Curio SmartChat : A system for Natural Language Question Answering for Self-Paced K-12 Learning

Srikrishna Raamadhurai **Ryan S Baker*** **Vikraman Poduval**
Saal.ai, Al Bateen Offices, Abu Dhabi 112230
*University of Pennsylvania, PA 19104
{srikrishna, vikram} @saal.ai, ryanshaunbaker@gmail.com

Abstract

During learning, students often have questions which they would benefit from responses to in real time. In class, a student can ask a question to a teacher. During homework, or even in class if the student is shy, it can be more difficult to receive a rapid response. In this work, we introduce Curio SmartChat, an automated question answering system for middle school Science topics. Our system has now been used by around 20,000 students who have so far asked over 100,000 questions. We present data on the challenge created by students' grammatical errors and spelling mistakes, and discuss our system's approach and degree of effectiveness at disambiguating questions that the system is initially unsure about. We also discuss the prevalence of student "small talk" not related to science topics, the pluses and minuses of this behavior, and how a system should respond to these conversational acts. We conclude with discussions and point to directions for potential future work.

1 Introduction

Question asking is an important part of students' classroom learning. Through asking questions, students can clarify their confusions, address their doubts, and explore a topic in greater depth. Student questions, when framed appropriately, can form an important tool for learning in Science and other domains (Chin & Brown, 2002).

However, this same type of learning support is not available when students are working at home. Even in a classroom setting, teachers may not be able to answer all student questions, much less to say about some shy students who do not even register their questions in class. Increasing numbers of students now spend class time working one-on-one with adaptive learning platforms (Baker, 2016), and in these contexts, multiple students may have questions at the same time, and teachers may not be able to answer all questions at the same time (Schofield, 1995).

This challenge has led to the idea of automated question answering systems in education (Louwerse et al., 2002; Corbett et al., 2005; Milik et al., 2006; Jin et al., 2018), where students can ask questions in natural language. Different than simply a search engine, educational question answering systems attempt to provide answers focused on current content, set at an appropriate level for the student's current stage of learning. An 8th grader with a question about the Krebs Cycle needs different types of information than an undergraduate Biology major, for example.

However, despite research into the possibility of automated question answering in education, there has been little effort to scale these systems, with considerably more energy going into tutor-led tutorial dialogue systems (Wolfe et al., 2013; Ventura et al., 2018).

Building such a system is non-trivial for several reasons, first and foremost the complexity that arises from handling unforeseen queries that represent considerable variability in the use of human language. Several challenges must be solved in order for an automated question answering system to be optimally effective. It must recognize which questions are germane and which are off-topic (see, for instance, Corbett et al., 2005), and decide how to respond. It must be able to handle students' grammatical errors and spelling mistakes (a challenge in all NLP-based learning systems – see Chollampatt & Ng, 2017). It must be able to map from often ill-formed questions to the content in those questions. It must provide content at the right educational level.

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 336–342
Florence, Italy, August 2, 2019.

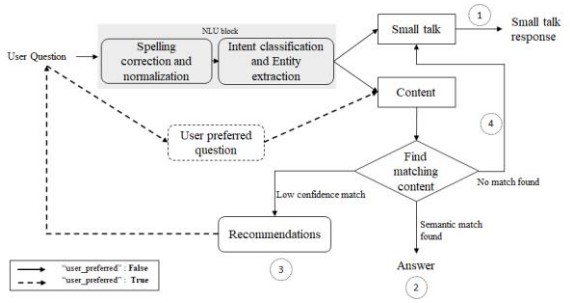

Figure 1: Architecture of QA Engine

In this paper, we discuss our system, Curio SmartChat for self-paced K-12 learning through Question Answering as a mode and is currently being used by around 20,000 students, who have asked over 100,000 questions, in the domain of middle school Science. K-12 stands for Kindergarten through 12th grade in many countries while some others refer to this as "All-through-school". While we only present our system in practice for middle school which forms a part of K-12 system, we believe the exact framework can be extended to serve across other grades provided the content is available in a similar format, as the focus of our system is more technology oriented (Question Answering) rather than content oriented. Within this paper, we will present data on the challenge created by students' grammatical errors and spelling mistakes, and discuss our system's approach and degree of effectiveness at disambiguating questions that the system is initially unsure about. We also discuss the prevalence of student "small talk" not related to Science topics, the pluses and minuses of this behavior, and how a system should respond to these conversational acts. We focus on our efforts to address these challenges, towards developing a system that can effectively give the right response to a student question, and thereby help them to progress rather than becoming frustrated or stuck (Beck & Rodrigo, 2014).

