# Multi-Document Answer Generation for Non-Factoid Questions

Valeriia Baranova-Bolotova
RMIT University
Melbourne, Australia
lurunchik@gmail.com

## ABSTRACT

Modern question answering (QA) systems have advanced to the point where they can provide answers for a wide range of user questions. However, when users ask a search system some complex questions that go beyond factoid question answering, for instance "how to come up with a great idea?", they, at best, see a summary of a single relevant HTML page. Typically the users would then have to manually go through several web-pages to satisfy their information needs. We envision that question answering systems should generate an answer as an aggregated summary of multiple relevant HTML pages. Our main research question is the following:

**RQ1** *How to generate an optimal answer to a non-factoid question from multiple supported documents?*

Currently, only one previous study addressed multi-document non-factoid question answering [2]. The authors presented a new long-form QA dataset ELI5 where each answer has a supported document created from several HTML pages. The study shows that even the best model performance was far worse than humans, as during the evaluation human-generated responses were preferred in 86% of cases over model-generated ones. Having experimented with the ELI5 dataset in our preliminary studies, we found that it has a key limitation connected with the unsupervised selection of relevant HTML pages. These HTML pages do not guarantee that the target answer could be created from them. In our work we propose a new dataset for multi-document non-factoid QA with answers and a set of supported documents generated by experts which covers one particular type of non-factoid questions: how-to or instrumental/procedural questions. This dataset, consisting of more than 100K QA pairs, is based on a dedicated web resource wikiHow[1] where answers are written by experts and have links to the source web-pages used by experts to create an answer.

The task of answer generation is as follows: given a set of texts and a question as input, the model should generate the target answer as an output. We will use a multi-task Seq2Seq model [2] as a baseline and compare it to our variants of Seq2Seq based architectures with different types of encoders and decoders and various text generation algorithms.

The evaluation of models for non-factoid QA is also a challenging task since such models are expected to generate abstract and long answers. Fan et al. [2] suggested using the summarization metric ROUGE [3] which is based on n-gram matching. Chen et al. [1] studies BERTScore along with a conditional BERTScore that incorporates the context and question into a metric. Chen et al. [1] shows that current metrics do not correlate well with human assessment on more difficult generative QA datasets. It is crucial for non-factoid QA to use metrics for evaluation which would be able to assign scores in the same way as humans. However, we first need to understand how people interact with answers to non-factoid questions and what are the criteria for a good answer. The subsequent research questions of our study are:

**RQ2** *How do people interact with answers to non-factoid questions? What are the criteria for an optimal answer and the target structure of an answer for different types of questions?*

**RQ3** *What metrics should be used for the evaluation of answers to non-factoid questions?*

In a preliminary study, we investigated how people interact with correct and incorrect passage-length answers for non-factoid questions. Users were tasked with evaluating the answer's quality, correctness, completeness, and conciseness. Words in the answer were also annotated, both explicitly through user mark up of salient words and implicitly through user gaze data obtained from eye-tracking. Our results show that the correctness of an answer strongly depends on its completeness, while conciseness is less important. To further investigate non-factoid answer evaluation, we are planning to run a large-scale editorial study on assessment of non-factoid QA-pairs. Participants will be asked to assess a question type and evaluate an answer on a broader number of answer criteria. We will base our future non-factoid questions taxonomy on the one suggested by Tawfik et al. [4]. The results will allow us to understand the most important answer criteria, build a new taxonomy of non-factoid question types and answer structures, improve existing metrics and, hopefully, generalize a QA-generative model to the other non-factoid question types.

[1] https://www.wikihow.com/wikiHow:About-wikiHow


*SIGIR '20, July 25–30, 2020, Virtual Event, China*

ACM ISBN 978-1-4503-8016-4/20/07.
https://doi.org/10.1145/3397271.3401449

## REFERENCES

[1] Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating Question Answering Evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, Hong Kong, China, 119–124.

[2] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In *Proceedings of ACL 2019*. Association for Computational Linguistics, Florence, Italy, 3558–3567.

[3] Chin-Yew Lin. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain.*

[4] Andrew Tawfik, Arthur Graesser, Jessica Gatewood, and Jaclyn Gishbaugher. 2020. Role of questions in inquiry-based instruction: towards a design taxonomy for question-asking and implications for design. *Educational Technology Research and Development* (01 2020), 1–25.