

ANTIQUÉ: A Non-Factoid Question Answering Benchmark

Helia Hashemi

Center for Intelligent Information Retrieval
University of Massachusetts Amherst
Amherst, MA 01003
hhashemi@cs.umass.edu

Hamed Zamani

Center for Intelligent Information Retrieval
University of Massachusetts Amherst
Amherst, MA 01003
zamani@cs.umass.edu

Mohammad Aliannejadi

Faculty of Informatics
Università della Svizzera italiana (USI)
Lugano, Switzerland
mohammad.alian.nejadi@usi.ch

W. Bruce Croft

Center for Intelligent Information Retrieval
University of Massachusetts Amherst
Amherst, MA 01003
croft@cs.umass.edu

ABSTRACT

Considering the widespread use of mobile and voice search, answer passage retrieval for non-factoid questions plays a critical role in modern information retrieval systems. Despite the importance of the task, the community still feels the significant lack of large-scale non-factoid question answering collections with real questions and comprehensive relevance judgments. In this paper, we develop and release a collection of 2,626 open-domain non-factoid questions from a diverse set of categories. The dataset, called ANTIQUÉ, contains 34,011 manual relevance annotations. The questions were asked by real users in a community question answering service, i.e., Yahoo! Answers. Relevance judgments for all the answers to each question were collected through crowdsourcing. To facilitate further research, we also include a brief analysis of the data as well as baseline results on both classical and recently developed neural IR models.

1 INTRODUCTION

With the rising popularity of information access through devices with small screens, e.g., smartphones, and voice-only interfaces, e.g., Amazon’s Alexa and Google Home, there is a growing need to develop retrieval models that satisfy user information needs with sentence-level and passage-level answers. This has motivated researchers to study answer sentence and passage retrieval, in particular in response to *non-factoid* questions [1, 20]. Non-factoid questions are defined as open-ended questions that require complex answers, like descriptions, opinions, or explanations, which are mostly passage-level texts. Questions such as “what is the reason for life?” are categorized as non-factoid questions. We believe this type of questions plays a pivotal role in the overall quality of question answering systems, since their technologies are not as mature as those for factoid questions, which seek precise facts, such as “At what age did Rossini stop writing opera?”

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

arXiv:1905.08957v2 [cs.LG] 19 Aug 2019

© 2019 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

Despite the widely-known importance of studying answer passage retrieval for non-factoid questions [1, 2, 8, 20], the research progress for this task is limited by the availability of high-quality public data. Some existing collections, e.g., [8, 14], consist of few queries, which are not sufficient to train sophisticated machine learning models for the task. Some others, e.g., [1], significantly suffer from incomplete judgments. Most recently, Cohen et al. [3] developed a publicly available collection for non-factoid question answering with a few thousands questions, which is called WikiPassageQA. Although WikiPassageQA is an invaluable contribution to the community, it does not cover all aspects of the non-factoid question answering task and has the following limitations: (i) it only contains an average of 1.7 relevant passages per questions and does not cover questions that have multiple aspects in multiple passages; (ii) it was created from the Wikipedia website, containing only formal text; (iii) more importantly, the questions in the WikiPassageQA dataset were generated by crowdworkers, which is different from the questions that users ask in real-world systems; (iv) the relevant passages in WikiPassageQA contain the answer to the question in addition to some surrounding text. Therefore, some parts of a relevant passage may not answer any aspects of the question; (v) it only provides binary relevance judgments.

To address these shortcomings, in this paper, we create a novel dataset for non-factoid question answering research, called ANTIQUÉ,¹ with a total of 2,626 questions. In more detail, we focus on the non-factoid questions that have been asked by users of Yahoo! Answers, a community question answering (CQA) service. Non-factoid CQA data without relevance annotation has been previously used in [1], however, as mentioned by the authors, it significantly suffers from incomplete judgments.² We collected a set of four-level relevance judgments through a careful crowdsourcing procedure involving multiple iterations and several automatic and manual quality checks. Note that we, in particular, paid extra attention to collect reliable and comprehensive relevance judgments for the test set. Therefore, we annotated the answers after conducting result pooling among several term-matching and neural retrieval models. In summary, ANTIQUÉ provides annotations for 34,011 question-answer pairs, which is significantly larger than many comparable datasets.

¹ANTIQUÉ stands for answering non-factoid questions.

²More information on the existing collections is provided in Section 2.

We further provide brief analysis to uncover the characteristics of ANTIQUE. Moreover, we conduct extensive experiments with ANTIQUE to present benchmark results of various methods, including classical and neural IR models on the created dataset, demonstrating the unique challenges ANTIQUE introduces to the community. To foster research in this area, we release ANTIQUE for research purposes.³

2 EXISTING RELATED COLLECTIONS

Factoid QA Datasets. TREC QA [15] and WikiQA [19] are examples of factoid QA datasets whose answers are typically brief and concise facts, such as named entities and numbers. InsuranceQA [5] is another factoid dataset in the domain of insurance. ANTIQUE, on the other hand, consists of open-domain non-factoid questions that require explanatory answers. The answers to these questions are often passage level, which is contrary to the factoid QA datasets.

