

Virtual Pre-Service Teacher Assessment and Feedback via Conversational Agents

Debajyoti Datta¹*, Maria Phillips¹*, James P. Bywater², Jennifer Chiu³, Ginger S. Watson³,
Laura Barnes¹ Donald Brown¹

¹School of Engineering and Applied Science, University of Virginia

²College of Education, James Madison University

³School of Education and Human Development, University of Virginia

{dd3ar, mp6kv, jlc4dz, gw2b, lb3dp, deb}@virginia.edu

bywatejx@jmu.edu

Abstract

Conversational agents and assistants have been used for decades to facilitate learning. There are many examples of conversational agents used for educational and training purposes in K-12, higher education, healthcare, the military, and private industry settings. The most common forms of conversational agents in education are teaching agents that directly teach and support learning, peer agents that serve as knowledgeable learning companions to guide learners in the learning process, and teachable agents that function as a novice or less-knowledgeable student trained and taught by a learner who learns by teaching. The Instructional Quality Assessment (IQA) provides a robust framework to evaluate reading comprehension and mathematics instruction. We developed a system for pre-service teachers, individuals in a teacher preparation program, to evaluate teaching instruction quality based on a modified interpretation of IQA metrics. Our demonstration and approach take advantage of recent advances in Natural Language Processing (NLP) and deep learning for each dialogue system component. We built an open-source conversational agent system to engage pre-service teachers in a specific mathematical scenario focused on scale factor with the aim to provide feedback on pre-service teachers' questioning strategies. We believe our system is not only practical for teacher education programs but can also enable other researchers to build new educational scenarios with minimal effort.

1 Introduction

In the era of remote teaching due to governmental regulations and stay-at-home orders for COVID-19, remote teaching methodologies have come to the forefront of education and training. Virtual conversational agents have been used for a variety of

training scenarios to train teachers and nurses in many different contexts (Datta et al., 2016). Virtual conversational agents refers to an online interactive system where a user is able to dialogue and receive responses in turn. all future references to conversational agents will refer to this definition of virtual conversational agents. Conversational agents, often characterized as dialogue systems, are common in customer support representative applications (Yan et al., 2017). In education, they have been used extensively for interpreting the content of the conversation, integrating and assimilating information, and providing feedback (Chhibber and Law, 2019). However, as pointed out by Smutny and Schreiberova (2020) very few conversational agents used in education use recent advances in machine learning and deep learning, instead relying on simple decision trees. This reinforces the need for more research and development on artificial intelligence-based methods to support content-specific conversations. A conversational agent deployed through a web interface, as opposed to software implementations, for teacher learning has multiple benefits. Two of the prominent benefits are that this approach can easily be scaled to reach more in-service and pre-service teachers as well while also providing a cost-effective way to build systems that can be scaled to new teaching scenarios as multiple dialogue system components can be reused. The dialogue manager and semantic sentence-level components can be used for different mathematical scenarios as long as the assessment component remains unchanged. In this case the term assessment is referring to pre-service teacher evaluation as opposed to a students understanding of a mathematical topic. In this work, we combine the advances in NLP and deep learning research with a modified version of the Instructional Quality Assessment (IQA) (Boston, 2012) framework to build a scenario to be used in teacher

* First two authors have contributed equally

education settings. Our goal in this project is to assess and give teachers immediate feedback on the quality of their instructional moves using a specific scenario on a given mathematical topic.

Research demonstrates that the Instructional Quality Assessment provides a robust framework for evaluating teachers' instructional practice in mathematics classrooms. Our demonstration uses a modified version of the IQA that focuses on the strategy of teachers' questions during one-on-one or class discussions through a web-based platform that allows for pre-service teachers to receive real-time feedback on the quality of their questioning. Although our demonstration centers on using an adapted version of IQA as the assessment component, our modular architecture can be used to incorporate alternative evaluation schemes as well.

A critical challenge of developing dialogue systems in a new domain is the requirement of large data sets for the different components of the dialogue system (intent classification, slot filling for dialogue state tracking, and the dialogue policy). In education, where data collection can itself prove a significant challenge, especially given a need for increased domain expert annotations, our system is developed with only a small amount of annotated data: around two thousand sentences that are labeled with one of four adapted IQA classes. Note that this system is not necessarily better than existing approaches that use annotated training data for each component of the dialogue system, rather, this system is a compromise because dialogue systems are challenging to build in low- or no-data amount scenarios. What our system demonstrates is a solution to developing a highly-manipulable scenario given minimal domain-expert annotated data that can be used to support virtual feedback for pre-service teachers.

