iTA: A Digital Teaching Assistant

Vishnu Dutt Duggirala, Rhys Sean Butler and Farnoush Banaei-Kashani
Department of Computer Science and Engineering, University of Colorado Denver, U.S.A.

Keywords: Machine Reading Comprehension, Question-answering.

Abstract: Designed and implemented a question-answering chatbot, dubbed iTA (intelligent Teaching Assistant), which can provide detailed answers to questions by effectively identifying the most relevant answers in "long" text sources (documents or textbooks). iTA answers questions by implementing a two-stage procedure. First, the topmost relevant paragraphs are identified in the selected text source using a retrieval-based approach and scores for the retrieved paragraphs are computed. Second, using a generative model, extracted the relevant content from the top-ranked paragraph to generate the answer. Our results show that iTA is well suited to generate meaningful answers for questions posed by students.

1 INTRODUCTION

Online learning offers flexibility and availability, allowing students to continue their studies, even when it is difficult or impossible to receive in-person instruction. However, online learning creates challenges for tasks that have traditionally been carried out in-person. Guidance counseling, producing course feedback, and tutoring can all be more difficult without human interaction. The problems with online learning have heightened the interest in making digital tools that can help students face these challenges. Recently, chatbots have been designed that attempt to mimic human interaction.

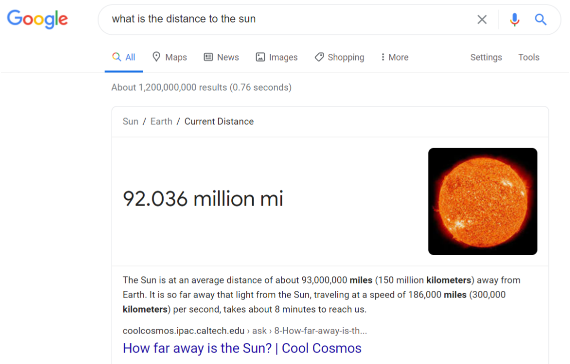
Chatbots can support students in several ways, but most have been designed to answer administrative questions or function like customer support systems. They can serve as a tutor which explains a topic in a chapter and quizzes the student. So far, research has not focused on developing a chatbot that can produce nuanced, multi-sentence responses in topic-specific domains. This encouraged us to create a tool where it takes the role of teaching assistant to help students understand academic concepts. *iTA: intelligent Teaching Assistant*, a tool that acts as a Teaching Assistant. iTA can provide detailed responses to user queries within a specific knowledge domain.

Machine Reading Comprehension (MRC), or the ability to read and understand the unstructured text and then answer questions about it, remains a challenging natural language processing task motivated by a wide variety of applications. For example, as shown in Figure 1, a search engine with MRC capabilities can return the correct response to users' questions in natural language instead of a progression of related web pages.

To develop iTA, we overcame two challenges: The first challenge is that most MRC solutions focus on comprehension from a passage no larger than 500 words. In contrast, a typical textbook may contain more than 15,000 words. MRC models, such as Match LSTM (Wang and Jiang, 2016), a basic comprehension neural network, are not designed to process large text bodies. To achieve a question-aware context representation without early summarization, BiDAF (Seo et al., 2018) is used. On the other hand, BERT (Devlin et al., 2019) is introduced to reduce computation time and extend machine reading comprehension beyond what LSTM can accomplish. That said, BERT still fails given documents longer than 512 tokens. The second challenge was producing a model that provides in-depth, easily understandable responses. Current question answering datasets provide extractive and, short responses. Student questions will usually require multi-sentence responses. One-word or two-word responses will not adequately help students. MS MACRO v2 (Bajaj et al., 2018) is introduced to address this problem by generating responses, with an average of 13.6 words. TriviaQA (Joshi et al., 2017) has multi-sentence support, but its answers are shorter than most of the datasets.

iTA is a system that can function as a digital teaching assistant. Toward this end, we have adopted and adapted a developed model for multiple passages

Duggirala, V., Butler, R. and Banaei-Kashani, F.
iTA: A Digital Teaching Assistant.
DOI: 10.5220/0010461002740281
In *Proceedings of the 13th International Conference on Computer Supported Education (CSEDU 2021) - Volume 2*, pages 274-281
ISBN: 978-989-758-502-9



what is the distance to the sun

About 1,200,000,000 results (0.76 seconds)

Sun / Earth / Current Distance

92.036 million mi

The Sun is at an average distance of about 93,000,000 **miles** (150 million **kilometers**) away from Earth. It is so far away that light from the Sun, traveling at a speed of 186,000 **miles** (300,000 **kilometers**) per second, takes about 8 minutes to reach us.

coolcosmos.ipac.caltech.edu › ask › 8-How-far-away-is-th...