Non-Factoid QA Datasets. There have been efforts for developing non-factoid question answering datasets [7, 8, 18]. Keikha et al. [8] introduced the WebAP dataset, which is a non-factoid QA dataset with 82 queries. The questions and answers in WebAP were not generated by real users. There exist a number of datasets that partially contain non-factoid questions and were collected from CQA websites, such as Yahoo! Webscope L6, Qatar Living [9], and Stack-Exchange. These datasets are often restricted to a specific domain, suffer from incomplete judgments, and/or do not contain sufficient non-factoid questions for training sophisticated machine learning models. The nfl6 dataset [1] is a collection of non-factoid questions extracted from the Yahoo! Webscope L6. Its main drawback is the absence of complete relevance annotation. Previous work assumes that the only answer that the question writer has marked as correct is relevant, which is far from being realistic. That is why we aim to collect a complete set of relevance annotations. WikiPassageQA is another non-factoid QA dataset that has been recently created by Cohen et al. [3]. As mentioned in Section 1, despite its great potentials, it has a number of limitations. ANTIQUE addresses these limitations to provide a complementary benchmark for non-factoid question answering.⁴ More recently, Microsoft has released the MS MARCO V2.1 passage re-ranking dataset [10], containing a large number of queries sampled from the Bing search engine. In addition to not being specific to non-factoid QA, it significantly suffers from incomplete judgments. In contrast, ANTIQUE provides a reliable collection with complete relevance annotations for evaluating non-factoid QA models.

Machine Reading Comprehension (MRC) Datasets. MRC has recently attracted a great deal of attention in the NLP community. The MRC task is often defined as selecting a specific short text span within a sentence, selecting the answer from predefined choices, or predicting a blanked-out word of a sentence. There exist a number of datasets for MRC, such as SQuAD [13], BAbI [16], and MS MARCO v1 [10]. In this paper, we study retrieval-based QA tasks, thus MRC is out of the scope of the paper.

The screenshot shows a web interface for a Human Intelligence Task (HIT). At the top, there is a 'Question' section with the text: 'How do you prevent chicken from drying out when you cook it?'. Below that is a 'Possibly Correct Answer' section with a paragraph of text: 'The dark meat of the chicken retains moistness more so than the breast meat. Try recipes using legs and thighs instead of breast meat. Also leave the skin on when cooking (you can always remove it later). When barbecuing I marinate chicken thighs 4-24 hours ahead of time/throw away the marinade afterwards to avoid food poisoning.' Below this is a 'Candidate Answer' section with the text: 'Candidate Answer: you need to stab you chicken between 5 and 10 depending on the size, with a fork. Then marinate it for at least 4 hours. When cooking, if in a pan, cook in the marinade. If on a grill brush the marinade on the chicken every couple of minutes'. There are four radio button options for labeling the candidate answer: 1. 'Is this a good answer?' with 'Yes, it looks reasonable, and convincing-label 4.' (selected), 2. 'Is this a bad answer?' with 'Yes, it talks about same general topic, but it doesn't provide the answer of question or it provides an unreasonable answer-Label 2.', 3. 'Yes, it's not convincing enough, but still it could be an alternative answer with lower quality-label 3.', and 4. 'Yes, it's completely off-topic, or it does not make any sense-Label 1.'. At the bottom, there is a 'Comment (optional)' section with a text input field and a 'Submit' button. A 'Task 4 out of 4' indicator is also present.

Figure 1: The HIT interface for answer relevance annotation.

3 DATA COLLECTION

In this section, we describe how we collected ANTIQUE. Following Cohen and Croft [1], we used the publicly available dataset of non-factoid questions collected from the Yahoo! Webscope L6, called nfl6.⁵

Pre-processing & Filtering. We conducted the following steps for pre-processing and question sampling: (i) questions with less than 3 terms were omitted (excluding punctuation marks); (ii) questions with no best answer (\hat{a}) were removed; (iii) duplicate or near-duplicate questions were removed. We calculated term overlap between questions and from the questions with more than 90% term overlap, we only kept one, randomly; (iv) we omitted the questions under the categories of “Yahoo! Products” and “Computers & Internet” since they are beyond the knowledge and expertise of most workers; (v) From the remaining data, we randomly sampled 2,626 questions (out of 66,634).

