ANTIQUE: A Non-Factoid Question Answering Benchmark

Helia Hashemi[1], Mohammad Aliannejadi[2*], Hamed Zamani[1**], and W. Bruce Croft[1]

[1] Center for Intelligent Information Retrieval, University of Massachusetts Amherst, Amherst, MA 01003
{hhashemi, zamani, croft}@cs.umass.edu

[2] Information and Language Processing Systems (ILPS), University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
m.aliannejadi@uva.nl

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 ANTIQUE, contains 34k 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 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, 18]. 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 like "How to cook burgers?" are non-factoid. 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?".

Despite the widely-known importance of studying answer passage retrieval for non-factoid questions [1, 2, 8, 18], the research progress for this task is limited by the availability of high-quality public data. Some existing collections, e.g., [8, 13], consist of few queries, which are not sufficient to train sophisticated machine learning models for the task. Some others, e.g., [1], significantly

* Work done while affiliated with Università della Svizzera italiana (USI), Switzerland.
** Hamed Zamani is currently affiliated with Microsoft.

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 question and does not cover many questions with multiple correct answers; (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 labels.

To address these shortcomings, in this paper, we create a novel dataset for non-factoid question answering research, called *ANTIQUE*, 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 (see Section 2 for more information on existing collections). We collected four-level relevance labels through a careful crowdsourcing procedure involving multiple iterations and several automatic and manual quality checks. Note that we 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, ANTIQUE provides annotations for 34,011 question-answer pairs, which is significantly larger than many comparable datasets.

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.[3]

2 Existing Related Collections

Factoid QA Datasets. TREC QA [14] and WikiQA [17] 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, 16]. 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

[3] https://ciir.cs.umass.edu/downloads/Antique/

number of datasets that partially contain non-factoid questions and were collected from CQA websites, such as Yahoo! Webscope L6, Qatar Living [9], and StackExchange. 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 (see Section 1). 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.

3 Data Collection

Following Cohen et al. [1], we used the publicly available dataset of non-factoid questions collected from the Yahoo! Webscope L6, called nfL6. 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 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 $\mathcal{A} = \{a_1, \ldots, a_n\}$. 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, 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 the task and its motivations, followed by detailed annotation guidelines. Since workers needed background knowledge for answering the

Table 1: Statistics of ANTIQUE.

# training (test) questions:	2,426 (200)	**# label 4:**	13,067	**# total workers:**	577
# training (test) answers:	27,422 (6,589)	**# label 3:**	9,276	**# total judgments:**	148,252
average question length:	10.51	**# label 2:**	8,754	**# rejected judgments:**	17,460
average answer length:	47.75	**# label 1:**	2,914	**% of rejections:**	12%

majority of the questions, we also included $\hat{a}$ in the instructions and called it a "possibly correct answer." In some cases, the question is very subjective and could have multiple correct answers. This is why it is called "possibly correct answer" to make it clear in the instructions that other answers could potentially be different from the provided answer, but still be correct.

Label Definitions. To facilitate the labeling procedure, we described labels in the form of a flowchart to users. Our aim was to preserve the notion of relevance in QA systems as we discriminate it with the typical topical relevance definition in ad-hoc retrieval tasks. The definition of each label is as follows: **Label 4**: It looks reasonable and convincing. Its quality is on par with or better than the "Possibly Correct Answer". Note that it does not have to provide the same answer as the "Possibly Correct Answer". **Label 3**: It can be an answer to the question, however, it is not sufficiently convincing. There should be an answer with much better quality for the question. **Label 2**: It does not answer the question or if it does, it provides an unreasonable answer, however, it is not out of context. Therefore, you cannot accept it as an answer to the question. **Label 1:** It is completely out of context or does not make any sense.

We included 15 diverse examples of annotated QA pairs with explanation of why and how the annotations were done. Overall, we launched 7 assignment batches, appointing 3 workers to each QA pair. In cases where the workers could agree on a label (i.e., majority vote), we considered the label as the ground truth. We then added all QA pairs with no agreement to a new batch and performed a second round of annotation. It is interesting to note that the ratio of pairs with no agreement was nearly the same among the 7 batches (˜13%). In the very rare cases of no agreement after two rounds of annotation (776 pairs), an expert annotator decided on the final label. To allow further analysis, we have added a flag in the dataset identifying the answers annotated by the expert annotator. In total, the annotation task costed 2,400 USD.

Quality Check. To ensure the quality of the data, we limited the HIT to the workers with over 98% approval rate, who have completed at least 5,000 assignments. 3% of QA pairs were selected from a set of quality check questions with obviously objective labels. It enabled us to identify workers who did not provide high-quality labels. Moreover, we recorded the click log of the workers to detect any abnormal behavior (e.g., employing automatic labeling scripts) that would affect the quality of the data. Finally, we constantly performed manual quality checks by reading the QA pairs and their respective labels. The manual inspection was done on the 20% of each worker's submission as well as the QA pairs with no agreement.