In the next section we will describe our system architecture and the QA engine's workflow in more detail. Section 3 will explain the challenges faced. Section 4 will present the discussion, followed by potential directions for future work.

2 System Description

Our system architecture comprises of three blocks: a semantic match engine (referred to as the QA engine), a content library and a web browser-based client for user interaction. The client is a simple chat interface with a text input field. The content library is where our entire collection of curriculum based text documents, metadata of pictures and other media exist. Our QA engine handles all the information processing tasks.

Let us go through a typical work flow and possible outcome scenarios. The user inputs his or her query in the text field. Based on the system's understanding of the user query, it tries to retrieve the *Answer* from the content library. When the system encounters complex user queries which are difficult to comprehend, it alternates to offering *recommendations*, to try to disambiguate what the student is asking. *Recommendations*, unlike *Answer* are a list of possible questions from the question bank that closely matches the initial user query. However, when a query has no potential recommendations with sufficiently high probability, the system responds with small talk: off-topic exchanges such as system level guidance, greetings, weather, sports and so on. The level of our small talk content has been designed to suit our target users, who are around 13-16 years old. At the moment, our small talk service is simple and stateless, meaning it does not remember the sequence of exchanges to respond to the query at hand.

2.1 Content library

The content library in this study pertains to middle school Science topics. The library includes a compilation of text documents and quick definitions collected based on the curriculum. This library also contains questions and answers tagged according to three levels of Bloom's (1956) Taxonomy: *Knowledge*, *Understanding* and *Application*. Content such as definitions are labeled as *Knowledge* since they could be understood without any other prerequisite (ex: *"What is energy?"*). *Understanding* level content are those where the students can relate to what they learned from *Knowledge* (ex: *"Cutting a tree with an axe is very easy. Why?"*). *Application* level content allow the students to test their understanding by way of more practical scenarios (ex: *"How do we separate oil from water?"*).

2.2 QA Engine

Figure 1shows the architecture of the QA engine which is the main component of our system. When a user asks a questions, the engine checks for

spelling mistakes and does spelling correction and spelling normalization including replacing contractions in informal English. For example, '*What's energy?'* is converted into '*What is energy*?'. As a next step the system predicts if the user is interested in the content or small talk.

We have trained our own custom taggers for intent and entity extraction by extending SpaCy taggers (https://spacy.io/). The intent classifier decides which service will provide the response, small talk or content library, while the entity extractor will retrieve the entities the user is interested in. For example, if the query is "*What is photosynthesis?*", then the tagged JSON would look like {"intent": "*content",* "entity": ["*photosynthesis"*]}. There could also be more than one desired entity but only one intent per query. If the query is assessed to be content related, the system then will look to retrieve the answer from the content library through a combination of semantic matches. Our main search methodology includes a Vector Space approach to look for related concepts in our content library to find out candidate responses.

A naive, search system would look for keywords (entities), however those methods suffer from out-of-vocabulary problem and cannot detect paraphrases. More recent Information Retrieval systems have moved to employing word vectors. Popular word vectors such as Word2Vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014) provide a fixed size representation for words, in a sense attempting to capture their *meaning* in the language space by providing synonymous words with similar vectors. Word vectors have been shown to be superior to simple keyword approaches towards understanding syntactic similarities (Mikolov et al., 2013b). However, there are still some shortcomings in terms of processing unknown words. A heuristic approach has been suggested to handle this problem by way of randomly initializing such unknown words (Sutskever et al., 2014). There are still concerns with respect to word sense disambiguation, however. For example, the word "mean" could be a Verb, Adjective or a Noun based on the sentence structure. Word vectors usually only offer one representation towards a word. To address the problem of polysemy, a model called sense2vec was trained as a deep bidirectional language model (Trask et al., 2015).

In our work we have used sentence level encoders instead of word level encoders. Sentence level encoders, similar to word vectors, provide a fixed size representation for an entire sentence instead of individual words. In principle, the embedding of a sentence and its paraphrase should be vectorially similar in a target language space even if those two sentences use different words to convey the same idea. In our system, we use a pretrained model released by Google called Universal Sentence Encoder (Cer et al., 2018) to detect paraphrases. We also use a combination of hash map lookups besides paraphrase detection to make the retrieval faster and scalable.

If the probability of our candidate response does not pass the confidence check, the system dynamically offers recommendations to the student that are conceptually related to that particular query. As our system consists of a Deep Learning model in production, we have made use of the Tensorflow framework and Docker containerization which are best practices in the industry for developing scalable, production grade software.