How far away is the Sun? | Cool Cosmos

Figure 1: An example of a Google search engine with an MRC technique.

MRC tasks (Clark and Gardner, 2017). With iTA, when a user has asks a question, the five paragraphs that contain the most relevant information for the question are selected. This process is completed using TF-IDF. Second, we use BART (Lewis et al., 2019), pre-trained on the ELI5 dataset (Fan et al., 2019) to generate an answer for the given question. BART is a denoising auto-encoder designed to train sequence to sequence models. ELI5 is a dataset that has a series of questions that have been answered with explanatory, multi-phrase responses. Here is an example of an iTA generated response for a given question:

Question: How is range different from an array?
Answer: A range is an array of numbers in increasing or decreasing order, each separated by a regular interval. An array is a collection of numbers separated by an interval. A range can be used to represent a very large number of numbers, so it's worthwhile to use it. Ranges are useful in a surprisingly large amount of cases, so they are worthwhile to be used in a very small number of cases. The difference is in how they are used. The range is used to describe the order of the numbers in the array, while the array is used as a representation of the order in which the numbers are in the range.

2 RELATED WORK

In this section, we briefly review the literature on MRC methods, and more generally, Conversational AI, as well as systems similar to iTA. Conversational AI represents a similar research space because it focuses on tasks like natural language retrieval and generation. The retrieval-based response is where the agent's answer is based on pattern matching and a question is matched to a pool of message-response candidates. The response whose message has the highest matching score with the question is selected. The generative-based response does not select or choose from any predefined response but instead creates a new one. A generative-based response can be more specific and instructive, while retrieval-based responses tend to be more grammatically correct because they are manually produced. iTA uses both retrieval and generation methods to evaluate questions and delivers the response.

2.1 MRC Methods

There has been a large amount of research produced on MRC methods that use a single paragraph to generate responses. That is to say when the response that is generated comes from a short passage. Existing work have explored developing various architectures such as Match-LSTM (Wang and Jiang, 2016), R-Net (Wang et al., 2017), and designing multiple attention mechanisms such as BiDAF (Seo et al., 2018), Transformers (Vaswani et al., 2017) to achieve more precise and improved answers. BERT (Devlin et al., 2019) has produced state-of-the-art results on 11 NLP tasks, which includes a single-paragraph MRC task.

A few researchers have focused on multi-passage question-answering. Longformer (Beltagy et al., 2020) was developed in response to the token limitation of BERT (Devlin et al., 2019). Longformer uses an attention mechanism which changes linearly with the sequence length. This network (Zhang et al., 2018) took the advantages of hierarchical-attention (Wang et al., 2018) to learn the paragraph level representation and implement the match-LSTM (Wang and Jiang, 2016) mechanism.

2.2 Educational Chatbots

Some education chatbots focus on assisting the user by producing a "Yes" or "No" response to a question. Conversational Agents to Promote Children's Verbal Communication Skills(Fabio Catania and Garzotto, 2020) have developed as a system that allows a 9-year-old kid to create a bitmoji by talking to a chatbot. Every time they responded to a question, they show the visual representation of the feature described directly on the avatar in the GUI. At every step, it asks the user if they are satisfied with the avatar. Suppose the answer is "No" or any negative expression. It will remove the feature and goes to the previous intent. Adaptive Conversations for Adaptive Learning: Sustainable Development of Educational Chatbots (Donya Rooein, 2020), helps students explain course modules and provide a reference for students on the total course progress. In this paper (Skjuve, 2020), Skjuve explains various stages of a student's learning progress. Students can use our chatbot in a similar way to this paper.

