Factoid and Open-Ended Question Answering with BERT in the Museum Domain

Md. Mahmud-uz-zaman[1], Stefan Schaffer[1], and Tatjana Scheffler[2]

[1] DFKI, Alt-Moabit 91c, 10559 Berlin, Germany
[2] German Department, Ruhr-Universität Bochum, Germany

Abstract. Most question answering tasks are oriented towards open domain factoid questions. In comparison, much less work has studied both factoid and open ended questions in closed domains. We have chosen a current state-of-art BERT model for our question answering experiment, and investigate the effectiveness of the BERT model for both factoid and open-ended questions in the museum domain, in a realistic setting. We conducted a web based experiment where we collected 285 questions relating to museum pictures. We manually determined the answers from the description texts of the pictures and classified them into answerable/un-answerable and factoid/open-ended. We passed the questions through a BERT model and evaluated their performance with our created dataset. Matching our expectations, BERT performed better for factoid questions, while it was only able to answer 36% of the open-ended questions. Further analysis showed that questions that can be answered from a single sentence or two are easier for the BERT model. We have also found that the individual picture and description text have some implications for the performance of the BERT model. Finally, we propose how to overcome the current limitations of out of the box question answering solutions in realistic settings and point out important factors for designing the context for getting a better question answering model using BERT.

Keywords: question answering · BERT · art museums.

1 Introduction

Despite recent technological advancement in conversational agents, the majority of museums still offer only prerecorded audio guides to visitors as an aid for a better experience. However, these audio scripts are long and visitors have no way to select information according to their needs. A study shows that use of chatbots can assist visitors better, educate them and help to improve their overall experience in museums [7]. A museum chatbot is different in nature from general chatbots because it is presented a picture and visitors are expected to ask questions related to the artwork. A comprehensive study with more than 5 thousand unique sessions in the Pinacoteca museum in Brazil was performed to discover the type of questions people ask using chatbots in a museum [2]. They

have found 8 types of questions are asked, including *fact, author, visual, style, context, meaning, play* and *outside*. Among these, *meaning* constitutes around 60% of the questions. We can additionally classify possible questions into factoid and open-ended, according to the amount of facts required to answer them [8]. According to this classification, open-ended questions are more frequent than factoid questions in the museum domain. Additionally, they have also discovered that different artworks did not have effect on the distribution of content types of the questions.

Any information providing chatbot can be considered as a question answering system because the fundamental components of a chatbot are processing the question and providing an answer. Question answering is a branch of information retrieval [3] where questions are automatically answered in natural language. It has three components, question processing, document processing and answer processing [1]. The most common way of processing the answer is by selecting the answer from a given context. This task is alternatively called reading comprehension [9]. A popular model for reading comprehension type question answering is BERT [6] which is built over some recent remarkable research works [13, 10, 18, 15]. It combines the concept of language modeling, transfer learning and bidirectional modeling and has performed even better than human level performance in some NLP tasks[3].

A literature study consisting of 1842 papers published up to 2017 suggests that any closed domain question answering is rare [1]. However, there are some BERT based implementations focusing on factoid [19] and open-ended questions [11, 12, 14] separately. For example, in open domain tasks which consist mostly of open-ended questions, a BERT implementation had the best performance [8]. Still, it would be quite rare to find any research to deal with both in the context of a realistic domain like the museum. This previous research motivates us to address the following research questions:

RQ 1: Is BERT able to tackle both factoid and open-ended questions or do we need specific modules?

RQ 2: What are the special characteristics when dealing with a specific test domain (museum) with a very small dataset?

RQ 3: Does the structure and content of the available data (in our case, the picture and its description) have any effect on the performance of the system?

2 BERT Notable properties

BERT introduced us actively to a new era of transfer learning [17], by transferring knowledge that has been already been learned from other tasks. To be more specific, its inductive transfer not only improves learning in standard supervised tasks but also helps overcome the problem of small datasets. BERT is

[3] https://rajpurkar.github.io/SQuAD-explorer/

trained in two stages. At the first pre-training stage, the model is trained using semi supervised learning [4] using a huge online dataset. The model is trained on a certain task that enables it to grasp patterns in language. Next, there is supervised training on a specific task with a labeled dataset.