3 Data

Since the time of launching the service, the system has served over 100,000 questions from around 20,000 students, mostly 13-16 years old. We focus our analyses on the quality of the served responses. User logs comprising the input user query, the response (either direct answer or small talk) and/or the recommendations have been collected.

Any user query could have one of the three possible outcomes as shown in Figure 2; (i) A direct answer obtained through exact or semantic match, (ii) Recommendations, (iii) Small talk exchange.

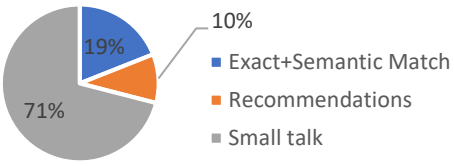

Figure 2: Distribution of Responses

Assessing the correctness of (i) is straightforward and it means that the query was understood by the system and had passed all the necessary confidence checks. Given our architecture's two-step

Content Available	Count	Mean valid recommendation (count) Rater 1	Mean valid recommendation (count) Rater 2
No	28	0.70 (19)	0.64 (18)
Yes	107	0.55 (59)	0.73 (78)
Query not clear	65	0.49 (31)	0.47 (29)
	200	**0.55**	**0.63**

Table 1: Quality of Recommendations

winnowing approach as shown in Figure 1, a drop in confidence for direct answers alternates to recommendations and an even lower confidence defaults to small talk. Thanks to our carefully chosen parameters for checks, we have hardly ever found the system presented a wrong answer to the user. Hence we are only left with (ii) and (iii) to be evaluated as follows.

To evaluate the quality of recommendations, we randomly chose 200 sample queries and employed two human raters to independently rate the recommendations. As for small talk analysis, we analyze only the misspelt queries linguistically and share our findings.

3.1 Analyses of quality of recommendations

The purpose of our analyses is to check the validity of our recommendations. The raters were asked to rate every single recommendation for each query as valid/ invalid depending on what was asked by the user. For a given user query even if one recommendation from the list of recommendations was validated by the rater, we count that as valid and represent the mean grouped by content availability in Table 1.

The reader is required to note that recommendations could happen due to two primary reasons; either the user asked a query where the system lacked content about or the query could not be clearly disambiguated by the system. We have observed that the system very rarely responded with wrong answers when it lacked content, instead it responded with recommendation, thanks to the confidence check.

With the above method of estimation, the average quality of recommendations by both raters is around 60% which could already offer a good level of user engagement. It is also important to note that despite human raters finding it difficult to precisely understand the purpose of 65 out of 200 user queries (32.5%), the system had offered valid recommendations to retain the users' interest. Such a situation in class would have required the student having to rephrase the question until the teacher was confident of responding. For example, a user query as "*as the pressure increases, the solubility of solids and liquids"* is vague but also implies that the student is interested in knowing more about *solubility* and *saturated solution.* The system offered recommendations along these lines.

Category	User Query
Insertion	you are frome wich contry
	can you tell me what acide are the produce
	what is nonpoar moleculens
Omission	what is mixtur
	wate is a substsnce with a sour taste
	wat is meen saturated solution
Substitution	what is a substence
	what are the eyeglasses called that habe both concave and convex lenses
	yas i neeed halp
Inversion	waht is meen pressure
	waht maen acid ph
Duplication	can u tell me about definning solubility?
	wich termes deescribe a substance that has an equal amount of hydronium and hydroxide ions
Phonetic errors	but i don't find the anser
	Hi. Wat your name
	were you are

Table 2: Analyses of Spelling mistakes

3.2 Analyses of Spelling mistakes

We present the analysis of common spelling mistakes our users make by manually categorizing at a character level. Our scheme for categorization follows linguistic cues such as insertion, omission, substitution, inversion and so on.

We find that a majority of students tend to heavily depend on phonetics to spell words, with examples of student misspellings such as *"anser"*, *"meen"*, *"wich"*. Among these queries, it appears that the silent characters have almost consistently been dropped such as *"w"* and *"h"* in *"anser"* and *"wich"* respectively. Since "a" and "e" have different forms of phonetic presence in the English language, students tend to interchangeably use these characters. Learning to correctly spell words is lacking with our users which makes serving

User 1 says	Curio SmartChat responds
What is a mammal?	Mammals are animals that give birth to babies rather than laying eggs and feed their young with milk. Examples of mammals include human beings, dogs, whales and bears.
What is an animal?	An animal is a living organism that feeds on organic matter. It typically has specialized sense organs and a nervous system, which makes it able to respond rapidly to stimuli.
I want to just chill	Please ask me about middle school topics in Science.
fine	Okay.
can you explain bird	A bird is a warm-blooded vertebrate with feathers.
User 2 says	**Curio SmartChat responds**
Dim	Dim means not shining brightly or clearly.
good	Let's just try again, we'll get it right eventually.
:(	! You are not too nice.
i mean :)	Never mind. Just kidding.
what does meiosis mean	The chromosome number is halved from diploid to haploid, resulting in genetically different cells. This process of reduction division is called meiosis.

