the QR metric as a proxy for the QA results: often questions do not have to be the same as the ground truth questions to solicit the same answers. The blank area in the top-left of the diagonal shows that a lexical overlap is required to produce the same answer set, which is likely due to the candidate filtering phase based on the bag-of-word representation matching.

We also compared ROUGE and Jaccard similarity for the tokens in reading comprehension results but they showed only a weak correlation (Pearson 0.31). This result confirms our observation that the extractive model tends to be rather sensitive to a slight input perturbation but will also often provide the same answer to very distinct questions.

Thus, our results show that the existing QR metrics have a limited capacity in predicting the performance on the end-to-end QA task. ROUGE correlates with the answer recall for the retrieval model, but cannot be used to reliably predict answer correctness for the individual question rewrites. Since ROUGE treats all the tokens equally (except for the stop-word list) it is unable to capture importance of the individual tokens that modify question semantics. The correlation of the QA performance with the embedding-based USE metric is even lower than with ROUGE for the both QA models.

6 Discussion

We showed that QA results can identify question paraphrases. However, this property directly depends on the ability of the QA model to match equivalent questions to the same answer set.

Wrong answer or wrong question? Our passage retrieval model is using BM25 as a filtering step, which relies on the lexical match between the terms in the question and the terms in the passage. Hence, synonyms, like "large" and "big", cannot be matched with this model. This effect explains that dissimilar questions are never matched to the same answers in Figure 2. The drawback of this model is that it suffers from the "vocabulary mismatch" problem (Van Gysel, 2017).

A one-word difference between a pair of questions may have a very little as well as a very dramatic effect on the question interpretation. This class of errors corresponds to the variance evident from Figure 2.

Our error analysis indicates that small perturbations of the input question, such as anaphora resolution, often cause a considerable change in the answer ranking. For example, the pair of the original question: "*Who are the Hamilton Electors and what were **they** trying to do?*", and the human rewrite: "*Who are the Hamilton Electors and what were the **Hamilton Electors** trying to do?*" produce ROUGE = 1 but R@1000 = 0.33. We also identified many cases in which inability of the QR component to generate apostrophes resulted in incorrect answers (original question: "*Describe **Netflixs** subscriber growth over time*", and the human rewrite: "*Describe **Netflix's** subscriber growth over time*").

In contrast, our reading comprehension models is based solely on the dense vector representations, which should be able to deal with paraphrases. In practice, we see from Figure 3 that this feature may also introduce an important drawback, when the model produces a correct answer even when given an incorrectly formulated question. This stability and undersensitivity of the reading comprehension model may indicate biases in the dataset and evaluation setup. When the answers no longer depends on the question being asked, but can be also predicted independently from the question, the QA model evidently fails to learn the mechanisms underlying QA. Instead, it may learn a shortcut in the benchmark dataset that allows to guess correct questions, such as likely answer positioning within an article.

Our experiments show that the retrieval-based setup is more adequate in judging model robustness. The size of the answer space is sufficiently large so as to exclude spurious cues that the model can exploit for shortcut learning. The role of QR is, therefore, much more evident in this task, since to be able to find the correct answer the question has to be formulated well by disambiguating conversational context. While it may be redundant to overspecify the question when given several passages to choose the answer from, it becomes of a vital importance when given several million passages. This argument holds, however, only when assuming a uniform answer distribution, which is often not the case unless for sufficiently large web collections. This particular observation brings us to the next question that we would like to discuss in more detail.

Rewrite or not to rewrite? Our experiments demonstrated the challenges in the quality control of the QR task itself. Human rewrites are not perfect themselves since it is not always clear whether and what should be rewritten. Considering the

human judgment of the QR quality independent from the QA results, little deviations may not seem important. However, from the pragmatic point of view, they may have a major impact on the overall performance of the end-to-end system.

The level of detail required to answer a particular question is often not apparent and depends on the dataset. However, we can argue that inconsistencies, such as typos and paraphrases, should be handled by the QA component, since they do not originate from the context interpretation errors.

Further on, we evaluated our assumption about human rewrites as the reliable ground truth. Our evaluation results indeed showed that human rewriting was redundant in certain cases. There are cases in which original questions without rewriting were already sufficient to retrieve the correct answers from the passage collection (see last row of the Table 1). In particular, we found that 10% of the questions in TREC CAsT were rewritten by human annotators that did not need rewriting to retrieve the correct answer. For example, original question: "*What is the functionalist theory?*", human rewrite: "*What is the functionalist theory in* ***sociology****?*". However, in another question from the same dialogue, omitting the same word from the rewrite leads to retrieval of an irrelevant passage, since there are multiple alternative answers.

The need for QR essentially interacts with the size and diversity of the possible answer space, e.g., collection content, with respect to the question. Some of the questions were correctly answered even with underspecified questions, e.g., original question: "*What are some ways to avoid injury?*", human rewrite: "*What are some ways to avoid* ***sports*** *injuries?*", because of the collection bias.