Our contributions are as follows:

- An open-source web-platform for assessment of the quality of teachers' mathematical questioning
- A process that allows for scenario development with minimal training data
- Direct feedback to pre-service teachers on the quality of mathematical questioning by relying on the state of the art NLP components
- A framework in which new scenarios can be deployed with minimal change in components

by relying on transfer learning and weak supervision

2 Related Work

Virtual human-based simulations can provide meaningful, deliberate practice for learning a wide range of teaching skills during teacher preparation as well as extended practice for advanced skills for in-service teachers. Dialogue systems for conversational agents can be built in two different approaches: End-to-end approaches that combine all of the dialogue system stages require large-scaled labeled training data and component-based systems which require less data, but each component needs to be trained separately. Component-based systems have the advantage of allowing certain components to be reused for similar contextual scenarios. Our work utilizes the latter approach leveraging unstructured data collected from web and textbooks as knowledge bases. Our dialogue policy is a combination of handcrafted rules by education domain experts, and dialogue states are tracked through a reading comprehension based approach highlighted by [Gao et al. \(2019\)](#). The most common approach for dialogue state tracking is the "slot-value" pairs approach. In this scenario, different stages of the dialogue are often framed as a multi-class classification task ([Mrkšić et al., 2015](#)). While this approach is well-studied, the decision to use the reading comprehension based approach in this system is based on the ability to incorporate pre-trained models for dialogue state tracking.

The assessment component for the pre-service teachers' questioning strategy is an adapted version of IQA. The IQA has been developed by the Learning Research and Developmental Center at the University of Pittsburgh since 2002 ([Matsumura et al., 2006](#)). The IQA offers a holistic assessment of mathematical instruction, including the academic rigor of the specific tasks and student and teacher discussion surrounding the task ([Pianta and Hamre, 2009](#)). The IQA has been further validated by its developers in subsequent research ([Boston and Candela, 2018](#)). More recent developments suggest that the IQA can be used not only as an assessment tool but also as a feedback tool to help teachers actively improve their instruction ([Boston and Candela, 2018](#)).

The questions that teachers ask are essential for promoting students' meaningful mathematical discourse. The academic rigor component ([Junker](#)

et al., 2005; Boston, 2012) of the IQA builds on earlier classifications of teacher questions (e.g. Boaler and Brodie (2004), distinguishes between "probing and exploring" questions that ask students to clarify their ideas or the connections between them, and "procedural and factual" questions that elicit facts or yes/no responses). The IQA is intended to be used in contexts where cognitively demanding mathematical tasks are implemented and is well suited for fine-grained teacher professional development such as that which focuses on teacher questioning (Boston et al., 2015). Another advantage of utilizing IQA is the limited categories as a high number of categories results in very complex dialogue policies (Yan et al., 2017).

Given the limited amount of time domain experts may have for annotating data, several methods to improve label efficiency were explored. Weak supervision techniques, as highlighted by Ratner et al. (2016) provides the two-fold benefit in that it requires less human labeling than would otherwise be required for training. An additional benefit of Weak supervision is that noisy data and the accuracy of each annotator can be taken into account for classification. Ratner et al. (2016) has shown that weak supervision systems are better than generic majority vote approaches. Noisy label data for model classification has also been studied by deep-learning-based approaches (Guan et al., 2018) and proven effective.

3 Data and Tools

3.1 Classification Model Data

A primary purpose of this conversational agent is in its ability to provide feedback for pre-service teachers. This requires the ability to classify each statement or question using the selected assessment rubric which in this implementation is an adapted IQA rubric. The categories of the adapted IQA measure were set by education domain experts and iterated on over the course of several months for the purposes of classification and feedback for pre-service teachers. The adapted IQA measure includes the following categories of questions:

- **Probing and Exploring:** Clarifies student thinking, enables students to elaborate their own thinking for their own benefit and the class. Points to underlying mathematical relationships and meanings and makes links among mathematical ideas. (e.g., Explain to me how you got that expression?)

- **Procedural and Factual:** Elicits a mathematical fact or procedure; Requires a yes/no or single response answer; Requires the recall of a memorized fact or procedure (e.g., What is the square root of 4?)
- **Expository and Cueing:** Provides a mathematical cueing or mathematical information to students. (e.g: To solve this problem you need to double this side, then take that number and multiply it by 3.)
- **Other** This refers to all other conversations not related to the above topics.(e.g: Close your books, Why didn't you use graph paper?)

Annotators also had the opportunity to flag any data as a "data issue" which would represent a transcript preprocessing error or another issue indicating the data could not be labeled such as incoherent text or blank text.

The data used for the adapted IQA evaluation rubric was developed from transcriptions of audio recordings of teachers in whole-class and teacher-student conversations that took place in elementary mathematics classrooms using different mathematics curricula across the United States. The de-identified dataset was shared from an NSF-sponsored project that had previously collected the recordings to answer separate research questions. Students engaged with a project purposed to help them understand different geometry concepts like scale factor, dimensions, surface area, and volume of rectangular prisms. The students recorded the observations from a given visualization and explained the impact of the scale factor. The data collected for the development of this scenario contained 2826 questions. The unique question along with the context, or speaking turn, in which the question was uttered were both provided as reference for the annotators to use during labeling.