Tutor presented by the (Hobert, 2019) uses a predefined learning path. First, a student receives some instruction on a specific topic. Then, Tutor will ask the students questions about the instruction they have just received. The student is also asked if they need any additional content for better understanding of the concept.

FIT-EBot (Hien et al., 2018) is an administrative support chatbot that produces information about course registration, course score, prerequisite courses, exam schedules, and other administrative information about a course. None of the aforementioned chatbots are capable of producing detailed answers to a set of dynamic questions. iTA provides this functionality and can produce dynamic responses for user-generated questions.

3 METHODOLOGY

A mixed methodological approach was used in this study. iTA uses a paragraph selection module (Clark and Gardner, 2017) to extract the highest scored passage and a sequence-to-sequence model for answer generation module that uses BART (Lewis et al., 2019). Below, in Section 3.1 an overview of the system is discussed. In Section 3.2, we will go through an in-depth review of each model in our system.

3.1 System Overview

iTA supports multi-paragraph generative-based responses, avoids noisy labels, and uses an unstructured text as its source. We use a Data Science textbook, stored as a .txt file as the source from which answers are generated. Figure 2 depicts the overall architecture and data flow in iTA. iTA is a two-tier model which has a passage selection module and answer generation module. We have scrubbed the data in the text "The fundamentals of Data Science" (Adhikari and John DeNero, 2019), by removing the mathematical equations, tables, and python code from the text. When a student asks a question, the question and the textbook are passed through TF-IDF in the first module to select the top 5 paragraphs which contain information most relevant to the question. This step is employed to save the computationally expensive step of calculating each paragraph's confidence score in the entire text document. We use the top 2 candidate paragraphs in BART model to generate the response.

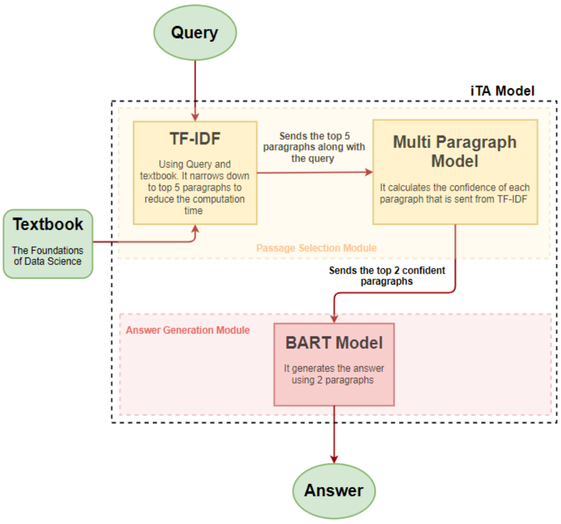

Query
iTA Model
TF-IDF
Using Query and textbook. It narrows down to top 5 paragraphs to reduce the computation time
Sends the top 5 paragraphs along with the query
Multi Paragraph Model
It calculates the confidence of each paragraph that is sent from TF-IDF
Textbook
The Foundations of Data Science
Passage Selection Module
Sends the top 2 confident paragraphs
Answer Generation Module
BART Model
It generates the answer using 2 paragraphs
Answer

Figure 2: Overview of iTA.

3.2 Model

3.2.1 Passage Selection Module

The paragraph selection module selects the paragraphs with the smallest TF-IDF cosine distance with a user-generated question. iTA uses a modified method for calculating TF-IDF. iTA currently uses only one-long format document to generate responses. When iTA calculates TF-IDF, instead of using the term-frequency from a large number of documents, iTA's term-frequency is calculated using the frequency of each term within the Data Science textbook. This approach gives more weight to the least common words in the question and extracts the relevant paragraphs. iTA's implementation uses an attention mechanism. This mechanism, presented in detail in (Clark and Gardner, 2017), computes a confidence score for each paragraph containing a relevant passage and selects the best passages based on the confidence score.

