Logistic Regression (in Weka) to predict the correctness label. If there were more than one reference answers, we chose one with the highest similarity score with the student answer. We then set different weights (from 1.0 to 0.0) for the words found in tutor utterance (we considered a word was found in the previous utterance if its base form or the synonym found in WordNet 3.0 (Miller, 1995) matched with any of the words in the previous utterance). We changed the weight in the student answer as well as in the reference answer and the impact of weight change in the classification results were assessed using 10-fold cross validation approach. The changes in classification accuracy with changing weights are presented in Figure 2.

Giving weight of 1.0 to each word is equivalent to aligning words in student answer with the reference answer without looking at the context. But we can see the improvement in classification accuracy after reducing word weights up to 0.4 (accuracy 49.33%; kappa = 0.22) for the words found in the previous utterance and then decreases. It indicates that the words found in previous utterance should get some weight but new words should get more importance. This approach is somewhat intuitive. But deeper semantic understanding is required in order to improve the performance. For instance, sometimes this word weighting approach infers more information and gives higher weight to the incomplete utterance where students true understanding of the context is hard to predict. Furthermore, it is non-trivial to use additional context, such as problem description including assumptions and graphical illustrations.

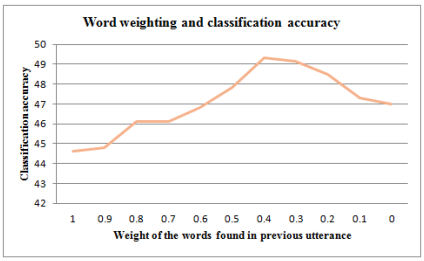

Figure 2: Classification accuracy and weight of the words that are found in the last utterance.

5 Conclusion

We presented a corpus called DT-Grade which contains student answers given to the intelligent tutoring system and annotated for their correctness in context. We explicitly marked whether the contextual information was required to properly understand the student answer. We also annotated whether the answer contains extra information. That additional information can be correct or incorrect as there is no specific reference to compare with but the answer grading systems should be able to handle them.

We also presented a baseline system in which we used semantic similarity generated using optimal alignment with contextual word weighting as feature in the classifier for predicting the correctness label. However, there is enough room for the improvements and using additional features in the classifier or developing a joint inference model such as Markov Logic Network incorporating different linguistic phenomena can be two future directions.

Acknowledgments

This research was supported by the Institute for Education Sciences (IES) under award R305A100875 to Dr. Vasile Rus. All opinions and findings presented here are solely the authors'.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Carmen Baneab, Claire Cardiec, Daniel Cerd, Mona Diabe, Aitor Gonzalez-Agirrea, Weiwei Guof, Inigo Lopez-Gazpioa, Montse Maritxalara, Rada Mihalceab, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Rajendra Banjade, Nobal B Niraula, Nabin Maharjan, Vasile Rus, Dan Stefanescu, Mihai Lintean, and Dipesh Gautam. 2015. Nerosim: A system for measuring and interpreting semantic textual similarity. *SemEval-2015*, page 164.

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, 1:391–402.

Jaime G Carbonell. 1983. Discourse pragmatics and ellipsis resolution in task-oriented natural language interfaces. In *Proceedings of the 21st annual meeting on Association for Computational Linguistics*, pages 164–168. Association for Computational Linguistics.

Myroslava O Dzikovska, Johanna D Moore, Natalie Steinhauser, Gwendolyn Campbell, Elaine Farrow, and Charles B Callaway. 2010. Beetle ii: a system for tutoring and computational linguistics experimentation. In *Proceedings of the ACL 2010 System Demonstrations*, pages 13–18. Association for Computational Linguistics.

Myroslava O Dzikovska, Rodney D Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa T Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. Technical report, DTIC Document.

Lushan Han, Abhay Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. Umbc ebiquity-core: Semantic textual similarity systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, volume 1, pages 44–52.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Michael Mohler and Rada Mihalcea. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 567–575. Association for Computational Linguistics.

Rodney D Nielsen, Wayne Ward, James H Martin, and Martha Palmer. 2008. Annotating students' understanding of science concepts. In *LREC*.

Nobal B Niraula, Vasile Rus, Rajendra Banjade, Dan Stefanescu, William Baggett, and Brent Morgan. 2014. The dare corpus: A resource for anaphora resolution in dialogue based intelligent tutoring systems. In *LREC*, pages 3199–3203.

Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162. Association for Computational Linguistics.

Vasile Rus, Sidney DMello, Xiangen Hu, and Arthur Graesser. 2013. Recent advances in conversational intelligent tutoring systems. *AI magazine*, 34(3):42–54.

Vasile Rus, Nobal Niraula, and Rajendra Banjade. 2015. Deeptutor: An effective, online intelligent tutoring system that promotes deep learning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Jana Z Sukkarieh and Eleanor Bolge. 2010. Building a textual entailment suite for the evaluation of automatic content scoring technologies. In *LREC*. Citeseer.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. Dls@cu: Sentence similarity from word alignment and semantic vector composition. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 148–153.