We had 5799 total labeled data instances. There were five total annotators: three expert teachers as well as two pre-service teachers. The total number of annotators fluctuated during different stages of the annotation process resulting in varied amounts of labels generated by each annotator. The time to label each data point averaged between 5.2 to 6.7 seconds per annotator. The total number of unique labeled sentences was 2826. The total distribution of labels between the four assessment categories ranged from 856 to 2133. We used weak supervision based approaches to combine the labeled

Reference Information

Context

['Alright nice job, thank you very much.', 'Alright Joelle please give her a high five, cheer her on, thank you for leading that for us today.', 'Alright, so this morning our job was to fill out our post it note.', 'Alright, we were to write our own definition for area, our own definition for perimeter, because we've been talking about this for about the past four days now.', 'Alright and we're going to be jumping in to how to measure volume today.', 'Alright?', 'What did our target say about volume?', 'Are we measuring liquid volume today?', 'What did our target say?', 'I like that a lot of you are turning around to refer to that.', 'What are we measuring?', 'Arian?']

Question

What did our target say?

- Probing and Exploring^[1]
- Procedural and Factual^[2]
- Expository and Cueing^[3]
- Other^[4]
- Data Issues^[5]

Select order in case you have a preference

↻ Update

Task ID: 5636952

Figure 1: Labeling interface for annotation

data from multiple annotators over majority vote approaches.

3.2 Labeling Platforms

Two labeling platforms were used extensively for this project: Labelbox (Labelbox, 2020) and Label Studio (Tkachenko et al., 2020). While both platforms were simple and straightforward to use, Label Studio enabled the building of custom user interfaces with several improved features such as keyboard shortcuts that allowed annotators to more easily onboard and complete labeling tasks more efficiently.

Each individual question was labeled with a context reference that allowed annotators to see the entire speaking turn of the teacher. The decision to include context came after previous iterations of labeling questions resulted in an inter-annotator agreement of below 0.50, which subsequently increased to 0.66 after including context. An example of the Label Studio labeling interface is shown in Figure 1.

Our data collection approach relied on weak supervision and learning with noisy labels strate-

gies. In this paradigm noisy labels acquired either through human labels or machine learning models are cost effective to acquire. In the scenario in which domain expert annotators are available (in our case, expert teachers), noisy disagreements between annotators can be leveraged to build high accuracy models (Ratner et al., 2016; Guan et al., 2018). Weak supervision approaches are scalable, enabling easy adaptation to multiple mathematical scenarios, one of the key contributions and focuses of this project.

3.3 Knowledge Base Data

The knowledge-base of dialogue systems can be very complex depending on the scenario for which the dialogue system is being built (Yan et al., 2017). For this initial demonstration scenario the conversational agent represents a student with some level of understanding of the topic scale factor. The knowledge base relies on unstructured knowledge about scale factor collected from the web and textbooks in which text compiled is in the format of plain text. As our system is intended to reflect a students understanding of a topic, which is reasonably im-

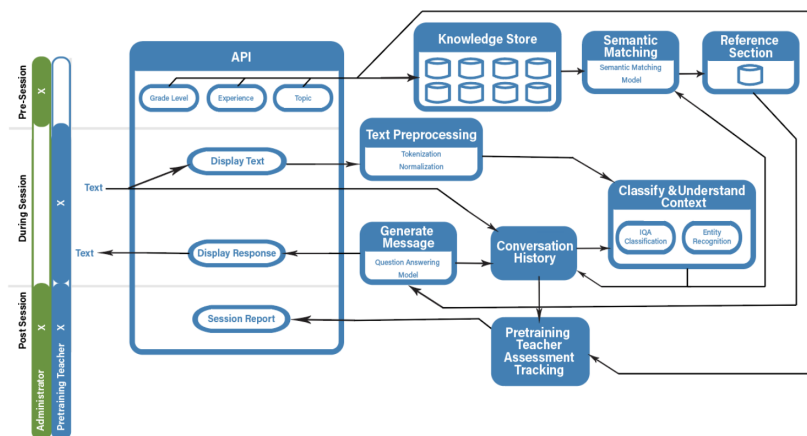


Figure 2: Conversational Agent Implementation Architecture

perfect, contradictory sources of information are not a primary concern. In fact a knowledge base with contradictory information may be leveraged to support more robust answering. Additionally, the expected level of understanding of a student for a given topic is likely to be documented in instructional materials readily available on the web and therefore collecting this data is a simple way to develop a knowledge base. Future efforts will address identifying grade-appropriate filtering that may improve interaction by generating a more realistic student profile with a grade-reflective knowledge-base.

The text collected from the web was not cleaned, labeled, or annotated. Basic pre-processing included removing references to figures or hyperlinks to other web pages. Once the plain text reference base was compiled, we then separated the text into sections of no more than 512 words. This processing step was done so that an entire section could be directly used as the input in the response generation discussed further in the methods section.

Relying on unstructured knowledge bases is critical to rapidly developing and deploying new conversational agent scenarios. Unstructured texts on varying mathematical topics are readily available from websites and video transcripts. Our framework allows for a simple way to incorporate newly generated external knowledge bases as a way to scale to additional scenarios. The ability to use unstructured knowledge as the key input of our knowledge base is possible due to the recent advances in question-answering models, reading comprehension tasks, and readily available libraries such as Huggingface Transformers (Wolf et al., 2020).