The weight in TF-IDF is a statistical measure used to evaluate how important a word in a collection of documents or corpus. The weights are a multiplication of TF, IDF. The definition for TF-IDF in our approach (Clark and Gardner, 2017) is the number of times a term occurred in a paragraph and divided by the total number of paragraphs.
GloVe(Pennington et al., 2014) has been used to vectorize the question and context for each student question. A shared bi-directional GRU(Cho et al., 2014) maps the context to the question. Attention mechanisms build a representation between question and context. Self-attention mechanisms enhance machine comprehension of complex contexts. The last layer of

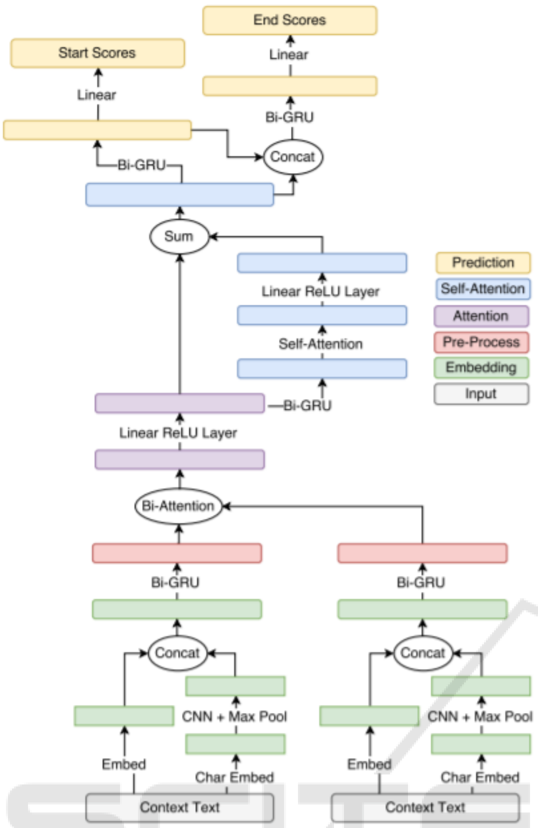


Figure 3: Multi-Paragraph Reading Comprehension architecture; figure courtesy of (Clark and Gardner, 2017).

iTA's first module applies a softmax operation to calculate the relevant passages' confidence scores. The highest scored paragraphs are sent to the iTA answer generation module.

3.2.2 Generative Module

iTA's language module (Wolf et al., 2020) was pretrained using BART (Figure 4) on the ELI5 dataset (Fan et al., 2019). BART (Lewis et al., 2019) is a sequence-to-sequence model with a bidirectional encoder similar to BERT (Devlin et al., 2019) BART uses an auto-regressive decoder as with GPT (Radford, 2018).
iTA modules are made up of Transformers (Vaswani et al., 2017). BART uses multi-headed-attention. This allows for a reliable sequence to sequence modeling. Using the multi-headed-attention mechanisms allows for greater accuracy in understanding context and generating responses. BERT encoder randomly replace the tokens with masks, and missing tokens predicted independently, so cannot use it for generation.

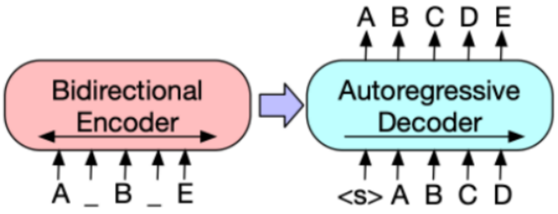


Figure 4: BART architecture; figure courtesy of (Lewis et al., 2019).

BART decoder uses transformer decoder block; its decoder has an extra layer which is masked self-attention which does not allow a position to peak at tokens to its right. The key difference with the transformer decoder is that it outputs one token at a time, just like traditional language models that use auto-regression. It uses a beam search to create the response. Larger beam widths result in better performance for iTA. Multiple candidate sequences increase the likelihood of better matching a target sequence. Performance is inversely related to beam-width.

