An Answer Verbalization Dataset for Conversational Question Answerings over Knowledge Graphs

Endri Kacupaj
kacupaj@cs.uni-bonn.de
University of Bonn, Germany

Kuldeep Singh
kuldeep.singh1@cerence.com
Zerotha Research and Cerence GmbH, Germany

Maria Maleshkova
maria.maleshkova@uni-siegen.de
University of Siegen, Germany

Jens Lehmann
jens.lehmann@cs.uni-bonn.de
University of Bonn, Germany

ABSTRACT
We introduce a new dataset for conversational question answering over Knowledge Graphs (KGs) with verbalized answers. Question answering over KGs is currently focused on answer generation for single-turn questions (KGQA) or multiple-turn conversational question answering (ConvQA). However, in a real-world scenario (e.g., voice assistants such as Siri, Alexa, and Google Assistant), users prefer verbalized answers. This paper contributes to the state-of-the-art by extending an existing ConvQA dataset with multiple paraphrased verbalized answers. We perform experiments with five sequence-to-sequence models on generating answer responses while maintaining grammatical correctness. We additionally perform an error analysis that details the rates of models’ mispredictions in specified categories. Our proposed dataset (Verbal-ConvQuestions) extended with answer verbalization is publicly available with detailed documentation on its usage for wider utility.

CCS CONCEPTS
• Information systems → Question answering.

KEYWORDS
conversations, verbalization, question answering, KG

1 INTRODUCTION
Question answering (QA) over Knowledge Graphs (KGs) is an essential task that maps a user’s utterance to a query over a KG in order to retrieve the correct answer [29]. Recently, with the increasing popularity of intelligent personal assistants, the research focus of the scientific community has shifted to conversational question answering over KGs (ConvQA) within multi-turn dialogues [5, 13, 24]. However, existing open-source KGQA systems are restricted to only generating or producing answers without verbalizing them in natural language [10]. The lack of verbalization makes the interaction with the user unnatural and often leaves the users with ambiguity [8]. Let us consider the first question in Figure 1, “What countries did the main character travel in the book Eat, Pray, Love?”, where existing QA systems will only respond with the countries as an answer, with no further explanation. In such cases, the user might need to refer to external data sources to verify the answer. Nevertheless, a verbalized answer would allow the user to confirm that the answer is related to the context, since it also includes additional characteristics that indicate how it was determined. For example, in task oriented dialogues, verbalized answers are a common phenomenon [23].

Recent efforts introduce answer verbalization in a QA dataset [3, 12]. The associated empirical results indicate the effectiveness of answer verbalization [12, 14]. Albeit effective, answer verbalization for ConvQA is an open research direction due to unavailability of dataset(s) (cf. Table 1). This is precisely one of the reasons why this critical information retrieval task is narrowly studied [16]. Hence, in this paper, we address the task of answer verbalization in conversational question answering over knowledge graphs. In particular, we extend an existing English ConvQA dataset [5] with multiple paraphrased natural language answers, using a semi-automated framework that employs back-translation as a paraphrasing technique. Moreover, we perform experiments with various models on generating the verbalized answer. We further provide

1https://github.com/endrikacupaj/Verbal-ConvQuestions
Table 1: Comparison of Verbal-ConvQuestions with existing conversational KGQA datasets. The lack of answer verbalization and paraphrased utterances remains a key gap in the literature, which is a key novelty of proposed dataset.

| Dataset                      | Train | Val. | Test |
|------------------------------|-------|------|------|
| # Conversations              | 6,720 | 2,240| 2,240|
| # Paraphrased Question       | 68,447| 22,368| 22,400|
| # Paraphrased Answer         | 68,447| 22,368| 22,400|
| Avg. Question length         | 8.48  | 8.75 | 8.01 |
| Avg. Answer length           | 8.82  | 9.19 | 8.39 |

Table 2: Verbal-ConvQuestions statistics.