The goal of QR is to learn patterns that correct question formulation independent from the collection content. This approach is similar to how humans handle this task when formulating search queries, i.e., based on their knowledge of language and the world. Humans can spot ambiguity regardless of the information source by modeling the expected content of the information source even without the need to have direct access to it. Clearly, this expectation may not be optimal or even sufficient to be able to formulate a single precise question, which is exactly the point at which the need for an information-seeking dialogue naturally arises.

When the question formulation procedure is designed to be independent of the collection content, it also allows for querying several sources with the same question. This property is especially helpful when the content of the collection is unknown, which also allows to work with 3rd party APIs to access information in a distributed fashion. We may see a similar effect as in human-human communication here as well, when the need for an automated information-seeking dialogue between distributed systems may arise to better negotiate the information need and disambiguate the question further with respect to the content of each remote information source.

7 Related Work

Several recent research studies reported that state-of-the-art machine learning models lack in robustness across several NLP tasks (Li and Specia, 2019; Zeng et al., 2019). In particular, these models were found to be sensitive to input perturbations.

Jia and Liang (2017) analysed this problem in the context of the reading comprehension task. They showed in the experimental evaluation that QA models suffer from overstability, i.e., they tend to provide the same answer to a different question, when it is sufficiently similar to the correct question. Lewis and Fan (2019) also showed this deficiency of the state-of-the-art models for reading comprehension that learn to attend to just a few words in the question.

Our evaluation results support these findings using the conversational settings. We showed that many ambiguous questions in QuAC can be answered correctly without the conversation history (see Table 2). In contrast, this effect is absent in the passage retrieval setup (see Table 1 or Figure 1). These results suggest that the passage retrieval evaluation setup is more adequate for training robust QA models.

Previous work that focused on analysing and explaining performance of Transformer-based models on the ranking task used word attention visualisation and random removals of non-stop words (Dai and Callan, 2019; Qiao et al., 2019). The analysis framework proposed here provides a more systematic approach to model evaluation using ambiguous and rewritten question pairs, which is a by-product of applying QR in conversational QA.

Our approach is most similar to the one proposed by Ribeiro et al. (2018), who used semantically equivalent adversaries to analyse performance in several NLP tasks, including reading comprehen-

sion, visual QA and sentiment analysis. Similarly, in our approach, question rewrites, as paraphrases with different levels of ambiguity, are a natural choice for evaluation of the conversational QA performance. The difference is that we do not need to generate semantically equivalent questions but can reuse question rewrites as a by-product of introducing the QR model as a component of the conversational QA architecture.

8 Conclusion

QR is a challenging but a very insightful task designed to capture linguistic patterns that identify and resolve ambiguity in question formulation. Our results demonstrate the utility of QR as not only enabler for conversational QA but also as a tool that helps to understand when QA models fail.

We introduced an effective error analysis framework for conversational QA using QR and used it to evaluate sensitivity of two state-of-the-art QA model architectures (for reading comprehension and passage retrieval tasks). Moreover, the framework we introduced is agnostic to the model architecture and can be reused for performance evaluation of different models using other evaluation metrics as well.

QR helps to analyse model performance and discover their weaknesses, such as oversensitivity and undersensitivity to differences in question formulation. In particular, the reading comprehension task setup is inadequate to reflect the real performance of a question interpretation component since ambiguous or even incorrect question formulations are likely to result in a correct answer span. In passage retrieval, we observe an opposite effect. Since the space of possible answers is very large it is impossible to hit the correct answer by chance. We discover, however, that these models tend to suffer from oversensitivity instead, i.e., when even a single character will trigger a considerable change in answer ranking.

In future work, we should extend our evaluation to dense passage retrieval models and examine their performance using the conversational QA setup with the QR model. We should also look into training approaches that could allow QA models to further benefit from the QR component. As we showed in our experiments QR provides useful intermediate outputs that can be interpreted by humans and used for evaluation. We believe that QR models can be also useful for training more robust QA models. Both components can be trained jointly, which is inline with Lewis and Fan (2019), who showed that the joint objective of question and answer generation further improves QA performance.

Acknowledgements

We would like to thank our colleagues Srinivas Chappidi, Bjorn Hoffmeister, Stephan Peitz, Russ Webb, Drew Frank, and Chris DuBois for their insightful comments.

References

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2020. Open-Domain Question Answering Goes Conversational via Question Rewriting. *arXiv preprint*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 169–174.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2174–2184.

Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 985–988.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. CAsT 2019: The Conversational Assistance Track Overview. In *Proceedings of the 28th Text REtrieval Conference*. 13–15.

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*. 4171–4186.

Laura Dietz, Ben Gamari, Jeff Dalton, and Nick Craswell. 2018. TREC Complex Answer Retrieval Overview. TREC.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can You Unpack That? Learning to Rewrite Questions-in-Context. In *Proceedings of the*