The capability of BERT lies in the pre-training tasks. It was trained on these two pre-training tasks:

1. Masked language model: Some parts from the input sentences were masked and the task was to predict the missed word. This task helped the model to learn the word from the context.
2. Next sentence prediction: In this tasks two sentences were fed in the system and the training objective was to predict if the two sentences were consecutive. This task also helps to understand the context better.

The training enables the model to identify the role of a word in a sentence and to learn connections between sentences. Both are important for question answering, since both the question and the context are passed to the model. The answer is a span from the context. In question answering, BERT uses a start and end token classifier to extract the answer from the context. From the pre-training task, the model learns the language in general which helps to extract answer from the question. The clue from the question can be identified in the context. In our experiment we also going to explore how this relationship between the question and the context works.

Lastly it is important to mention that BERT is a huge model. If we have a specific NLP task for which it has been trained, we are going to get high quality results. But if it does not fit exactly with the training paradigm, it is quite unlikely to expect the same outcome. Surely we can train a new model from scratch with our own dataset but it will require the dataset and computational resources to be huge [5]. The other additional problem we have to deal with is that BERT limits the input to 512 tokens, so it becomes hard for longer contexts. We describe how we deal with this problem in Section 3.2.

3 Experimental methods

3.1 Task and dataset description

We carry out an online evaluation of a museum domain chatbot, where users are able to freely ask questions related to pictures they are seeing. We collaborate with the German Städel[4] museum and are able to use their carefully curated textual description of the meaning and significance of each picture as information base. The texts cover basic, religion, culture and artist information about each picture, and are used as contexts from which to retrieve answers.

In our online experiment, we presented participants with 5 pictures in turn and prompted them to ask 3 questions for each by freely typing into a text box. 19 participants fully completed this experiment, so the total number of questions

[4] https://www.staedelmuseum.de/en

collected was 285 (19*5*3). Since we used a pretrained English BERT model, we automatically translated the questions from German into English first. There were some translation errors in the automatic translations. We assume that the model is robust to minor linguistic errors in the questions, but we excluded 9 translations that contained major translation errors. This leaves a total of 276 questions submitted to BERT.

At this stage of data collection, the first author manually classified the questions into 2 classes, factoid and open-ended. In addition, we manually checked whether the submitted questions are answerable from the context or unanswerable. Answerable questions are those questions whose answers are found in the context (provided by the museum description). Finally, we also manually mark the correct answer span for each answerable question from the context.

All questions were then processed to generate answers from the BERT model. If the generated answer is meaningful and matches exactly or partially with our manually annotated span, we consider it as "correct", otherwise "incorrect".

3.2 Model description

We used the default 24 layer BERT model pre-trained with the SQuad question answering dataset [16]. The default model is pre-trained with English texts while we need to process both inputs and outputs in German. We used the Google translation API[5] to translate between German and English.

BERT needs both question and context as an input. When a picture was chosen, the corresponding context is selected. The maximum number of tokens allowed for any BERT model is 512. Most of the contexts were more than 512 tokens, in which case we divided the context into sub parts and tested the question with each context part separately. Both question and context are then passed to the BERT model. The model returns start and end token positions of the answer in softmax distributions. We then extract the best span from the contexts. When the start and end token does not create a meaningful span (e.g., both point to position 0 or the same index), we consider the output as "none". In the case of non-answerable questions, only a "none" output will be considered correct. On the contrary, when we retrieve a "none" output for answerable questions, it is considered incorrect.

4 Evaluation and analysis

4.1 Overall data distribution

Table 1 shows the overall performance of the experiment classified into factoid and open-ended questions. Out of 276 questions asked, 174 were factoid and 102 were open-ended questions. The overall performance of the system in the factoid questions was significantly better than open-ended questions. Around 70% of factoid questions were answered correctly, compared to only 36% of open-ended questions. This leads to an overall performance in the full experiment of 58%.

[5] https://pypi.org/project/googletrans/

Table 1: Overall statistics of the full data classified in two classes, factoid and open-ended.

	Total questions	Correct	Wrong	% Correct
Factoid	174	124	50	71.26
Open-ended	102	37	65	36.27
Total	276	161	115	58.33

Table 2: Overall accuracy among the questions which are answerable from the provided text, classified into 2 classes, factoid and open-ended.