3.4 Platform Interface and Hosting

The application was built with Django, a Python web application development framework. Credentials must be generated and are required prior to using the interface. The interface includes an Institutional Review Board (IRB) consent form as well as a description of the scenario to include topic, expected student understanding of topic (Beginner, Intermediate, Advanced), and student grade level. Screenshots of the application are included in the Appendix.

4 Methods

The overall architecture of our conversational agent is depicted in Figure 2. As discussed, a central component of our demonstration is evaluating and providing feedback of pre-service teacher instruction. For this initial version of the scenario, domain experts provided a specified rubric for pre-service teachers to meet that clarifies the types of IQA categories desired within a session. One sample rubric evaluates if the teacher is asking at least one "probing and exploring" question, one "expository and cuing" question, and one "procedural and factual" question. This rubric can be changed easily in the demonstration through a separate JSON file. In our current evaluation, we do not evaluate the order in which the questions are asked, but we plan to include more sophisticated evaluation protocols in future work.

The current implementation incorporates text interactions between a conversational agent, representing a student, and a pre-service teacher. The pre-service teacher can interact with the conversational agent by providing new knowledge and

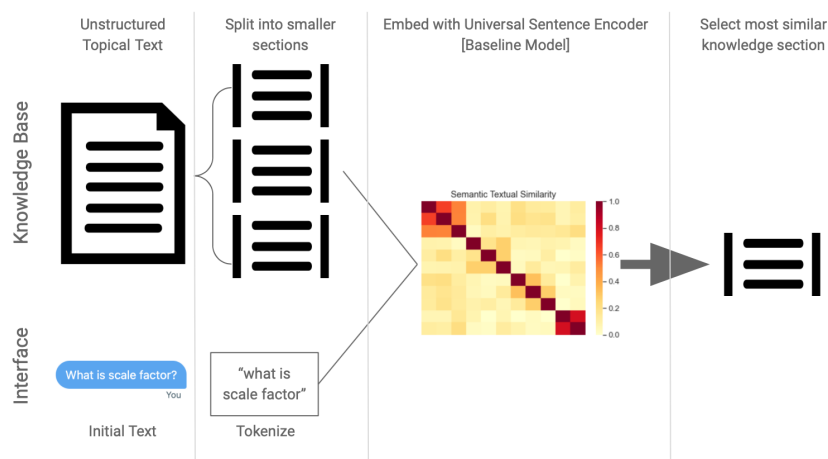


Figure 3: Dialogue System Selection of Knowledge Base Reference Section via Semantic Similarity

testing understanding as well as testing knowledge of the topic “scale factor”. The pre-service teacher types a statement, question, multiple statements or multiple questions in the text box, and the conversational agent responds by taking into account the conversation context as well as the pre-service teacher’s utterance. Each component of the dialogue system is described in detail in the subsequent subsections.

4.1 Assessment Metric Classifier

In our demonstration, we utilize the entire speaking turn of the pre-service teacher as the input text. This formulation is very similar to the two-sentence classification task like what is used in the Stanford Natural Language Inference corpus (Bowman et al., 2015). We frame our input in a similar format for classification and fine-tuning. The input text undergoes basic cleaning and is tokenized prior to being used as the input to our classifier model. We experimented with multiple text classification approaches to include Convolutional Neural Network (CNN)—based text classification (Kim, 2014), Long Short-Term Memory (LSTM)—based text classification (Liu et al., 2016), and newer approaches that rely on Transformer Architectures (Devlin et al., 2018; Liu et al., 2019) and perform well with small amounts of labeled data. Transfer learning models tend to perform well with less labeled data than other models because of the pretraining with unsupervised text that encodes knowledge and semantic meaning of words and sentences. This demonstration incorporates a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model for our adapted IQA classification task.

4.2 Semantic Matching

For the conversational agent to respond to input text, the first step is to identify the most relevant section of the knowledge base. To do this, the pre-processed input text is used in combination with the Universal Sentence Encoder (Cer et al., 2018) to find the most relevant, or more specifically, semantically similar section of the knowledge base. Semantic similarity refers to identifying the degree to which two texts have the same meaning. As discussed in the data treatment section, the knowledge base section is split into smaller sections that are a more optimized size for the semantic similarity tool used: the Universal Sentence Encoder. This process is depicted in Figure 3.

The Universal Sentence Encoder is optimized for short phrases or paragraphs and outputs a 512-dimensional vector. Semantic similarity computation is accomplished by computing the inner product between the input and knowledge base text. The semantic similarity between generated embeddings was computed using normalized cosine similarity of the embeddings. Semantic similarity computation at the sentence level is more accurate than the aggregate of word-level similarities and is therefore preferred in this application. Models trained to understand words in context are often better suited for identifying semantic similarities of phrases and sentences. In application, we may take input such as “What is scale factor?” By finding the most semantically similar section in the knowledge base, we can use this section to input the response generation.

If the text’s semantic similarity and the knowledge base sections do not achieve a pre-defined

threshold, the system responds from the unknown category. For this demonstration the threshold was set to 0.80 after empirical evaluation of semantic coherence. This threshold value will be further tested and empirically evaluated in future iterations of this system.