Training Set. In the training set, we annotate the list $\mathcal{A}$ (see Section 3) for each query, and assume that for each question, answers to the other questions are

irrelevant. As we removed similar questions from the dataset, this assumption is fair. To test this assumption, we sampled 100 questions from the filtered version of nfL6 and annotated the top 10 results retrieved by BM25 using the same crowdsourcing procedure. The results showed that only 13.7% of the documents (excluding $\mathcal{A}$) were annotated as relevant (label 3 or 4). This error rate can be tolerated in the training process as it enables us to collect significantly larger amount of training labels. On the other hand, for the test set we performed pooling to label all possibly relevant answers. In total, the ANTIQUE's training set contains 27,422 answer annotations as it shown in Table 1, that is 11.3 annotated candidate 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

Here, 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.

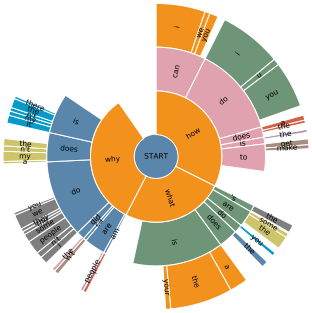

Fig. 1: Distribution of the top trigrams of ANTIQUE questions.

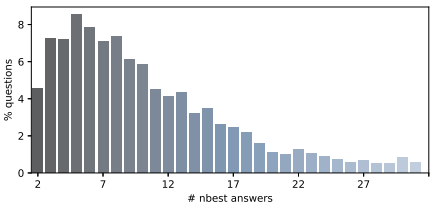

Fig. 2: Distribution of the length of $\mathcal{A}$ (i.e., nbest answers) per question.

Workers Performance. Overall, we launched 7 crowdsourcing batches to collect ANTIQUE. This allowed us to identify and ban less accurate workers. As reported 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 1 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

Table 2: The benchmark results for a wide variety of retrieval models on ANTIQUE.

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 [15]	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 1, 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 2, 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. We report the results for a wide range of retrieval models in Table 2. In this experiment, we report a wide range of standard precision- and recall-oriented retrieval 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 (we mapped our 1 to 4 labels to 0 to 3).

As shown in the table, the neural models significantly outperform 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

This paper introduced ANTIQUE; a non-factoid 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 for non-factoid question answering.

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

References

1. Cohen, D., Croft, W.B.: End to end long short term memory networks for non-factoid question answering. In: ICTIR '16. pp. 143–146 (2016)
2. Cohen, D., Croft, W.B.: A hybrid embedding approach to noisy answer passage retrieval. In: ECIR '18 (2018)
3. Cohen, D., Yang, L., Croft, W.B.: Wikipassageqa: A benchmark collection for research on non-factoid answer passage retrieval. In: SIGIR '18 (2018)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR (2018)
5. Feng, M., Xiang, B., Glass, M.R., Wang, L., Zhou, B.: Applying deep learning to answer selection: A study and an open task. CoRR (2015)
6. Guo, J., Fan, Y., Ai, Q., Croft, W.B.: A deep relevance matching model for ad-hoc retrieval. In: CIKM '16 (2016)
7. Habernal, I., Sukhareva, M., Raiber, F., Shtok, A., Kurland, O., Ronen, H., Bar-Ilan, J., Gurevych, I.: New collection announcement: Focused retrieval over the web. In: SIGIR '16 (2016)
8. Keikha, M., Park, J., Croft, W.B.: Evaluating answer passages using summarization measures. In: SIGIR '14. pp. 963–966 (2014)
9. Nakov, P., Hoogeveen, D., Màrquez, L., Moschitti, A., Mubarak, H., Baldwin, T., Verspoor, K.: SemEval-2017 task 3: Community question answering. In: SemEval '17. pp. 27–48 (2017)
10. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A human generated machine reading comprehension dataset. CoRR **abs/1611.09268** (2016)
11. Nogueira, R., Cho, K.: Passage re-ranking with BERT. CoRR **abs/1901.04085** (2019)
12. Padigela, H., Zamani, H., Croft, W.B.: Investigating the successes and failures of bert for passage re-ranking. CoRR **abs/1903.06902** (2019)
13. Shah, C., Pomerantz, J.: Evaluating and predicting answer quality in community qa. In: SIGIR '10 (2010)
14. Wang, M., Smith, N.A., Mitamura, T.: What is the jeopardy model? a quasi-synchronous grammar for qa. In: EMNLP '07 (2007)
15. Yang, L., Ai, Q., Guo, J., Croft, W.B.: anmm: Ranking short answer texts with attention-based neural matching model. In: CIKM '16. pp. 287–296 (2016)
16. Yang, L., Ai, Q., Spina, D., Chen, R.C., Pang, L., Croft, W.B., Guo, J., Scholer, F.: Beyond factoid qa: Effective methods for non-factoid answer sentence retrieval. In: ECIR '16 (2016)
17. Yang, Y., Yih, S.W., Meek, C.: Wikiqa: A challenge dataset for open-domain question answering. In: ACL '15 (2015)
18. Yulianti, E., Chen, R., Scholer, F., Croft, W.B., Sanderson, M.: Document summarization for answering non-factoid queries. TKDE (2018)