COMPARING BERT WITH AN INTENT BASED QUESTION ANSWERING SETUP FOR OPEN-ENDED QUESTIONS IN THE MUSEUM DOMAIN

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

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

Abstract: BERT-based models achieve state-of-the-art performance for factoid question answering tasks. In this work, we investigate whether a pre-trained BERT model can also perform well for open-ended questions. We set up an online experiment, from which we collected 111 user-generated open-ended questions. These questions were passed to a pre-trained BERT QA model and a dedicated intent recognition based module. We have found that the simple intent based module was around 25% more often correct than the pre-trained BERT model, indicating that open-ended questions still require different solutions compared to factoid questions.

1 Introduction

Museums play an important role in enabling people of all ages to study history, culture and contemporary society. Many museums, in turn, employ interactive systems such as chatbots to provide this kind of information to visitors. A detailed study of more than 5 thousand unique sessions in the Pinacoteca Museum in Brazil was conducted to discover the type of questions people pose using chatbots in the museum [1]. They have identified 8 types of questions, including *fact, author, visual, style, context, meaning, play* and *outside*. Among these, *meaning* questions constitute around 60% of the questions. If we categorize the collected questions into two broad groups, factoid and open-ended, open-ended questions are more common than factoid questions in the museum domain.

A chatbot which answers a user question in natural language can alternatively be viewed as a question answering (QA) system. It has three components: question processing, document processing and answer processing [2]. As we proceed to the era of deep learning, we can find these three components tend to merge. BERT [3] is considered a current state-of-art QA model. It incorporates the principle of language modeling, transfer learning and bi-directional modeling and has performed even better than human level output in some NLP tasks[1]. BERT for QA has been studied mostly in relation to factoid questions and performs very well on them.

In this study, we analyze the performance of BERT-based QA models for open-ended questions in a realistic museum domain setting. We investigate the research question whether a pre-trained BERT model is also sufficient for open-ended questions, and how it compares to a focused intent-recognition based module. Our task is situated within a spoken dialog system in the museum domain: a chatbot that can answer free user questions about pictures displayed in an art museum.

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

2 Open-ended QA with BERT.

Question answering is a branch of information retrieval [4] where questions are automatically answered in natural language. Work in QA has focused on extracting answer spans from related passages for reading comprehension questions. Several corpora exist as training data (e.g., SQuAD[5], QuAC[6]), but it is known that they are dominated by factoid questions (e.g., about dates or numbers). In contrast, open-ended questions have traditionally been approached as a passage retrieval task [7, 8, 9, 10, 11, 12].

For this task setup, BERT was used for binary passage re-ranking, where the question and several passages are fed into the model to find the best passage [13, 14, 15, 16]. This approach can perform well, but a fine tuning process is needed to optimize for this task with an in-domain training data set. In addition, a suitable training set is often not available and expensive to produce (it includes chunking the individual answer passages).

In this work, we therefore test the performance of regular pre-trained BERT for answer span extraction QA also on open-ended questions.

3 Experiment

3.1 Task

We carried 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 museum[2] 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 picture by freely typing into a text box. 19 participants fully completed this experiment, so the total number of questions we collected was 285 (19*5*3).

3.2 Dataset Description

At this stage of data collection, the first author manually classified the questions into 2 classes, factoid and open-ended. Factoid questions are questions which can be answered by providing a specific piece of information (for example, "Who painted the picture?"), while open-ended questions usually require a longer contribution to be fully answered (e.g., "What is the historical significance of the picture?"). Among the collected user questions, 111 were open-ended questions. We have sent these to the two models, a pre-trained BERT QA span extraction model and a small intent based model trained on the museum data. The outputs from both models were manually checked by one of the authors for correctness. Additionally, we determined whether the target answers to the users' questions are available in the description text of the pictures (i.e., whether the questions are answerable or not).