4 EXPERIMENTATION

4.1 Datasets

Extractive datasets such as SQuAD (Rajpurkar et al., 2018), CoQA (Reddy et al., 2019), HotpotQA (Yang et al., 2018) restrict responses to a word or short expression from the input, TiviaQA (Joshi et al., 2017) offers which challenges the models to perform reasoning across multiple paragraphs with an average of 2895 words in a document but the answer is still short. ELI5 (Fan et al., 2019) has overcome this long answer challenge. It can be seen in Table 1 that TriviaQA answers are drawn from are long-form documents. However, TriviaQA responses are limited to one or two words. TrivaQA was used to train the language model of iTA, and ELI5 was used to train the generative-based response model of iTA.

Table 1: Comparison of QA datasets and how ELI5 is better; table courtesy of (Fan et al., 2019).

	Average # of Words			
Dataset	Question	Document(s)	Answer	#Q-A Pairs
ELI5 (Fan et al., 2019)	**42.2**	857.6	**130.6**	**272K**
MS MARCO v2 (Bajaj et al., 2018)	6.4	56	13.8	183K
TriviaQA (Joshi et al., 2017)	14	**2895**	2.0	110K
CoQA (Reddy et al., 2019)	5.5	271	2.7	127K
SQuAD (2.0) (Rajpurkar et al., 2018)	9.9	116.6	3.2	150K
HotpotQA (Yang et al., 2018)	17.8	917	2.2	113K

TriviaQA (Joshi et al., 2017) is used to train the paragraph selection module with the shared-normalization confidence method (Clark and Gardner, 2017), stated that the shared-normalization confidence method fetched more transparent results than the merge method. To test our application, we used a data science textbook (Adhikari and John DeNero, 2019) as a long document. Pre-processing is done on the data to remove mathematical equations and any high-level coding language syntax.

4.2 Experimental Setup

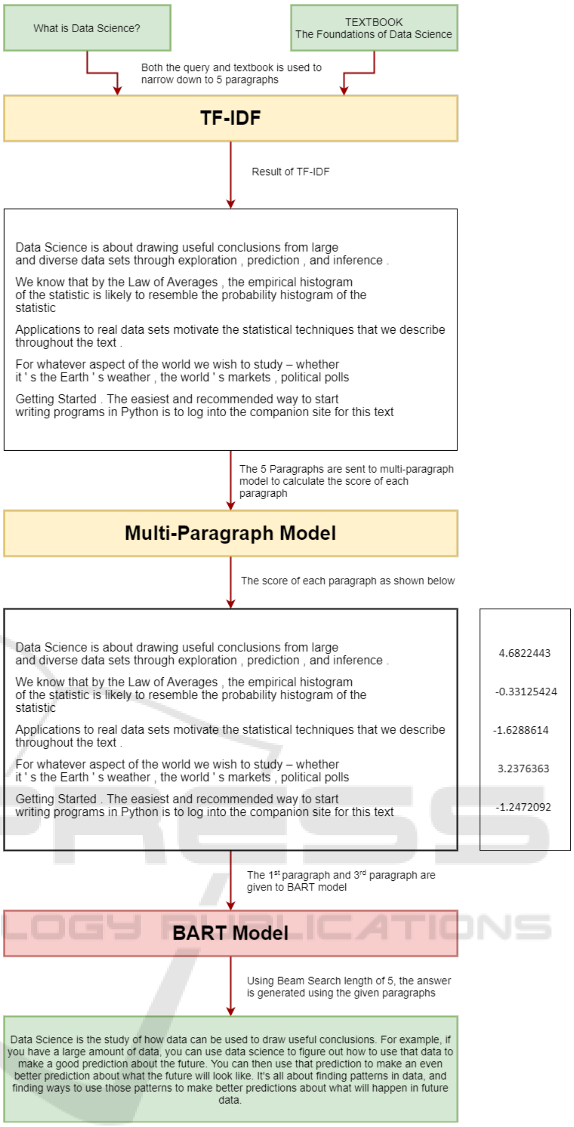
The iTA model was trained using an Adadelta optimizer with a batch size of 60 and used TriviaQA-unfiltered data. This choice was made because the dataset does not specify which documents contain the answer. In Clark (Clark and Gardner, 2017). Three different approaches are used in calculating confidence scores for relevant paragraphs. In Simple and Effective Multi-Paragraph Reading Comprehension (Clark and Gardner, 2017), they noted that the shared-norm approach gave superior results for paragraph retrieval. Once we had this trained model and the BART model, we sent our test dataset (textbook), which is prepossessed to remove the stop words as they do not contribute to calculating TF-IDF and a question to fetch the top five paragraphs using the TF-IDF approach. Forwarded to the shared-norm model to get each paragraph's confidence, here, more noisy labels will have less confidence value. BART generative model takes the top two high confidence value paragraphs to get a detailed answer. The chatbot is run on cloud infrastructure, leveraging 12 G.B. of Nividia Tesla K80.