2 DATA-GENERATION AND AUGMENTATION

We inherit questions from ConvQuestions [5], which is a high quality and large-scale benchmark for conversational QA over Wikidata KG [32]. The dataset contains 11,200 conversations, and it was compiled from inputs of crowd workers on Amazon Mechanical Turk with conversations from five domains: Music, Soccer, TV Series, Books, and Movies. The questions incorporate challenging phenomena such as aggregations, compositionality, temporal reasoning, and comparisons.

2.1 Initial Answer Verbalization

The first step is to generate the initial answer verbalization from the seed answers given in the original dataset. We group all similar questions or question templates and reword them using a rule-based approach. To maintain consistency across all answers, we cover the question and answer entities using a few general tokens (ENT, ANS). We substitute the tokens back to the original position after the first version is generated. Similarly to other works [8], we use box brackets to distinguish the seed answer from the remaining sentence; this is helpful when experimenting with the verbalized answers. For example, for the question “What countries did the main character travel in the book Eat, Pray, Love?” the step would provide an initial verbalized answer: “The main character travels in the book Eat, Pray, Love to [Italy, India, and Indonesia].” Here, the answer is mainly a paraphrased version of the question that includes the seed answers.

2.2 Paraphrasing Questions & Answers

Once we generate the first verbalized answer, inspired by [7, 9], we employ back-translation to produce multiple instances for the paraphrased answer. In particular, the back-translation is done using a transformer-based model [31] as translators. We produce paraphrases using translators for two sets of languages (English-Russian, English-German). For the selection, we considered the models’ performance on the WMT’18 dataset [4], including translations between different languages. The same paraphrasing process was applied to the rest of the conversational questions of the dataset. As output, we get paraphrased verbalized answers and questions.

2.3 Result Validation

Finally, to ensure the grammatical correctness of all the generated questions and answers, we include two rounds of a peer-review process. Similar to [8], the first round included a set of annotators where we asked them to assert the produced results and rephrase them if needed. This step ensures more natural and fluent dialogues. Next, another set of annotators were asked to validate the previous step and rephrase if needed. In particular, for the initial verbalization, 55% required some human interventions. While for paraphrased utterances, 47% required corrections during the peer-reviewed process. Table 2 indicates statistics of our generated dataset for all three sets.

3 EXPERIMENTS

3.1 Modeling Approaches

While the proposed dataset refers to the new task of “KGQA with verbalized answers”, our experiments focus only on generating the verbalized answers using the following inputs 1) the current question utterance and 2) conversation history. We assume the seed answer is given. Hence, we intend to study the ability of different models in generating the verbalized answer instead of producing the correct seed answer. We employ the following models:

- CNNSeq2seq [11]: a convolutional based encoder-decoder model.
- Transformers [31]: a model based solely on attention mechanism.
Table 3: Baselines on Verbal-ConvQuestions dataset. We report BLEU-4 and METEOR scores for generating the verbalized answer. The best result is in bold and the second best is underlined.

| Domain   | Books | Music | Movies | TV Series | Soccer |
|----------|-------|-------|--------|-----------|--------|
| Models   | BLEU-4 | METEOR | BLEU-4 | METEOR | BLEU-4 | METEOR | BLEU-4 | METEOR | BLEU-4 | METEOR |
| CNNSeq2seq | 0.1331 | 0.3030 | 0.2388 | 0.5199 | 0.1176 | 0.4534 | 0.1151 | 0.4595 | 0.0556 | 0.4040 |
| Transformer | 0.3751 | 0.6893 | 0.3594 | 0.6892 | 0.2633 | 0.6496 | 0.2301 | 0.6310 | 0.1995 | 0.5801 |
| BERTSeq2seq | 0.4231 | 0.7266 | 0.4403 | 0.7620 | 0.3117 | 0.6815 | 0.3063 | 0.6799 | 0.2396 | 0.6150 |
| BART     | 0.6088 | 0.8577 | 0.6524 | 0.8962 | 0.5925 | 0.8268 | 0.5416 | 0.8484 | 0.4616 | 0.7132 |
| T5       | 0.5540 | 0.8264 | 0.4033 | 0.7435 | 0.3974 | 0.7542 | 0.4799 | 0.8156 | 0.3323 | 0.6626 |