Table 3: Good examples episodes of small talk and default response

answers a lot more challenging. In practice, smarter text input devices correct user input in real time, although depriving the users of learning from their mistakes. Please see Table 2 for more analyses on spelling mistakes.

3.3 Analyses of small talk

Although the primary focus of Curio SmartChat is to talk about content, our work follows recommendations from Bickmore and Cassell (2000) to use small talk as a means to create student-system rapport that ultimately builds student willingness to engage in learning with the system. While education researchers have long argued that off-topic conversation takes time away from learning (Carroll, 1963), there is evidence that small talk is associated with more effective collaboration in human-human learning (Kreijns, 2004). Similar rapport has been created by conversational agents (Crossley & Kostyuk, 2017). In our work we employed small talk along with recommendations to gently nudge the user into content. Though users of Curio SmartChat do engage in small talk with the system, they have continued to produce content-based queries as well. So far we have observed that roughly one in

User 3 says	Curio SmartChat responds
I am nicer than u	I would have to agree.
U should agree	I guess I can appreciate that.
no ur not	Are we debating?
ya we r	Please ask me about middle school topics in Science.
b bbbbdjv cvmnsnv vdklv sd ,bfvheeubrh	Please ask me about middle school topics in Science.
User 4 says	**Curio SmartChat responds**
Hello! How can I help you?	Hey! That's my line!
Hey! That's my line!	Please ask me about middle school topics in Science.
v	Please ask me about middle school topics in Science.
;vojoevjverfd	Please ask me about middle school topics in Science.

Table 4: Bad example episodes of small talk and default response

every three queries are still content based as shown in Figure 2.

As seen in systems that use wizard of oz approaches to generate small talk (e.g. Crossley & Kostyuk, 2017), students develop social relationships with the system, explicitly asking Curio SmartChat questions about its family, friends and hobbies. When a question is beyond the capacity of Curio SmartChat to answer, a default response- *"Please ask me about middle school topics in Science"* is provided. This default response has seen mixed follow-up reactions from the students. Some students gracefully react to this default response by returning to asking about the content (~22.97%) as shown by examples in Table 3 while other students appear to become upset or respond with nonsense strings of letters (~29.66%) as shown by examples in Table 4. Given the scope of this paper, we will not psycho-analyze the user behavior, hence we simply report our findings.

4 Challenges

There are several technical challenges involved in developing and maintaining a chat service of this nature for students. Students do not always provide grammatically correct queries. Especially in the UAE where Curio SmartChat is primarily used, English is the language of instruction for Science but is not the native language. Hence good modules for spelling correction and spelling normalization are necessary to handle misspelt user queries. Every student has his or her own way of phrasing a question, however the response to a particular question has to be consistent across all students unless the input is irrecoverably broken. Even after spelling correction and normalization, there are still inputs that cannot be even understood by

human raters. Some of these appear to represent nonsense strings that were never intended to communicate (see Table 4) but others may represent difficulty in communicating ideas, sometimes due to lack of mastery in English-language communication, and sometimes due to the difficulty of the Science content and ideas. These utterances would be difficult for any system to parse accurately. It would be better to develop a mechanism where the students learn to properly spell alongside auto-correction rather than the system overriding the user with correct replacements.

Our system as of now either offers recommendations or responds with small talk when it does not completely understand what the user is asking. It is not very clear as to what is the best way to serve more content based queries as against small talk between building user models or developing stateful dialog managers at this scale. As with any chat service, some users tend to use profanity and insults. There are still some doubts about how to best deal with such inputs in the context of an education chat agent.

5 Conclusion

We introduced Curio SmartChat, our Natural Language Question Answering system for K-12 learning and analyzed its performance while serving over 100,000 queries for around 20,000 middle school students on Science topics. Curio SmartChat is capable of performing both content based and off topic conversations with students. Given the scope of the system we have analyzed the user queries for spelling mistakes, off topic chats and validity of offered recommendations. The system is able to either directly answer or at the very least offer relevant recommendations to the users at least 60% of the time. We showed that even when humans were not able to precisely understand the queries, the system was still able to provide relevant recommendations 50% of the time thereby saving the time for both students and teachers alike. We only expect such benefits to grow with more content and better spelling correction mechanisms added to our system as future work. As we have shown the most common forms of spelling mistakes students make, developing such systems could be crucial for improved quality of answer retrieval. The pluses and minus of having a default response appear to be roughly similar, in other words not very harmful. Perhaps there are smarter ways of nudging the student back into content that could make the experience more productive.