4.3 Results

BLEU score (Papineni et al., 2002), a metric used for evaluating machine-translated text, was used to assess the iTA generated responses. It measures the similarity of the machine-generated text to a set of human-created reference text. A value of 0 means that the output has no overlap with the reference text (low quality). In comparison, a value of 1 means there is perfect overlap with the reference (high quality). The score between 0.3 to 0.4 is understandable to good translations. For easy interpretations, we have multiplied with 100. Perplexity is one more quantitative measure that calculates how well a probability distribution predicts a sample.

We use both BLEU scores and perplexity to evaluate the strength of the iTA responses.

We gathered the top two paragraphs, each with 400 words from the passage selection module, and fed it to the generated answer model, which uses beam search, and we set beam length to 5, the minimum size of the answer to 96. We calculated the BLEU score and Perplexity for each response according to the parameters mentioned earlier.


What is Data Science?
TEXTBOOK
The Foundations of Data Science
Both the query and textbook is used to narrow down to 5 paragraphs
TF-IDF
Result of TF-IDF

Data Science is about drawing useful conclusions from large and diverse data sets through exploration , prediction , and inference .

We know that by the Law of Averages , the empirical histogram of the statistic is likely to resemble the probability histogram of the statistic

Applications to real data sets motivate the statistical techniques that we describe throughout the text .

For whatever aspect of the world we wish to study – whether it ' s the Earth ' s weather , the world ' s markets , political polls

Getting Started . The easiest and recommended way to start writing programs in Python is to log into the companion site for this text

The 5 Paragraphs are sent to multi-paragraph model to calculate the score of each paragraph
Multi-Paragraph Model
The score of each paragraph as shown below

Data Science is about drawing useful conclusions from large and diverse data sets through exploration , prediction , and inference .

We know that by the Law of Averages , the empirical histogram of the statistic is likely to resemble the probability histogram of the statistic

Applications to real data sets motivate the statistical techniques that we describe throughout the text .

For whatever aspect of the world we wish to study – whether it ' s the Earth ' s weather , the world ' s markets , political polls

Getting Started . The easiest and recommended way to start writing programs in Python is to log into the companion site for this text

4.6822443
-0.33125424
-1.6288614
3.2376363
-1.2472092
The 1st paragraph and 3rd paragraph are given to BART model
BART Model
Using Beam Search length of 5, the answer is generated using the given paragraphs

Data Science is the study of how data can be used to draw useful conclusions. For example, if you have a large amount of data, you can use data science to figure out how to use that data to make a good prediction about the future. You can then use that prediction to make an even better prediction about what the future will look like. It's all about finding patterns in data, and finding ways to use those patterns to make better predictions about what will happen in future data.

Figure 5: Data flow of iTA.

Explanation of how our two-step process works: When a question is asked, "What is Data Science?" for example, it is sent to the iTA's first module. There TF-IDF uses the entire corpus (textbook) along with the submitted question to select 5 paragraphs that may contain the answer. These five paragraphs are sent to the Multi-Paragraph model to calculate the confidence of each paragraph. Choosing the highest two confident paragraphs, the BART model will generate the answer as shown in Figure 5.

The textbook "The Foundations of Data Science." is used as the text from which responses are generated. Four students with different majors asked questions that they came across while reading the textbook. We have chosen 75 significant questions, and graphs are plotted for perplexity, BLEU score, and time-taken for each response.