	Total questions	Correct	Wrong	% Correct
Factoid	138	107	31	77.53
Open-ended	72	26	46	36.11
Total	210	133	77	63.33

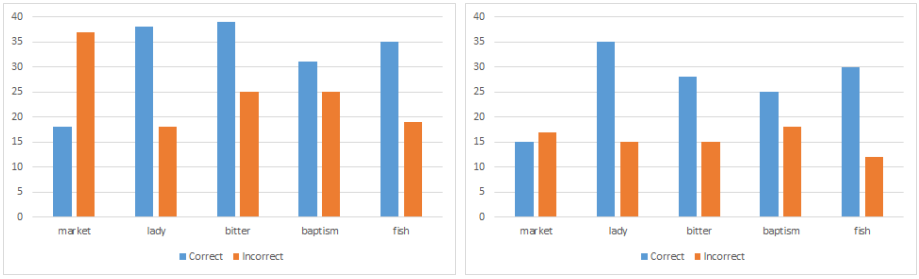
We can subdivide the overall data into answerable and non-answerable questions. Tables 2 and 3 show the performance of answerable and non-answerable questions. Out of all 276 questions, around 3 quarters were answerable. Among these questions, around two third were factoid questions. On the other hand in non-answerable questions, the number of factoid and open-ended questions was quite similar. Out of 66 non-answerable questions, 36 were factoid and 30 open-ended.

The accuracy in the case of answerable questions is around 5 points higher than in the full data, while in case of non-answerable questions, it fell from 58% to 42%. The accuracy for factoid questions in the answerable class also increased from around 71% to 78%. On the other hand, the accuracy in open-ended questions remains similar. In the case of non-answerable questions, the accuracy for factoid questions fell significantly compared to the overall accuracy while the number slightly increased for the open-ended questions.

Next we investigate the performance of the experiment in relation to the pictures (Figure 1). We are providing two graphs, one considering all questions (Figure 1a) and the other considering only the answerable ones (Figure 1b). We see that in both figures four of the pictures had more correctly answered questions than incorrect questions. In both figures, the number of incorrect answers is higher for the “market” picture. But it is also evident that the difference between the frequency of correct and incorrect responses was reduced for answerable questions. When we studied the details of the un-answerable, incorrectly answered questions for “market” picture, we found that these questions were mainly related to visual aspects of the picture and some random facts which were not present in the context. If we consider the artwork itself, the picture was about a market scene with at least 10–15 people in it which was quite different from other pictures in its visual complexity. On the other hand, the provided context from the museum was divided into a general description and religious background. Apparently, the type of picture or the content of the picture may

Table 3: Overall accuracy among the questions that are un-answerable from the provided text, classified into 2 classes, factoid and open-ended.

	Total questions	Correct	Wrong	% Correct
Factoid	36	17	19	47.22
Open-ended	30	11	19	36.67
Total	66	28	38	42.42



(a) Overall (b) Answerable

Fig. 1: Correct and incorrect counts among the questions(overall or answerable) distributed among the pictures

have an influence on the questions that users tend to ask. Additionally, in designing the knowledge base one also needs to take care to align the contexts with the possible questions which might be asked.

4.2 Answerable questions: detailed analysis

In this section, we are going to uncover the subtypes of questions which play a role in the performance of the model. This will enable future improvements of question answering solutions on real-world document sets. We are considering only answerable questions because the answer can be identified by a human from the given context. This analysis also uncovers the most common questions asked in factoid and open-ended domains. We will further analyze the questions with connection to the context in the next section which will also help us identify any useful patterns connecting the question and the context.

Table 4 shows the questions which are answered correctly in answerable questions. Out of a total of 133 correct questions, most of the questions are factoid. The question type "Who painted the picture / when was the picture painted" is the most common question overall. We consider these two questions under one category because the answer to these two questions resides in same sentence in the context, typically the first. The next most asked factoid question is "Who is the person" (inquiring about a person depicted in the picture). Fact-1sent are questions that can be answered from a single source context sentence. In fact, all the categories fact-a, fact-b, fact-museum and fact-title belong to this

Table 4: Answerable questions answered correctly.

	Question	count	category
Factoid	Who painted / when painted	75	fact-a
	Who is the woman or person	17	fact-b
	Which museum is the picture located in	3	fact-museum
	What is the title of the picture	6	fact-title
	ex. Who is baptized	5	fact-1sent
	Are there any fish	1	fact-tracearound
	Total	107	
Open-ended	What is happening	9	open-happening
	Where is Christ in the picture	2	open-whowhere
	Why are these people unhappy	7	open-cluematch
	Why i watch the man so funny	6	open-cluepartial
	What is the significance of this painting?	2	open-meaning
	Total	26	

question category. The other category fact-tracearound captures questions where the answer does not occur in a single sentence. Rather there are traces pointing towards the answer in either the previous or following sentences or both.