Each question q in nfl6 corresponds to a list of answers named “nbest answers,” which we denote with $\mathcal{A} = \{a_1, \dots, a_n\}$. For every question, one answer is marked by the question author on the community web site as the best answer, denoted by \hat{a} . It is important to note that as different people have different information needs, this answer is not necessarily the best answer to the question. Also, many relevant answers have been added after the user has chosen the correct answer. Nevertheless, in this work, we respect the users’ explicit feedback, assuming that the candidates selected by the actual user are relevant to the query. Therefore, we do not collect relevance assessments for those answers.

3.1 Relevance Assessment

We created a Human Intelligence Task (HIT) on Amazon Mechanical Turk,⁶ in which we presented workers with a question-answer pair, and instructed them to annotate the answer with a label between 1 to 4. The instructions started with a short introduction to

³<https://ciir.cs.umass.edu/downloads/Antique/>

⁴More information can be found in Section 1.

⁵<https://ciir.cs.umass.edu/downloads/nfl6/>

⁶<http://www.mturk.com/>

Table 2: The benchmark results by a wide variety of retrieval models on the ANTIQUE dataset.

Method	MAP	MRR	P@1	P@3	P@10	nDCG@1	nDCG@3	nDCG@10
BM25	0.1977	0.4885	0.3333	0.2929	0.2485	0.4411	0.4237	0.4334
DRMM-TKS [6]	0.2315	0.5774	0.4337	0.3827	0.3005	0.4949	0.4626	0.4531
aNMM [17]	0.2563	0.6250	0.4847	0.4388	0.3306	0.5289	0.5127	0.4904
BERT [4]	0.3771	0.7968	0.7092	0.6071	0.4791	0.7126	0.6570	0.6423

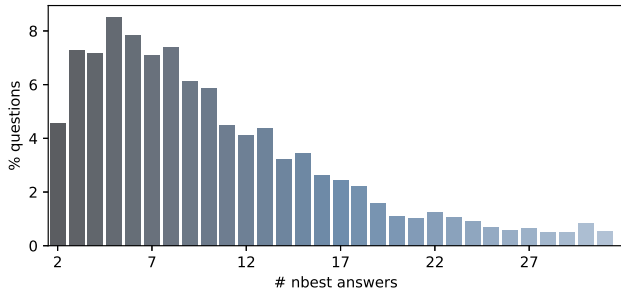


Figure 3: Distribution of the length of \mathcal{A} (i.e., nbest answers) per question.

answers per training question, which is significantly larger than its similar datasets, e.g., WikiPassageQA [3].

Test Set. The test set in ANTIQUE consists of 200 questions which were randomly sampled from nFL6 after pre-processing and filtering. Statistics of the test set can be found in Table 1. The set of candidate questions for annotation was selected by performing depth- k ($k = 10$) pooling. To do so, we considered the union of the top k results of various retrieval models, including term-matching and neural models (listed in Table 2). We took the union of this set and ‘nbest answers’ (set \mathcal{A}) for annotation.

4 DATA ANALYSIS

In this section, we present a brief analysis of ANTIQUE to highlight its characteristics.

Statistics of ANTIQUE. Table 1 lists general statistics of ANTIQUE. As we see, ANTIQUE consists of 2,426 non-factoid questions that can be used for training, followed by 200 questions as a test set. Furthermore, ANTIQUE contains 27.4k and 6.5k annotations (judged answers) for the train and test sets, respectively. We also report the total number of answers with specific labels.

Workers Performance. Overall, we launched 7 different crowdsourcing batches to collect ANTIQUE. This allowed us to identify and ban less effective workers. As we see in Table 1, a total number of 577 workers made over 148k annotations (257 per worker), out of which we rejected 12% because they failed to satisfy the quality criteria.

Questions Distribution. Figure 2 shows how questions are distributed in ANTIQUE by reporting the top 40 starting trigrams of the questions. As shown in the figure, majority of the questions start with “how” and “why,” constituting 38% and 36% of the questions, respectively. It is notable that, according to Figure 2, a considerable number of questions start with “how do you,” “how can you,” “what do you,” and “why do you,” suggesting that their corresponding answers would be highly subjective and opinion based. Also, we can see a major fraction of questions start with “how can I” and

“how do I,” indicating the importance and dominance of personal questions.

Answers Distribution. Finally, in Figure 3, we plot the distribution for the number of ‘nbest answers’ ($|\mathcal{A}|$). We see that the majority of questions have 9 or less nbest answers (=54%) and 82% of questions have 14 or less nbest answers. The distribution, however, has a long tail which is not shown in the figure.