3.3 Results

Table 3 summarizes the results of the baseline models. We perceive that BART significantly outperforms the other baselines on all five domains for both the BLEU-4 and METEOR scores. The T5 model performs equally well. The main reason is that both models, BART and T5, have been pretrained for several generation tasks. Interestingly, all models perform worse in the “Soccer” domain; this occurs due to similar terms (e.g. “plays for”) it shares with other domains (e.g. “Movies” and “Books”). Furthermore, all models perform relatively well on domains such as “Music” and “Books”. As stated, the presented results only refer to generating verbalized answers using the conversational context as an input to the models. The results indicate promising performance and support the creation of the dataset.

Table 4: Error rate of baseline models.

| Models | Grammatical | Semantic | Entity Related |
|--------|-------------|----------|----------------|
| CNN    | 86%         | 91%      | 77%            |
| Transf. | 41%         | 48%      | 35%            |
| BERT   | 33%         | 41%      | 32%            |
| BART   | 16%         | 19%      | 14%            |
| T5     | 18%         | 23%      | 21%            |

3.4 Error Analysis

We randomly sampled 200 incorrect model predictions. We detail the error rates for three different categories. The first one, “Grammatical”, refers to examples where the verbalized answer can be understood but contain some grammatical errors such as a mismatch between the noun and verb forms. Next, the “Semantic” error class refers to the cases where the primary meaning of the answer has changed; this can occur by introducing new content (e.g., entities) or by omitting essential parts of the content. “Entity Related” refers to generated answers where the model failed to copy the correct entities from the input utterance or replaces them with pronouns. Table 4 presents the error rates for the three categories. We can observe that the BART and T5 models contain the lowest error rates, and therefore, validate their superior performance. Our empirical study provides a glance at the different types of errors that may occur while targeting the task of KGQA with verbalized answers.
3.5 Case Study

We manually examine some examples from the proposed dataset and discuss the intentions behind their construction. Regularly, all the required information is provided with the question/query for the question answering task. Therefore, to generate verbalized answers in such scenarios, we have to concentrate only on the question context, considering that the seed answer is given. However, in conversational question answering, we have scenarios such as anaphora and ellipsis [5, 28], where the context from previous turns has to be incorporated in order to answer the given question. Hence, we had to consider all these cases when creating the dataset. Table 5 illustrates such examples from our dataset, where conversational context was incorporated to generate the verbalization. For instance, in the first conversational example, the user asks the question “What was the birth city of Lionel Messi?”; the dataset here includes the verbalized answer “The birth city of Lionel Messi was Rosario, Santa Fe.”. Alongside the question and the verbalized answer, the dataset also provides paraphrases for both. In the next turn, we have the question “Is he a member of the Colombian National soccer team?” where he refers to “Lionel Messi”. For such examples, we either provide verbalized answers using the relevant pronoun (e.g. “No, he is not a member of the Colombian National soccer team.”) or even the entity itself, (e.g. “No, Lionel Messi does not represent Colombia at the international level.”). Similarly, on the next conversation, which belongs to the “TV series” domain, we provide such answers. More precisely, on turn four, the user asks the question “And how many seasons did the show last?” referring to the seed entity “Dexter” from first turn. Even in such scenarios, the dataset provides answers that also contain the entity (e.g. “Dexter is an 8 season television series.”).

4 RELATED WORK

Our work relates to previous approaches on answer verbalization for KGQA. However, we also briefly refer to existing approaches for answer generation on reading comprehension QA. There exist three answer verbalization datasets for KGQA systems. The first dataset, named VQuAnD等领域 [15], extended an existing (complex) QA dataset by generating one verbalized response for each question. Following, the authors proposed [12] to add multiple paraphrased answer verbalizations and illustrated that multiple answers positively affect the performance of machine learning models. Work from Biswas et al. [3] proposed answer verbalization on existing large-scale QA datasets only for simple questions. However, these datasets do not cover multiple question turns involving anaphora/ellipses. To the best of our knowledge, the here proposed Verbal-ConvQuestions is the first dataset that provides answer verbalization for conversational KGQA.