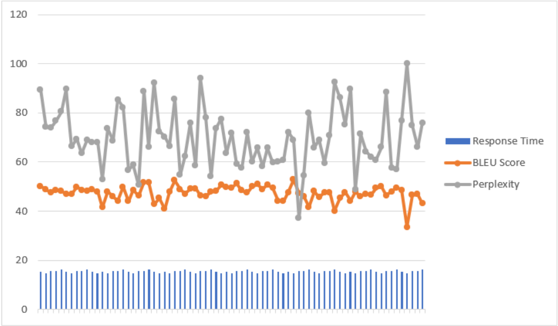


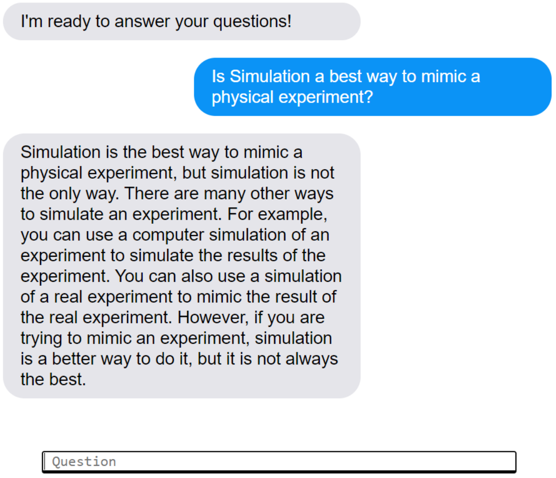
Figure 6: The Distribution of BLEU score, perplexity, and time taken to respond by the model for all 75 questions.

The average response time of iTA is 15 seconds. The distribution of BLEU score ranges between 40 and 50 as shown in Figure 6. This indicates that iTA can produce responses that human readers easily comprehend. The values of perplexity show that the text is understandable. Using human evaluation, Cohen's Kappa has been calculated by categorizing the generated response into correct, wrong, and acceptable answers. Acceptable is where the response has a grammatical error or repetitive sentence. The weighted kappa is 0.647 which is considered as a substantial agreement.

4.4 Discussion

iTA is one of its kind, so comparison with other systems is difficult. We measured the response time for each of iTA's two modules. The first module, the retrieval based model, has an average response time of 9.3 seconds. Once a question has been run through the retrieval model, the response-generation model takes over. The response generation module takes an average of 5 seconds to return a multi-sentence, explanatory answer. It is not computationally feasible to train an LSTM module on the ELI5 dataset. The Figure 7 shows the interface of iTA.

Experiments were initially performed on iTA using a set of 75 questions. Later, each question was rephrased and re-submitted to iTA. iTA generated similar answers for questions that have been rephrased. The BLEU score and perplexity of the responses to rephrased questions did not show any significant difference. Here is an example of iTA response to a question that asked in two different ways:



I'm ready to answer your questions!

Is Simulation a best way to mimic a physical experiment?

Simulation is the best way to mimic a physical experiment, but simulation is not the only way. There are many other ways to simulate an experiment. For example, you can use a computer simulation of an experiment to simulate the results of the experiment. You can also use a simulation of a real experiment to mimic the result of the real experiment. However, if you are trying to mimic an experiment, simulation is a better way to do it, but it is not always the best.

Figure 7: iTA interface.

Question: Why do we need Data Science?
iTA Answer: Data Science is the science of finding patterns in data. It's the study of how data can be used to find patterns in other data. For example, if you have a bunch of data, you can use data science to figure out how to use that data to make a prediction about what the next data is going to look like. You can then use that prediction to make an inference about what that next data will look like, and so on and so forth. This is how data science is used in the real world.
Re-phrased Question: What is the purpose of data science?
iTA Answer: Data Science is the study of data and the application of that data to real-world applications. It's not just about finding patterns in data, it's about finding ways to use those patterns to make an informed decision about what to do with that data. There are a lot of different types of data science, but they all boil down to the idea of using data to make informed decisions about how to use that data in real-life applications.