The other section in Table 4 categorizes correct, answerable, open-ended questions, whose total number is smaller than the number of factoid questions asked. Among the correctly answered open-ended questions, the most common question was "what is happening in the picture". We are calling this question type "open-happening". The category names open-happening, open-whowhere and open-meaning came directly from the question itself. The other 2 categories came from the question's connection with the context. When we have a clue about the topic of the question in the context, we call the question cluematch. As these questions belong to the open-ended category, we call them open-cluematch. When the clue is only partial, it is called cluepartial. In examples 3 and 4 of section 4.3, we give examples of direct clue and partial clue cases.

Table 5 shows the questions answered incorrectly among the answerable questions. Out of a total of 77 wrongly answered answerable questions, most are open-ended. Among the factoid questions, the question which is answered incorrectly most often is "Who is the person or lady". This category of question was also common in correctly answered questions. The next category where the count is 8 belongs to fact-a. This category question was the highest in the factoid correct class. Fact-partans means the question was answered partially. Fact-none questions are those where we get a "none" output although we do have an answer in the context. The other category fact-tracemissed is related to the context: when we have a trace in the context but BERT missed the trace and provided incorrect output.

The second part in Table 5 belongs to the incorrectly answered open-ended questions. The category names open-happening, open-whowhere, open-meaning, and open-summary come from the question itself. The other categories are named after their purposes. Open-tracemissed is the category of questions which is

Table 5: Answerable questions answered wrong.

	Question	count	category
Factoid	Who painted/ when painted	8	fact-a
	Who is the lady in the background	15	fact-b
	ex. it comes from Jordan	2	fact-partans
	ex. what's in the bottle	3	fact-none
	who has commissioned the image	3	fact-tracemissed
	Total	31	
Open-ended	What can be seen on the picture	21	Open-happening
	Who are the two people pictured in the foreground	2	Open-whowhere
	Ex. What fish symbolizes	7	Open-meaning
	Ex. What is the history of the image	8	Open-summary
	Why shouts the man	1	Open-cluepartial
	Ex. what does that mean medal	4	Open-partans
	Why are the proportions that weired	3	Open-tracemissed
	Total	46	

similar to factoid-tracemissed, where the model missed the clue given in the question. Open-partans relates to answers which were unacceptable due to incompleteness. The last category Open-cluepartial means those questions where the clue from the question partially matches the context. Like the correctly answered open-ended questions, the highest number of wrong answers are also in the open-happening category.

4.3 Question categories discovered in connection with the context

In this sections we will discuss the question categories separately for factoid and open-ended questions. First we group the categories into 3 sections according to how often questions from this category were answered incorrectly (Table 6). The value in the bracket denotes the count of correct or wrong outcomes. For the categories in the middle column we mention both the counts of correct and wrong responses. Then we derive the relationships of these categories with the context.

Factoid question categories The categories fact-a and fact-b are frequently answered both correctly and incorrectly. Out of a total of 174 factoid questions, 113 belong to these two categories. Fact-a is mostly answered correctly. The question categories which are always correct are fact-museum, fact-title, fact-1sent and fact-tracearound. The questions where we got a partial answer, got "none" output and where we missed the traces are incorrect in our experiment. It is quite obvious that when we get "none" as an output it must be considered wrong in the answerable question class. It is also notable that only 3 out of 174 factoid questions produced a "none" output.

The factoid question categories described are sometimes directly related to the question, but sometimes related to the purpose. For example, fact-museum is

Table 6: Question categories grouped into 3 sections, common, only correct and only wrong. The number in bracket denotes the number of questions within the category.