In the unknown category, one of the six random responses (pre-defined) is selected to convey to the pre-service teacher that the system does not understand the user input. This pre-defined selection represents a hand-crafted dialogue policy that was determined by domain experts.

If the semantic similarity is greater than or equal to the threshold for a given knowledge base section, the system then selects this knowledge section. The initial input text along with the selected knowledge base section are used as the inputs to a question-answering module. The question-answering module is a pre-trained BERT model that is fine-tuned on the Stanford Question Answering Dataset (SQuAD). This module is then used to generate a response in the user interface for the subsequent turn of the conversation. The dialogue manager retains the dialogue states of the conversation for record and reference within the conversation.

4.3 Dialogue Manager

4.3.1 Dialogue State Tracking

Dialogue State Tracking is a core component of the dialogue system. The goal of the dialogue state tracking system is to estimate the goal at each turn of the conversation. There are multiple formulations of dialogue state tracking systems like hand-crafted rules (Wang and Lemon, 2013) and web-style ranking (Williams, 2014). In our approach we use the most recent question-answering paradigm for dialogue state tracking (Gao et al., 2019). Unlike Gao et al. (2019) we do not train our reading comprehension based model, but instead use the question-answering paradigm to understand the dialogue states. We append each of the pre-service teacher utterances along with the student responses. Since conversational agent responses can either be mathematical (evaluating an expression) or responding to a pre-service teacher utterance, we evaluate each stage as a question-answer task. So at the end an utterance questions prompted include: “Did the teacher ask a probing and exploratory question?”, “Did the conversational agent answer the question correctly?”, “Did

the pre-service teacher acknowledge the answer?”. This iterative framework of question answering helps keep track of the dialogue state. Since this is a task-specific dialogue system being evaluated for a specific mathematical scenario, we are already aware of the dialogue states of the conversation. At each turn in the conversation, we use all the previous pre-service teacher utterances to determine the current dialogue state of the conversation.

4.3.2 Dialogue Policy

In our task-specific dialogue system, our dialogue policy is rule-based. Depending on the dialogue state accomplished up to turn n , our utterance at $n + 1$ depends on the dialogue states accomplished up to that point. The hand-crafted rules for our dialogue policy also enable the use of direct evaluation based on simple metrics such as the number of questions in each adapted IQA category or development of metrics like Initiate Response Evaluate (IRE) (Mehan, 1979).

4.4 Response Generation

The response generation component extracts relevant sections of the knowledge base as a question-answering task. A question-answering task, also referred to as a reading comprehension task, is a supervised learning problem, where given a segment of text of i tokens, a question of j tokens, it returns an answer segment of k tokens. The answer in question-answering tasks can be cloze-style as in CNN/Daily Mail (Hermann et al., 2015), span prediction (like SQuAD (Rajpurkar et al., 2016)), or be similar to NarrativeQA (Kočíšký et al., 2018). We retrieved our knowledge from semantic matching of web-text categories and thus our response generation pipeline matched closely to span prediction tasks. We implemented the response generation pipeline using the transformers library (Wolf et al., 2020), where a BERT model (Devlin et al., 2018) was fine-tuned on the SQuAD dataset. We did not fine-tune our question-answering system for the response generation module, rather relying on semantically-matched unstructured data sections to be used as inputs in generating answers to questions.

4.5 Session Feedback

All text input by the pre-service teacher is retained as well as the associated adapted IQA category classification. The compilation of classifications of the pre-service teachers’ input texts are provided in

an assessment report which is used as feedback for the pre-service teacher at the end of the session.

5 Results

This section will discuss the step by step interaction with the developed interface that is developed. Only one scenario currently exists and is defined as a 4th-5th grade average understanding of the mathematics topic scale factor.

5.1 IQA Classifier

To train the IQA classifier, we used BERT (Devlin et al., 2018) and DistillBERT (Sanh et al., 2019) for classification based on the open source Transformers (Wolf et al., 2020) implementation. We trained the classifiers on 80% of the data and sectioned the remaining data as a 10% validation set and 10% for the test set. The validation set accuracy for the BERT and DistillBert models were 75.8 and 74.3 respectively. Since, the performance increase with Bert was minimal we used Distillbert for faster inference.

5.2 Walk Through

This platform is currently hosted on Python Anywhere. Pre-generated credentials for instructors are generated for a given user ID. Each user uses the unique ID to create an associated password. Following the prompts the user is able to read through the IRB consent form and sign in to access the conversational interface. Figure 4 shows the first conversational agent interaction screen as well as an example inputs where the pre-service teacher may ask several questions to assess the conversational agents current level of understanding.

The pre-service teacher, or user, can then interact with the conversational agent by typing questions or statements. The Appendix contains additional screen captures of the developed interface. Currently the system only supports text-based conversation.

An example of a text snippet that does not meet the set semantic matching threshold with the generated conversational agent response is demonstrated in Figure 4. There are instances when the conversation breaks despite the teacher re-framing their questions or input differently. This is intended to reflect how a student may sometimes fail to answer a properly framed question or statement appropriately.