We illustrate our approach by the example given below in Figure 1. First, we determine whether the question can be answered from the context. Then we pass the question to the models and manually determine the outcome correctness. In this case, the BERT outcome is correct, while the intent based model was incorrect. The identified category was too broad to answer a specific question in this case.[3]

[2] `https://www.staedelmuseum.de/en`

[3] Please note that the museum and all its data are presented in German, and the evaluation was carried out in German. In order to use the pre-trained BERT model, we pass the user questions as well as the picture context

Question: why did the artist a black background chosen?
Context: Sandro Botticelli (Sandro Filipepi), Female Ideal Portrait (Portrait of Simonetta Vespucci as a Nymph), c. 1475. She is a larger than life beauty. Effectively placed against a dark background you can see her face almost in profile, while the upper body is slightly facing the viewer. Sandro Botticelli created a masterpiece of Italian Renaissance painting here. [...]
Data Available in context: Yes
BERT model outcome: you can see her face almost in profile , while the upper body is slightly facing the viewer
BERT outcome correct? Yes
Intent identified? What is the story behind the picture?
Intent based model outcome: In January 1849, the inspector of the Städel, Johann David Passavant, wrote an expert opinion in connection with the purchase of the picture for the museum. It reads: "All art lovers who know the works of Sandro Botticelli and have seen the picture in question agree with it beautiful original work by this master to be recognized. I also think I can predict that all art lovers will immerse themselves in the picture with real pleasure. Simonetta had died young. Maybe the picture should keep her memory. Through the portrayal as an ancient nymph, she was simultaneously raptured and glorified.
Intent outcome correct? No

Figure 1 – A sample record including an excerpt of the picture description context, a sample question and responses from the two models.

4 Models in the Experiment

4.1 BERT Model

We used the default 24 layer BERT model pre-trained with the SQuAD question answering dataset [5]. 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[4] to translate between German and English.

BERT needs both question and context as an input. When a picture was chosen, the corresponding description context is selected. The maximum number of tokens allowed for any BERT model is 512. Most of the contexts were longer 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.

4.2 Intent based model

The intent based model is a combination of the RASA [17] intent identification and a backup similarity module. RASA consists of a pair of tools, Rasa NLU and Rasa Core, which are open

to the Google Translate API for automatic translation into English. While this introduces some minor linguistic errors (note the ungrammatical question in Figure 1), we assume that the QA model is generally robust wrt. such errors.

[4] `https://pypi.org/project/googletrans/`

source python libraries to create conversational applications. The RASA intent identification module which is part of the RASA NLU tool classifies the intent and a corresponding confidence value from the input question. If the confidence surpasses a manually chosen threshold of 20%, we return the related section from the database. The similarity module, on the other hand, acts as a backup module for two cases: either if the identified intent confidence is too low, or if the associated intent data is not found in the database for the given picture (e.g., the picture description is missing discussion of religious significance of the artwork). The model architecture is depicted in Figure 2. In the remainder of this section, we describe some of the important aspects of each part of the model.

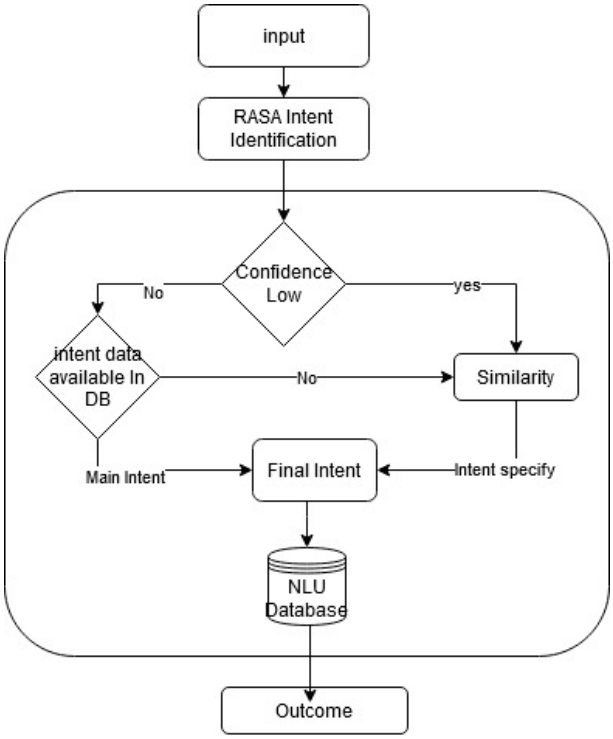

Figure 2 – A model in combination with RASA intent identification and similarity module.