	Only correct	Correct and wrong	Only wrong
Factoid	fact-museum(3) fact-title(6) fact-1sent(5) fact-tracearound(1)	fact-a(75c 8w) fact-b(17c 15w)	fact-partans(2) fact-none(3) fact-tracemissed(3)
Open-ended	Open-cluematch(7)	Open-happening(9c 21w) Open-whowhere(2c 2w) Open-cluepartial(6c 1w) Open-meaning(2c 7w)	Open-summary(8) Open-partans(9) Open-tracemissed(3)

directly a museum question. But fact-none can be any question where the output is none. All the factoid categories can be further divided into 2 classes based on how many sentences from the context are needed to answer them: fact-1sent and fact-multi. Fact-1sent means questions where the answer comes in a single sentence from the context. Fact-multi is where the answer combines multiple sentences. Fact-a, fact-b, fact-museum, fact-title and fact-1sent always belong to fact-1sent. Fact-tracearound is a typical example of the fact-multi class. These types of questions have traces from the context in multiple sentences. Fact-partans, fact-none and fact-tracemissed categories can be either fact-multi and fact-1sent questions. This categorization will help us understand and generalize the performance of the questions in relation with the context.

To have a better understanding of these question categories, let us explain the two categories with two examples. In example 1, the answer of the question comes directly from a single sentence. On the other hand in example 2, the answer to the question "Is it a historical person?" requires multiple sentences to have a meaningful answer. The answer could be directly given either yes/no but since we are picking a span from the context, it needs more than a single sentence.

Example 1 (Fact 1 sentence). A single sentence is good enough for the answer.
Question : Which artist has painted the picture?
Context: [...] **Adriaen Brouwer** painted this picture around 1636/38. [...]

Example 2 (Fact multi sentence). Multiple sentences are needed for answering a factoid question
Question : Is it a historical person?
Context: [...] The portrait depicts Simonetta Vespucci, a beauty praised throughout Florence – not in the sense of a portrait, but as an idealized figurine of an ancient nymph. [...]

Table 7 reflects how these broad two categories affect the performance on the factoid questions. Though we have few instances of fact-multi questions, we can see that it is unlikely to get a good output in these cases. On the contrary,

Table 7: Factoid question categories summarized into 2 groups, fact-1sent and fact-multi.

	Fact-1sent	Fact-multi
Correct	106	1
Wrong	27	4

where we have traces in a single sentence in factoid questions, we get a higher number of correct answers.

Open-ended question categories In case of open-ended questions, there are fewer instances of correct answers compared with the wrong answers. Four categories, open-happening, open-whowhere, open-cluepartial and open-meaning, have both correct and incorrect intances. Open-cluematch is the only category which has all the questions asked (7) correct. Open-whowhere has equal numbers in correct and incorrect. But the other two categories, open-happening and open-meaning, have more incorrect than correct questions in the experiment. In our experiment we have 6 instances of open-cluepartial cases where we got correct answers. The other three categories open-summary, open-partans and open-tracemissed have only incorrect outputs.

All the open-ended questions require multiple sentences from the context to be answered correctly, so the broad factoid classification will not work for them. However, we can differentiate them based on the clue given in the question. For example, if the question asks for something specific, like "Why is the sky blue?" we can mark these as DirectClue, since a direct clue ("sky") is given which may match some part of the context explicitly. On the other hand, when the question asks "What is the history of the picture?", its answer can cover a large part of the context. But it is unlikely that a mention of the word "history" will be there in the context. We call these cases IndirectClue. In examples 3 and 4, we have examples of these questions.

Example 3 (Direct Clue). "Painting a fish" is considered a direct clue. We are here looking for specific information.
Question: why did the artist paint fish?
Outcome: on behalf of a guild who wanted to decorate their rooms with the image.

Example 4 (Indirect clue). The question does not directly relate to the sentence in the context.
Question: Does it have a deeper meaning?
Outcome: [...] the picture should perhaps keep her memory . through the representation as an ancient nymph she was simultaneously raptured and glorified .

The categories open-whowhere, open-cluepartial and open-cluematch belong to the DirectClue group, because we typically have a specific clue in the question.

Table 8: Open-ended question categories summarized into 2 groups, DirectClue and IndirectClue.

	DirectClue	IndirectClue
Correct	15	11
Wrong	8	38

On the other hand, open-meaning, open-happening and open-summary do not seek any specific information. Open-partans and open-tracemissed will appear in both broad categories.

In Table 8, we depict the performance for open-ended questions divided into the two classes, DirectClue and IndirectClue. Unsurprisingly, we receive better output in the case of DirectClue questions, while we got correct answers in just one fifth of the IndirectClue cases.