For reading comprehension approaches, work from Baheti et al. [1] studied the task of generating fluent QA responses in the context of building conversational agents. The authors proposed a framework that modifies the SQuAD 2.0 dataset [26] so that it includes conversational answers, which is used to train sequence-to-sequence based generation models. Another work [22] presents a BART-based [18] model for conversational answer generation and evaluate the validity of generated responses using NLI entailment.

Table 5: Verbal-ConvQuestions conversation examples. Each conversation in the dataset consists of five turns.

For question paraphrasing, we point readers to recent works by [30, 34].

5 CONCLUSIONS AND PREDICTED IMPACT

Conversational QA over KGs has been a trending research topic in scientific literature since Saha et al. [28] introduced the first dataset in this domain. However, unlike task-oriented dialog systems [19], the lack of verbalized answers was a significant research gap in existing datasets, hindering the development of approaches involving more natural conversations. We introduce a dataset, Verbal-ConvQuestions, that contains verbalized answers and multiple paraphrased utterances for each conversation. We further provide experiments with several baseline models to generate the answer utterances. Our error analysis illustrates error rates for model mispredictions in different categories. We believe that our empirical study, which highlighted gaps in the state-of-the-art models for answer verbalization, will serve as a basis for researchers. At the same time, to develop novel techniques for the introduced task.

REFERENCES

[1] Ashutosh Baheti, Alan Ritter, and Kevin Small. 2020. Fluent Response Generation for Conversational Question Answering. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 191–207. https://doi.org/10.18653/v1/2020.acl-main.19
[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://www.aclweb.org/anthology/W05-0909
[3] Debanjali Biswas, Mohinh embeddings, Md Rashad Al Haan Rony, and Jens Lehmann. 2020. VANiLLa: Verbalised ANswers in Natural Language at Large scale. (2020). https://sda.tech/projects/vanilla/
[4] Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Mark Finkel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Christof Kacupaj, et al.
An Answer Verbalization Dataset for Conversational Question Answerings over Knowledge Graphs

Mzon, et al. 2018. Proceedings of the Third Conference on Machine Translation: Shared Task Papers. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers.

[5] Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abuajabal, Jyotshaa Singh, and Gerhard Weikum. 2019. Look before You Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion. In Proceedings of the 2019 ACM International Conference on Information and Knowledge Management (Beijing, China) (CIKM ’19). Association for Computing Machinery, New York, NY, USA, 729–738. https://doi.org/10.1145/3357384.3358006

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[7] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, 489–500. https://doi.org/10.18653/v1/D18-1045

[8] Arash Einolghozati, Anchit Gupta, Keith Diedrick, and Sonal Gupta. 2020. Sound Translation, and Comprehension. In Proceedings of the 35th International Conference on Machine Learning - Volume 70 (Sydney, NSW, Australia) (ICML ’17). JMLR.org, 1243–1252.

[9] Endri Kacupaj, Barshana Banjere, Kuldeep Singh, and Jens Lehmann. 2021. Par(a)QA: A Question Answering Dataset with Paraphrase Responses for Single-Turn Conversation. In Eighteenth Extended Semantic Web Conference - Resources Track. https://opendataview.net/forum/id_GmcNeCEAaA3

[10] Endri Kacupaj, Joan Plepi, Kuldeep Singh, Harsh Thakkar, and Jens Lehmann. 2021. VOGUE: Answer Verbalization Through Multi-Task Learning. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 563–579. https://www.acmweb.org/anthology/2021.ecml-main.72

[11] Endri Kacupaj, Hamid Zafar, Jens Lehmann, and Maria Maleshkova. 2021. Conversational Question Answering over Knowledge Graphs with Transformer and Graph Attention Networks. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume Association for Computational Linguistics, Online, 850–862. https://www.aclweb.org/anthology/2021.eacl-main.703

[12] Endri Kacupaj, Javier Zafar, Jens Lehmann, and Maria Maleshkova. 2020. VOCE: Answer Verbalization Through Multi-Task Learning. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 563–579.