4.2.1 RASA-based Intent Identification

In order to use the RASA intent identification model we manually segmented the museum descriptions of the pictures into paragraphs and assigned each paragraph to one of 14 common topic (= intent) categories chosen in collaboration with the museum experts, such as “history”, “religious significance”, etc. The RASA intent identification model is trained to classify freely entered user questions into these 14 intent categories. We trained the RASA model with 20 user questions for each intent category. This model returns the best matching intent along with the confidence value.

4.2.2 NLU Database

The database is populated with the texts collected from the Staedel museum experts. For this project, we chose 30 pictures for the experiment. The text corpus consists of 769 sentences. On average, the text describing a picture could be annotated with 3 to 5 intent categories.

4.2.3 Similarity

The similarity module is used as a backup for the intent based module and works in a two step process. At first, the similarity module matches the question with each of the sentences describing the selected image. The similarity between the user question and a sentence in the script is determined by the cosine similarity [18] measure. We collect the intent of the best matched sentence from the NLU database. Finally, all sentences that belong to this intent are returned.

5 Results and Discussion

We did the analysis in two parts. Table 1 and 2 summarizes the performance comparison between the BERT based and the intent based model. We present the findings from two annotators and present an average measure of accuracy. First, we looked at all the open-ended questions and assessed the outcome of both models. en we viewed at all 111 open-ended questions, the intent-based model was more frequently correct, 51.35% compared to 37% (average accuracy score of annotators) for BERT (Tab.1).

Table 1 – Overall data performance of the two models, Intent based and BERT

Model	**Total**	**Correct**		**% Acc.**		
		Ann. 1	**Ann. 2**	**Ann. 1**	**Ann. 2**	**Avg.**
Intent based	111	57	57	51.35	51.35	51.35
BERT	111	37	46	33.33	41.44	37.22

Table 2 – Answerable data performance of the two models, Intent based and BERT

Model	**Total**		**Correct**		**% Acc.**		
	Ann. 1	**Ann. 2**	**Ann. 1**	**Ann. 2**	**Ann. 1**	**Ann. 2**	**Avg**
Intent based	81	85	48	52	59.26	61.17	60.21
BERT	81	85	26	32	32.1	38.82	35.4

Next, we considered only the answerable questions in the analysis in table 2. Out of 111 open-ended questions, 81 from annotator 1 and 85 from annotator 2 questions were answerable from the description text of the picture. BERT was able to retrieve the correct output span in 26 cases (32.1%) and 32 cases (38.82%) (annotator wise). In contrast, the intent based model was correct around 50 cases (60.21%) (on average). If we compare the accuracy of the BERT model among the overall and answerable data, the value of the accuracy remains similar. While the average accuracy of the intent based model increase up to 9%.

The performance difference between the pre-trained BERT model and the intent based model is remarkable for two reasons. First, the training dataset for intent recognition was quite small. The performance of the intent recognition can be further improved, which will eventually widen the gap between the performance of these two models. Second, on average 3 to 5 out of the total of 14 intent categories could be found in the NLU database. So in most of the cases we had to depend on the backup similarity based answer retrieval. However, a single paragraph can serve multiple intents which can further improve the accuracy of the intent based system. On the other hand, the BERT model could only reach up to a third time for 81 questions where

we could manually identify the answer in the context. In short, if we consider the deficiencies, the outcome of the intent-based module can be considered to be more promising than the pre-trained BERT model.

6 Conclusion

In this study, we analyzed the performance of a BERT-based QA model for open-ended questions in a realistic museum domain setting. We investigated the research question whether a pre-trained BERT model is also sufficient for open-ended questions, and how it compares to a focused intent-recognition based module.

Our results show that the intent based model is about 30 percent more accurate than the BERT model on answerable questions. Even though we trained the intent based open-ended module with a limited training data set, it was relatively better than the pre-trained BERT model, and shows promise for further improvements.