5 Experiments for improving outputs

We can use the detailed analysis of error categories from the previous sections to improve the performance of the outcome for factoid and open-ended questions (see Table 6). Among the factoid questions, fact-partans, fact-none and fact-tracemissed questions are always incorrect. The same is the case for open-ended questions open-summary, open-partans and open-tracemissed. Partial answers and missing traces from the question are common in both groups. In this Section we report on our experiments to solve partial answer and fact-none problems. We are leaving open-summary for future work because the question has little information to elaborate. The other category which we are leaving is tracemissed. This type of question, like cluematch or cluepartial questions, is hard to answer because the helpful information for answering these questions is implicit.

In the category of partial answers, the model actually found the point where the answer resides. But due to incompleteness, the answer becomes unacceptable. In example 5, we can see the answer was initially very short. But when we added the sentence consisting of the word and the next sentence, the answer becomes acceptable.

Example 5 (Adding context). Adding more context with the partial answer can result in better output
Question: What kind of fish?
Category: Open-partans
Context: Still life with fish on a kitchen bench. A few eels meander on the left side of the sales bench, in the middle a shimmering carp hangs on a thread, on the right the rich flesh of a sliced salmon lights up. The Antwerp painter Jacob Foppens van Es presents a selection of different fish species on this virtuoso still life. [...] *Outcome*: salmon
Operation: including the current sentence and next sentence in output
Outcome: A few eels meander on the left side of the sales bench, in the middle a

shimmering carp hangs on a thread, on the right the rich flesh of a sliced salmon lights up. The Antwerp painter Jacob Foppens van Es presents a selection of different fish species on this virtuoso still life.

Another category of incorrectly answered questions is fact-non. This is because the system could not find any hints towards the answer in the context. In example 6, we see that there is no reference of the word "bottle" in the context. Absence of a clear trace resulted in a "none" output. The sentence close to the answer is elliptical in that the bottle is not mentioned. If we adapt the provided context to make this explicit, the answer is correctly retrieved. This example shows an important characteristic of the BERT span picking method. We can see that there exists a cause-effect relationship in the model. If we add additional information in a coherent manner or break the internal rigidity, It can be achievable to enrich the context without affecting earlier performance.

Example 6 (Adding traces can improve output). Having an explicit trace is crucial for retrieving an output. If we add a trace in the clue sentence we get an output.
Question: What's in the bottle?
Category: fact-none
Context: Adriaen Brouwer, The Bitter Potion (1076). The bitter potion that the ragged young man has just consumed makes his facial features derailed. You can almost taste it. [...] *Outcome*: none
Operation - Adding a trace in the next sentence: Adriaen Brouwer, The Bitter Potion (1076). The bitter potion that the ragged young man has just consumed makes his facial features derailed. You can almost taste it **from the bottle**.
Outcome: the bitter potion

6 Conclusions and future directions

We have studied the effectiveness of BERT for both factoid and open-ended questions, applied to one real-world domain. We have focused mainly on those questions which are answerable. We have found that we get much better results for factoid questions. But in our experiment we have also found that BERT was able to answer open-ended questions in around 36% of cases. Based on these results we can answer our first research question, "Is BERT able to tackle both factoid and open-ended questions", negatively.

For the second research question, "What are the special characteristics when dealing with a specific domain and small dataset", we carried out a detailed error analysis to identify which types of questions pose specific problems. First we categorised different subtypes of questions. Then we further grouped the categories according to their difficulty. Among the factoid questions, we have found two groups: Questions which can be answered from one sentence in the context (fact-1sent), and questions which need multiple sentences from the context

(fact-multi). In our experiment, questions which can be answered from just one sentence are more often answered correctly. Answers which involve multiple sentences from the context are quite difficult in this experiment. In the light of this, we can predict that the performance on open-ended questions will be comparatively lower, because they need multiple sentences from the context due to the nature of the question. In our observation from the experiment, this expectation was confirmed. Overall less than 40% of open-ended answers were acceptable.

With respect to the connection to the context, open-ended questions can be divided into two broad categories: questions which explicit clues which are mentioned in the context (DirectClue) and those which ask for more broader answers without any lexical hints (IndirectClue). In our experiment, open-ended questions with direct clues had more acceptable outcomes. We also had positive outcomes in 11 cases out of 72 broader indirect open-ended questions. Among these questions, 9 were from open-happening and 2 from Open-meaning.