Once the session is completed, the pre-service

teacher receives a session report which incorporates the adapted IQA rubric and the number of their interactions that would be classified within each of the sections. They are also provided a report of the entire conversation that is shared with expert teachers as well in an effort to identify how the conversational agent stages can be designed better.

6 Conclusion and Future Work

Our goal in this paper is to demonstrate the implementation of a conversational agent with very little training data that relied on foundational and well-studied metrics like IQA. By leveraging state of the art modules for natural language processing and deep learning we could build a functional prototype that is now going to be used as a pilot for training for a specific mathematical scenario of "scale factor". By integrating pre-trained models such as SQuAD, BERT, and the Universal Sentence Encoder as well as using weak supervision approaches in data treatment we have leveraged minimal amounts of domain-expert-labeled data and knowledge base data in order to create a usable interface. Approaches like this help evaluate pre-service teachers in a scalable fashion and can also be deployed across the web for large scale participation. There are several areas we are pursuing to improve this interface in order to provide a more robust interface as well as a more useful assessment tool. Some of the current features that we are working on are as follows:

- **Improved IQA model:** We plan to continue collecting domain-expert labeled data that can then be used to improve the trained classification model. The improved classification model will better reflect a realistic assessment of pre-service teacher sessions using the IQA-developed categories as a rubric.
- **Increased Knowledge Base:** This first scenario is limited to a specific knowledge base on the topic "Scale Factor". We plan to incorporate a wider variety of data related to expected knowledge on this topic (from 4th grade to 10th grade) as well as associated topics such as perimeter, volume, and ratio. With these knowledge bases compiled we will be able to test techniques that can assist in generating responses that are representative of a limited understanding of the desired topic to a

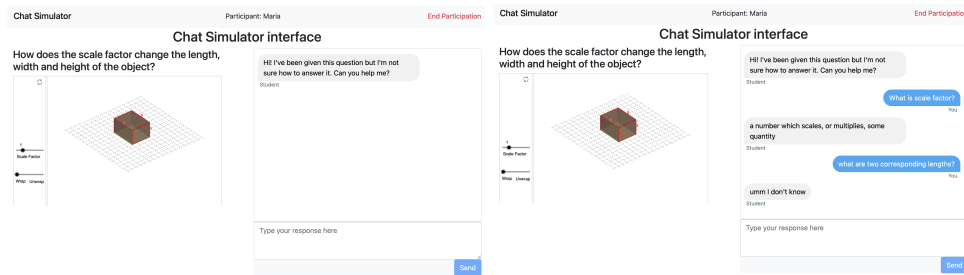


Figure 4: Conversational Agent Interface Screen. Left image shows initial screen. Right image shows interaction example where conversational agent has identified a statement that does not meet semantic matching threshold within the knowledge base.

more advanced level of understanding on the topic.

- **Response Generation:** Currently there are some features within the response that appear to represent the way a student is more likely to respond. We would like to further develop these features by incorporating more student-like speech features. This would allow for more realistic conversational agent interaction and result in less formal or textbook-like responses. Additionally, current responses generated are most coherent when responding to a question which is not necessarily reflective of all student-teacher interactions. Future developments are planned to improve the robustness of responses to better account for the different forms of inputs.

Finally, we plan to be deploy this tool with a group of pre-service teachers under the direction of expert teachers in order to test the qualitative aspects and realism of this system.

Acknowledgments

This work was funded in part by the Robertson Foundation. The authors wish to acknowledge the use of de-identified classroom dialogue from NSF 1535024.