At present, the experiment was carried out in a small scale, using only our own judgments for evaluation (for example of the correctness of system responses). In the future, we plan to include domain experts from the museum in this process in order to be more reliable. Furthermore we would also like to extend our work to compare it with other types of open-ended questions (e.g. Ubuntu chat task).

We conclude that the high performance of pre-trained BERT-based QA systems does not extend to open-ended questions, which therefore warrant new research into suitable approaches.

Acknowledgements

This research is part of the ChiM project of the research initiative "KMU-innovativ: Mensch-Technik-Interaktion", which is funded by the Federal Ministry of Education and Research (BMBF) of the Federal Republic of Germany under funding number 16SV8331.

References

[1] BARTH, F., H. CANDELLO, P. CAVALIN, and C. PINHANEZ: *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.

[2] ANTONIO, M., C. SOARES, and F. PARREIRAS: *A literature review on question answering techniques, paradigms and systems. Journal of King Saud University-Computer and Information Sciences*, pp. 806–809, 2018.

[3] DEVLIN, J., M.-W. CHANG, K. LEE, and K. TOUTANOVA: *Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805*, 2018.

[4] CAO, Y.-G., J. J. CIMINO, J. ELY, and H. YU: *Automatically extracting information needs from complex clinical questions. Journal of Biomedical Informatics*, 6(43), pp. 962–971, 2010.

[5] RAJPURKAR, P., J. ZHANG, K. LOPYREV, and P. LIANG: *Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250*, 2016.

[6] CHOI, E., H. HE, M. IYYER, M. YATSKAR, W. T. YIH, Y. CHOI, P. LIANG, and L. ZETTLEMOYER: *Quac: Question answering in context*. In *2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 2174–2184. Association for Computational Linguistics, 2020.

[7] COHEN, D., L. YANG, and W. B. CROFT: *Wikipassageqa: A benchmark collection for research on non-factoid answer passage retrieval*. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1165–1168. 2018.

[8] HASHEMI, H., M. ALIANNEJADI, H. ZAMANI, and W. B. CROFT: *Antique: A non-factoid question answering benchmark*. In *European Conference on Information Retrieval*, pp. 166–173. Springer, 2020.

[9] COHEN, D. and W. B. CROFT: *End to end long short term memory networks for non-factoid question answering*. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, pp. 143–146. 2016.

[10] COHEN, D. and W. B. CROFT: *A hybrid embedding approach to noisy answer passage retrieval*. In *European Conference on Information Retrieval*, pp. 127–140. Springer, 2018.

[11] KEIKHA, M., J. H. PARK, and W. B. CROFT: *Evaluating answer passages using summarization measures*. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pp. 963–966. 2014.

[12] YULIANTI, E., R.-C. CHEN, F. SCHOLER, W. B. CROFT, and M. SANDERSON: *Document summarization for answering non-factoid queries*. *IEEE transactions on knowledge and data engineering*, 30(1), pp. 15–28, 2017.

[13] NOGUEIRA, R. and K. CHO: *Passage re-ranking with bert*. *arXiv preprint arXiv:1901.04085*, 2019.

[14] LIU, X., P. HE, W. CHEN, and J. GAO: *Multi-task deep neural networks for natural language understanding*. *arXiv preprint arXiv:1901.11504*, 2019.

[15] MASS, Y., H. ROITMAN, S. ERERA, O. RIVLIN, B. WEINER, and D. KONOPNICKI: *A study of bert for non-factoid question-answering under passage length constraints*. *arXiv preprint arXiv:1908.06780*, 2019.

[16] QIAO, Y., C. XIONG, Z. LIU, and Z. LIU: *Understanding the behaviors of bert in ranking*. *arXiv preprint arXiv:1904.07531*, 2019.

[17] BOCKLISCH, T., J. FAULKNER, N. PAWLOWSKI, and A. NICHOL: *Rasa: Open source language understanding and dialogue management*. *arXiv preprint arXiv:1712.05181*, 2017.

[18] HUANG, A.: *Similarity measures for text document clustering*. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, vol. 4, pp. 9–56. 2008.