For the last research question, "Does the structure and content of the available data have any effect on the performance of the system", we compared the performance of the questions across different pictures. In our experiment we have found the performance of just one picture was different from the other pictures. For this picture, most of the questions were related to visual facts (e.g., "how many people are there?") or facts which were not present in the provided context. When we analyzed it more carefully, we found that the picture was different (market scene consisting of several people) and most of the questions asked were un-answerable from the context. This leads us to conclude that the specific data provided can have a large effect on the performance of a question answering system applied to a real world domain.

From the results of the experiment we can also draw connections with the training objective of BERT. The masked language model has a greater influence on identifying answers in a single sentence, whereas the next sentence prediction can be related to identifying the context. It creates a kind of cause and effect relation which we also mentioned in our analysis as clue. So if the clue from the question is matched in the context, it is more likely to give an acceptable answer. But when the clue is much broader, like "the history" or "hidden meaning", we are less likely to get a good answer, because the model is optimized to point to a specific clue reference. When we need multiple clues, the model can not retrieve the answer. In our observation, when we need more than two sentences from the context to answer a question, we expect to get unacceptable output.

Finally, finding a partial answer is a problem for both factoid and open-ended questions. For example, when we ask "Who painted", the model retrieves the painter because in the context "x was painted by y", the part "painted by" plays an important role to determine the outcome. This phenomenon is obviously great in factoid questions, but for open-ended questions the scenario is often not so straight forward. In our experiment we showed that if we add to the context to make it more explicit, we can get an acceptable outcome. In open-ended questions, questions which are very common but yield incorrect answers are those which require broader answers consisting of a span of multiple

sentences. We name this as indirect Clue. In the future it can be further analyzed whether we gain performance benefits if instead of providing a single indirect clue question, we generate multiple questions in relation to the main question and combine the generated outcome.

References

1. Antonio, M., Soares, C., Parreiras, F.: A literature review on question answering techniques, paradigms and systems. Journal of King Saud University-Computer and Information Sciences pp. 806–809 (2018)
2. Barth, F., Candello, H., Cavalin, P., Pinhanez, C.: Intentions, meanings, and whys: Designing content for voice-based conversational museum guides. In: Proceedings of the 2nd Conference on Conversational User Interfaces. pp. 1–8 (2020)
3. Cao, Y.g., Cimino, J.J., Ely, J., Yu, H.: Automatically extracting information needs from complex clinical questions. Journal of Biomedical Informatics **6**(43), 962–971 (2010)
4. Dai, A.M., Le, Q.V.: Semi-supervised sequence learning. In: Advances in neural information processing systems. pp. 3079–3087 (2015)
5. Dettmers, T.: Tpus vs gpus for transformers (bert) — tim dettmers. https://timdettmers.com/2018/10/17/tpus-vs-gpus-for-transformers-bert/ (10 2018), (Accessed on 09/20/2020)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.N.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. pp. 4171–4186 (2019)
7. Gribanova, A.: Chatbots as a tool to enhance visitor experience in museums. a case study in the panorama museum. In: ISCONTOUR 2020 Tourism Research Perspectives: Proceedings of the International Student Conference in Tourism Research. p. 62. BoD–Books on Demand (2020)
8. Hashemi, H., Aliannejadi, M., Zamani, H., Croft, W.B.: Antique: A non-factoid question answering benchmark. In: European Conference on Information Retrieval. pp. 166–173. Springer (2020)
9. Hirschman, L., Light, M., Breck, E., Burger, J.D.: Deep read: A reading comprehension system. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics. pp. 325–332 (1999)
10. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 328–339 (2018)
11. Liu, X., He, P., Chen, W., Gao, J.: Multi-task deep neural networks for natural language understanding. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4487–4496 (2019)
12. Nogueira, R., Cho, K., Scholar, C.A.G.: Passage re-ranking with bert. arXiv preprint arXiv:1901.04085 (2019)
13. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
14. Qiao, Y., Xiong, C., Liu, Z., Liu, Z.: Understanding the behaviors of bert in ranking. arXiv preprint arXiv:1904.07531 (2019)
15. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)

16. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
17. Torrey, L., Shavlik, J.: Transfer learning. In: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, pp. 242–264. IGI global (2010)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
19. Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., Xiong, K., Li, M., Lin, J.: End-to-end open-domain question answering with bertserini. In: NAACL-HLT (Demonstrations) (2019)