5 BENCHMARK RESULTS

In this section, we provide benchmark results on the ANTIQUE dataset. To this aim, we report the results for a wide range of retrieval models (mostly neural models) in Table 2. In this experiment, we report a wide range of standard retrieval metrics, ranging from precision- to recall-oriented metrics (see Table 2). Note that for the metrics that require binary labels (i.e., MAP, MRR, and P@k), we assume that the labels 3 and 4 are relevant, while 1 and 2 are non-relevant. Due to the definition of our labels (see Section 3), we recommend this setting for future work. For nDCG, we use the four-level relevance annotations.⁹

As shown in the table, the neural model significantly outperforms BM25, an effective term-matching retrieval model. Among all, BERT [4] provides the best performance. Recent work on passage retrieval also made similar observations [11, 12]. Since MAP is a recall-oriented metric, the results suggest that all the models still fail at retrieving all relevant answers. There is still a large room for improvement, in terms of both precision- and recall-oriented metrics.

6 CONCLUSIONS

In this paper, we introduced ANTIQUE; a non-factoid community question answering dataset. The questions in ANTIQUE were sampled from a wide range of categories on Yahoo! Answers, a community question answering service. We collected four-level relevance annotations through a multi-stage crowdsourcing as well as expert annotation. In summary, ANTIQUE consists of 34,011 QA-pair relevance annotations for 2,426 and 200 questions in the training and test sets, respectively. Additionally, we reported the benchmark results for a set of retrieval models, ranging from term-matching to recent neural ranking models, on ANTIQUE. Our data analysis and retrieval experiments demonstrated that ANTIQUE introduces unique challenges while fostering research in the domain of non-factoid question answering.

7 ACKNOWLEDGEMENT

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF IIS-1715095. Any opinions, findings and conclusions or recommendations expressed in this

⁹Note that we mapped our 1 to 4 labels to 0 to 3 for computing nDCG.

material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

- [1] D. Cohen and W. B. Croft. 2016. End to End Long Short Term Memory Networks for Non-Factoid Question Answering. In *ICTIR '16*. 143–146.
- [2] D. Cohen and W. B. Croft. 2018. A Hybrid Embedding Approach to Noisy Answer Passage Retrieval. In *ECIR '18*.
- [3] D. Cohen, L. Yang, and W. B. Croft. 2018. WikiPassageQA: A Benchmark Collection for Research on Non-factoid Answer Passage Retrieval. In *SIGIR '18*.
- [4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* (2018).
- [5] M. Feng, B. Xiang, M. R. Glass, L. Wang, and B. Zhou. 2015. Applying Deep Learning to Answer Selection: A Study and An Open Task. *CoRR* (2015).
- [6] J. Guo, Y. Fan, Q. Ai, and W. B. Croft. A Deep Relevance Matching Model for Ad-hoc Retrieval. In *CIKM '16*.
- [7] I. Habernal, M. Sukhareva, F. Raiber, A. Shtok, O. Kurland, H. Ronen, J. Bar-Ilan, and I. Gurevych. New Collection Announcement: Focused Retrieval Over the Web. In *SIGIR '16*.
- [8] M. Keikha, J. Park, and W. B. Croft. 2014. Evaluating Answer Passages using Summarization Measures. In *SIGIR '14*. 963–966.
- [9] P. Nakov, D. Hoogeveen, L. Márquez, A. Moschitti, H. Mubarak, T. Baldwin, and K. Verspoor. 2017. SemEval-2017 Task 3: Community Question Answering. In *SemEval '17*. 27–48.
- [10] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. 2016. MS MARCO: A Human Generated MACHine Reading COMprehension Dataset. *CoRR* abs/1611.09268 (2016).
- [11] R. Nogueira and K. Cho. 2019. Passage Re-ranking with BERT. *CoRR* abs/1901.04085 (2019).
- [12] H. Padigela, H. Zamani, and W. B. Croft. 2019. Investigating the Successes and Failures of BERT for Passage Re-Ranking. *CoRR* abs/1903.06902 (2019).
- [13] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *CoRR* (2016).
- [14] C. Shah and J. Pomerantz. 2010. Evaluating and Predicting Answer Quality in Community QA. In *SIGIR '10*.
- [15] M. Wang, N. A. Smith, and T. Mitamura. 2007. What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA. In *EMNLP '07*.
- [16] J. Weston, A. Bordes, S. Chopra, and T. Mikolov. 2015. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. *CoRR* (2015).
- [17] L. Yang, Q. Ai, J. Guo, and W. B. Croft. 2016. aNMM: Ranking Short Answer Texts with Attention-Based Neural Matching Model. In *CIKM '16*. 287–296.
- [18] L. Yang, Q. Ai, D. Spina, R.-C. Chen, L. Pang, W. B. Croft, J. Guo, and F. Scholer. 2016. Beyond Factoid QA: Effective Methods for Non-factoid Answer Sentence Retrieval. In *ECIR '16*.
- [19] Y. Yang, S. W. Yih, and C. Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. *ACL '15*.
- [20] E. Yulianti, R. Chen, F. Scholer, W. B. Croft, and M. Sanderson. 2018. Document Summarization for Answering Non-Factoid Queries. *TKDE* (2018).