References

- Jo Boaler and Karin Brodie. 2004. The importance, nature and impact of teacher questions. In *Proceedings of the twenty-sixth annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, volume 2, pages 774–782.
- Melissa Boston. 2012. Assessing instructional quality in mathematics. *The Elementary School Journal*, 113(1):76–104.
- Melissa Boston, Jonathan Bostic, Kristin Lesseig, and Milan Sherman. 2015. A comparison of mathematics classroom observation protocols. *Mathematics Teacher Educator*, 3(2):154–175.
- Melissa D Boston and Amber G Candela. 2018. The instructional quality assessment as a tool for reflecting on instructional practice. *ZDM*, 50(3):427–444.
- Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.
- Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. [Universal sentence encoder](#).
- Nalin Chhibber and Edith Law. 2019. Using conversational agents to support learning by teaching. *arXiv preprint arXiv:1909.13443*.
- Debajyoti Datta, Valentina Brashers, John Owen, Casey White, and Laura E Barnes. 2016. A deep learning methodology for semantic utterance classification in virtual human dialogue systems. In *International Conference on Intelligent Virtual Agents*, pages 451–455. Springer.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tur. 2019. Dialog state tracking: A neural reading comprehension approach. In *Proceedings of the 20th Annual SIG-Dial Meeting on Discourse and Dialogue*, pages 264–273.
- Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. 2018. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701.
- Brian William Junker, Yanna Weisberg, Lindsay Clare Matsumura, Amy Crosson, Mikyung Wolf, Allison Levison, and Lauren Resnick. 2005. *Overview of the instructional quality assessment*. Regents of the University of California.
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.
- Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.
- Labelbox. 2020. [Labelbox: Training data platform](#).
- Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.
- Lindsay Clare Matsumura, Brian Junker, Yanna Weisberg, and Amy Crosson. 2006. Overview of the instructional quality assessment.
- Hugh Mehan. 1979. ‘what time is it, denise?’: Asking known information questions in classroom discourse. *Theory into practice*, 18(4):285–294.
- Nikola Mrkšić, Diarmuid O Séaghdha, Blaise Thomson, Milica Gašić, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2015. Multi-domain dialog state tracking using recurrent neural networks. *arXiv preprint arXiv:1506.07190*.
- Robert C Pianta and Bridget K Hamre. 2009. Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational researcher*, 38(2):109–119.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.
- Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In *Advances in neural information processing systems*, pages 3567–3575.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Pavel Smutny and Petra Schreiberova. 2020. Chatbots for learning: A review of educational chatbots for the facebook messenger. *Computers & Education*, page 103862.
- Maxim Tkachenko, Mikhail Malyuk, Nikita Shevchenko, Andrey Holmanyuk, and Nikolai Liubimov. 2020. [Label Studio: Data labeling software](#). Open source software available from <https://github.com/heartexlabs/label-studio>.
- Zhuoran Wang and Oliver Lemon. 2013. A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information. In *Proceedings of the SIGDIAL 2013 Conference*, pages 423–432.
- Jason D Williams. 2014. Web-style ranking and slu combination for dialog state tracking. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 282–291.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jian-she Zhou, and Zhoujun Li. 2017. Building task-oriented dialogue systems for online shopping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Appendix

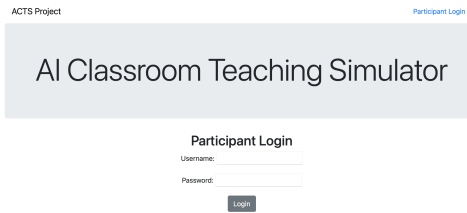


Figure 5: Conversational Agent Session Example: Login

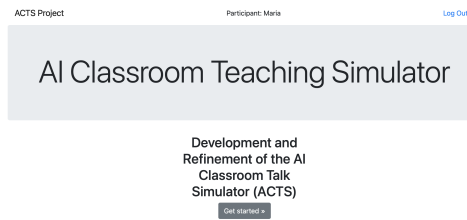


Figure 6: Conversational Agent Session Example: Acknowledgement

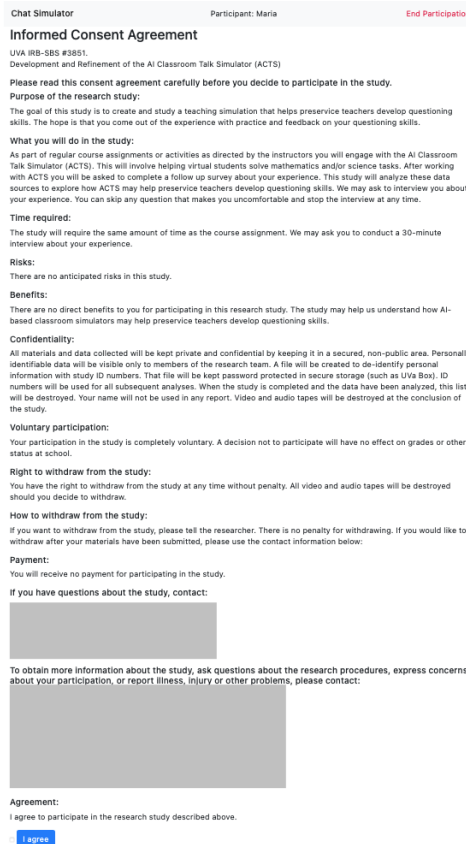


Figure 7: Conversational Agent Session Example: Institutional Review Board consent