[13] Endri Kacupaj, Hamid Zafar, Jens Lehmann, and Maria Maleshkova. 2020. VOCE: Answer Verbalization Through Multi-Task Learning. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 563–579.

[14] Endri Kacupaj, Hamid Zafar, Jens Lehmann, and Maria Maleshkova. 2020. VOCE: Answer Verbalization Through Multi-Task Learning. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 563–579.

[15] Stan Peshterliev, Barlas Oguz, Debojeet Chatterjee, Hakan Inan, and Vikas Bhardwaj. 2021. Conversational Answer Generation and Factuality for Reading Comprehension Question-Answering. arXiv e-prints, Article arXiv:2103.06500 (March 2021), arXiv:2103.06500 pages. arXiv:2103.06500 [cs.CL]

[16] Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Yousef, and Mona Diab. 2019. Multi-domain goal-oriented dialogues (multidog): Strategies toward curating and annotating large scale dialogue data. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 4518–4528.

[17] Joan Plepi, Endri Kacupaj, Kuldeep Singh, Harsh Thakkar, and Jens Lehmann. 2021. Context transformer with stacked pointer networks for conversational question answering over knowledge graphs. In European Semantic Web Conference. Springer, 356–371.

[18] Colin Raffel, Noah Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yang Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[19] Pranav Rupakur, Robin Jia, and Percy Liang. 2018. Know What You Don’t Know: Unanswerable Questions for SQuAD. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Melbourne, Australia, 784–789. https://doi.org/10.18653/v1/A18-2124

[20] Sascha Rothe, Shashar Narayani, and Aliakssei Severyn. 2020. Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. Transactions of the Association for Computational Linguistics 8 (2020), 264–280. https://doi.org/10.1162/tacl_a_00313

[21] Amrita Saha, Varadana Pahuja, Mitesh M. Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018. Complex Sequential Question Answering: Towards Learning to Converse Over Linked Question Answer Pairs with a Knowledge Graph. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), the 38th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI-18). New Orleans, Louisiana, USA, February 2-7, 2018. Shaila A. Mellilha and Kilian Q. Weinberger (Eds.). AAAI Press, 705–713. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17181

[22] Kuldeep Singh, Arun Sethupati Radhakrishna, Andreas Both, Saeedeh Shekarpour, Ioana Lytra, Ricardo Usbeck, Akhlesh Vyas, Aemal Khaliqmatullayev, Dharman Punjani, Christoph Lange, Maria Esther Vidal, Jens Lehmann, and Sören Auer. 2018. Why Reinvent the Wheel: Let’s Build Question Answering Systems Together. In Proceedings of the 2018 World Wide Web Conference ( Lyon, France) (WWW ’18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CH, 1247–1256. https://doi.org/10.1145/3178676.3186023

[23] Svitala Vukalenko, Shanye Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 355–363.

[24] Ashish Vaswani, Noah Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, umbldedfukasz Kaiser, and Ila Polosukhin. 2017. Attention is All You Need. In 3rd ICLR, 2017.

[25] Stan Peshterliev, Barlas Oguz, Debojeet Chatterjee, Hakan Inan, and Vikas Bhardwaj. 2021. Conversational Answer Generation and Factuality for Reading Comprehension Question-Answering. arXiv e-prints, Article arXiv:2103.06500 (March 2021), arXiv:2103.06500 pages. arXiv:2103.06500 [cs.CL]

[26] Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Yousef, and Mona Diab. 2019. Multi-domain goal-oriented dialogues (multidog): Strategies toward curating and annotating large scale dialogue data. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 4518–4528.

[27] Joan Plepi, Endri Kacupaj, Kuldeep Singh, Harsh Thakkar, and Jens Lehmann. 2021. Context transformer with stacked pointer networks for conversational question answering over knowledge graphs. In European Semantic Web Conference. Springer, 356–371.

[28] Colin Raffel, Noah Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yang Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html