Acknowledgments

We would like to extend our heartfelt thanks to our content team, the independent raters and our linguist towards ensuring a successful, timely submission.

References

Baker, R.S. 2016. *Stupid Tutoring Systems, Intelligent Humans*. International Journal of Artificial Intelligence and Education, 26 (2), 600-614.

Beck, J., & Rodrigo, M. M. T. 2014, June. *Understanding wheel spinning in the context of affective factors*. In International conference on intelligent tutoring systems (pp. 162-167). Springer, Cham.

Bickmore, T., & Cassell, J. 2000. "*How about this weather?" Social dialogue with embodied conversational agents*. Proceedings from the American Association for Artificial Intelligence (AAAI) Fall Symposium. North Falmouth, MA: AAAI Press.

Bloom, B. S. 1956. *Taxonomy of educational objectives. Vol. 1: Cognitive domain*. New York: McKay, 20-24.

Carroll, J. 1963. *A Model For School Learning*. Teachers College Record, 64, 723-733.

Cer, Daniel, et al. 2018. *"Universal sentence encoder."* arXiv preprint arXiv:1803.11175.

Chin, C., & Brown, D. E. 2002. *Student-generated questions: A meaningful aspect of learning in science.* International Journal of Science Education, 24(5), 521-549.

Chollampatt, S., & Ng, H. T. 2017, September. *Connecting the dots: Towards human-level grammatical error correction.* In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (pp. 327-333).

Corbett, A., Wagner, A., Chao, C. Y., Lesgold, S., Stevens, S., & Ulrich, H. 2005, May. *Student questions in a classroom evaluation of the ALPS learning environment.* In Proceedings of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology (pp. 780-782). IOS Press.

Crossley, S., & Kostyuk, V. 2017. *Letting the Genie out of the Lamp: Using Natural Language Processing*

tools to predict math performance. In International Conference on Language, Data and Knowledge (pp. 330-342). Springer, Cham.

Jin, L., King, D., Hussein, A., White, M., & Danforth, D. 2018. *Using Paraphrasing and Memory-Augmented Models to Combat Data Sparsity in Question Interpretation with a Virtual Patient Dialogue System.* In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 13-23).

Kreijns, K. 2004. *Sociable CSCL environments: Social Affordances, Sociability, and Social Presence.* Unpublished doctoral dissertation, Open University of the Netherlands, The Netherlands.

Louwerse, M. M., Graesser, A. C., Olney, A., & Tutoring Research Group. 2002. *Good computational manners: Mixed-initiative dialog in conversational agents.* In Etiquette for Human-Computer Work, Papers from the 2002 Fall Symposium, Technical Report FS-02-02 (pp. 71-76).

Mikolov, Tomas, et al. 2013. "*Distributed representations of words and phrases and their compositionality.*" Advances in neural information processing systems.

Mikolov, Tomas, et al. 2013. "*Efficient estimation of word representations in vector space*." arXiv preprint arXiv:1301.3781.

Milik, N., Marshall, M., & Mitrovic, A. 2006, June. *Responding to free-form student questions in ERM-Tutor*. In International Conference on Intelligent Tutoring Systems (pp. 707-709). Springer, Berlin, Heidelberg.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "*Glove: Global vectors for word representation.*" Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).

Schofield, J. W. 1995. *Computers and classroom culture*. Cambridge University Press
Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. "*Sequence to sequence learning with neural networks.*" Advances in neural information processing systems.

Trask, Andrew, Phil Michalak, and John Liu. 2015. "*sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings.*" arXiv preprint arXiv:1511.06388.

Ventura, M., Chang, M., Foltz, P., Mukhi, N., Yarbro, J., Salverda, A. P., … & Marvaniya, S. 2018, June. *Preliminary Evaluations of a Dialogue-Based Digital Tutor*. In International Conference on Artificial Intelligence in Education (pp. 480-483). Springer, Cham.

Wolfe, C. R., Widmer, C. L., Reyna, V. F., Hu, X., Cedillos, E. M., Fisher, C. R., ... & Weil, A. M. 2013. *The development and analysis of tutorial dialogues in AutoTutor Lite.* Behavior research methods, 45(3), 623-636.