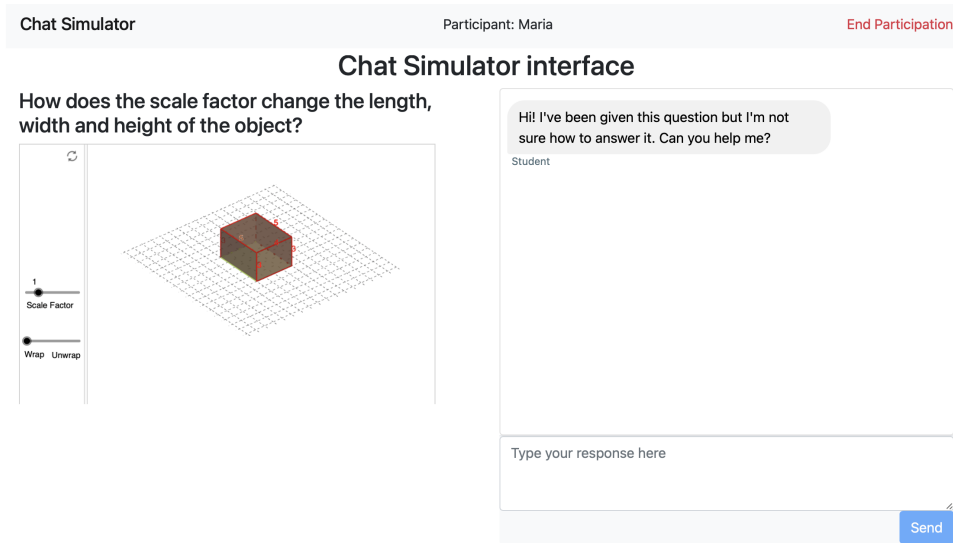


Figure 8: Conversational Agent Session Example: Initial Session Screen

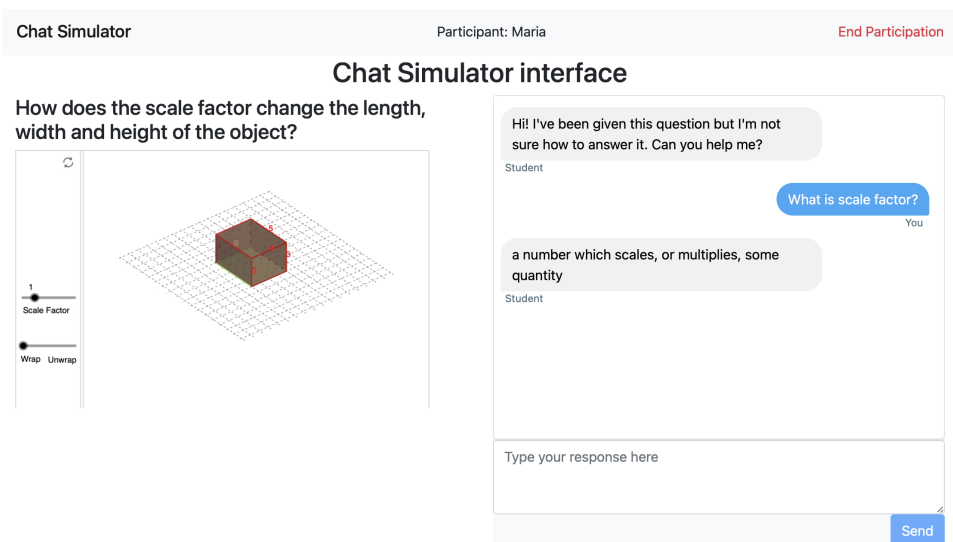
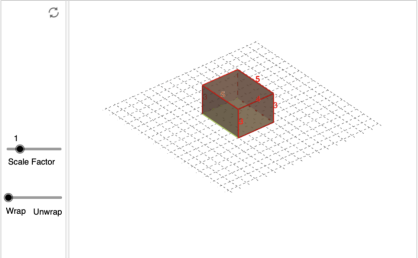


Figure 9: Conversational Agent Session Example: Testing Scale Factor Knowledge

Chat Simulator Participant: Maria End Participation

Chat Simulator interface

How does the scale factor change the length, width and height of the object?



Hi! I've been given this question but I'm not sure how to answer it. Can you help me?
Student

What is scale factor?
You

a number which scales, or multiplies, some quantity
Student

what are two corresponding lengths?
You

umm I don't know
Student

Type your response here

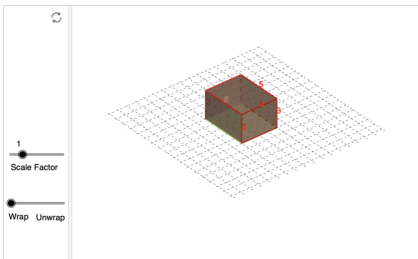
Send

Figure 10: Conversational Agent Session Example: Improperly Phrased Question

Chat Simulator Participant: Maria End Participation

Chat Simulator interface

How does the scale factor change the length, width and height of the object?



What is scale factor?
You

a number which scales, or multiplies, some quantity
Student

what are two corresponding lengths?
You

umm I don't know
Student

what is the ratio of two corresponding lengths?
You

scale factor
Student

Type your response here

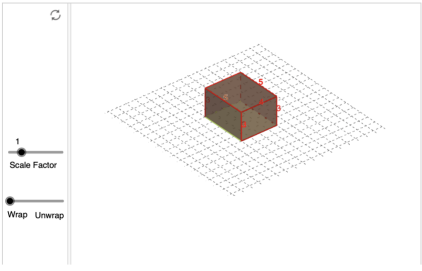
Send

Figure 11: Conversational Agent Session Example: Properly Phrased Question

Chat Simulator Participant: Maria [End Participation](#)

Chat Simulator interface

How does the scale factor change the length, width and height of the object?



Student

what are two corresponding lengths?
You

umm I don't know
Student

what is the ratio of two corresponding lengths?
You

scale factor
Student

what is the height of Mt. Everest?
You

[blank stare]
Student

Type your response here

[Send](#)

Figure 12: Conversational Agent Session Example: Irrelevant Question