Limitations:

1. If any keyword in the question is not present in the textbook it takes the first paragraph of the document and generates the answer.
2. To answer the question, the context in the document should explicitly state the keyword.
   For example given explanation for Data type in the textbook is "*Every value has a type, and the built-in type function returns the type of the result of any expression.*" we pose the following question:
   Example Question: What are Data types?
   Generated Answer: A Data Type is a type of

number. For example, an integer is a number, a float is a float, a decimal is a decimal, and a hexadecimal is a hex. Data Types can also be used to describe types of numbers. For instance, if you have a number of digits, you can use a Data Type to describe what those digits are, and how they relate to each other. So, for example, if I have a set of numbers, I can use Data Types to describe how those numbers relate to one another.

3. No "*Data Type*" and "*Python language*" keyword are used in the same paragraph in the textbook. iTA still generates an explanatory response that matches the context of the question. The retrieval model has extracted a passage which contains "*Data Type*" tokens.
   Example Question: What are Data Types in Python language?
   Generated Answer: Data types are a way of representing data in a way that the programmer can understand. For example, if you have a list of numbers, you can use a data type to represent the number in the list. For instance, if I have a set of numbers and I want to know how many of those numbers are in that set, I could use a DataType to represent that number in a list. You can also use Data Types to represent data in other ways, for example if I wanted to know the age of a person, I can use the DataType for that

5 CONCLUSIONS

Comprehending large documents is a challenging task for current machine learning systems. Existing work focuses on systems that can generate responses given a small body of text to work with. Most available generative-based response systems generate one or two-word replies. Short replies are not always acceptable, and many questions that a user may want to have answered will require more than one or two words. iTA can produce a detailed, multi-sentence response to user-generated questions by combining a language comprehension module and a generative-based response module. Noisy labels are reduced while selecting relevant paragraphs. Finally, a generative model replies with an explanatory response.

6 FUTURE WORK

In the future, we will add more documents and allow users to choose from which document to ask questions. Moreover, one of iTA's modules uses LSTM. It may be possible to significantly increase iTA's performance by using Transformer based methods instead of LSTM. At present, the response time of iTA is about 15 seconds. Improvement of the response time will be the primary focus of future work.

REFERENCES

Adhikari, A. and John DeNero, B. (2019). The foundations of data science.

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., and Wang, T. (2018). Ms marco: A human generated machine reading comprehension dataset.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation.

Clark, C. and Gardner, M. (2017). Simple and effective multi-paragraph reading comprehension.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Donya Rooein, P. P. (2020). Adaptive conversations for adaptive learning: Sustainable development of educational chatbots.

Fabio Catania, Micol Spitale, G. C. and Garzotto, F. (2020). Conversational agents to promote children's verbal communication skills.

Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., and Auli, M. (2019). Eli5: Long form question answering.

Hien, H. T., Cuong, P.-N., Nam, L. N. H., Nhung, H. L. T. K., and Thang, L. D. (2018). Intelligent assistants in higher-education environments: The fit-ebot, a chatbot for administrative and learning support. In *Proceedings of the Ninth International Symposium on Information and Communication Technology*, SoICT 2018, page 69–76, New York, NY, USA. Association for Computing Machinery.

Hobert, S. (2019). Say hello to 'coding tutor'! design and evaluation of a chatbot-based learning system supporting students to learn to program.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *In EMNLP*.

Radford, A. (2018). Improving language understanding by generative pre-training.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for squad.

Reddy, S., Chen, D., and Manning, C. D. (2019). Coqa: A conversational question answering challenge.

Seo, M., Kembhavi, A., Farhadi, A., and Hajishirzi, H. (2018). Bidirectional attention flow for machine comprehension.

Skjuve, M. (2020). "from start to finish": Chatbots supporting students through their student journey.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

Wang, S. and Jiang, J. (2016). Learning natural language inference with lstm.

Wang, W., Yan, M., and Wu, C. (2018). Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1705–1714, Melbourne, Australia. Association for Computational Linguistics.

Wang, W., Yang, N., Wei, F., Chang, B., and Zhou, M. (2017). Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198, Vancouver, Canada. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Huggingface's transformers: State-of-the-art natural language processing.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. (2018). Hotpotqa: A dataset for diverse, explainable multi-hop question answering.

Zhang, Y., Zhang, Y., Bian, K., and Li, X. (2018). Towards reading comprehension for long documents. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4588–4594. AAAI Press.