Iteration	Precision	Recall	F1
1	0.736	0.601	0.662
2	0.758	0.561	0.645
3	0.771	0.390	0.518
4	0.827	0.286	0.426
5	0.829	0.270	0.407

Table 1: Performance of the classifier on the annotated test set at the end of each iteration of the down-sampling procedure.

Iteration	Precision	Recall	F1
1	0.829	0.270	0.407
2	0.835	0.262	0.434
3	0.800	0.270	0.404
4	0.82	0.344	0.488
5	0.82	0.414	0.550

Table 2: Performance of the classifier on the annotated test set at the end of each iteration of the up-sampling procedure.

ate recall can still ensure the incorporation of large and diverse types of (p, q) tuples.

3 Experimental Results

This section describes the results of the iterative refinement strategy.

Test Set Creation: We first create a manually annotated test set to evaluate the effectiveness of the classifier at each step of the iterative refinement process. For this, we randomly sample 100 (p, q) tuples each from 7 different domains (Apple, cooking, gaming, money, photography, scifi, travel). These questions are either the last, second last or the third last comments of their corresponding posts. The annotated test set has a 7:3 ratio of positives to negatives.

Seed Dataset: It is created based on the method described in Section 2.2.2. It consists of 1,800,000 (p, q) tuples, amongst which 50% are randomly sampled negative instances. The classifier is then iteratively trained based on Algorithm 1.

3.1 Results of Iterative Refinement

The results of the down-sampling and the up-sampling procedure are discussed below:

3.1.1 Down-Sampling

Table 1 describes the performance of the classifier on the annotated test set during the down-sampling process. It can be clearly observed that the precision of the classifier increases with each iteration.

Metric	Without CQ	With CQ
P@1	0.751	0.791
P@2	0.399	0.416
P@3	0.278	0.287
P@4	0.214	0.220
P@5	0.174	0.178
MRR	0.791	0.816

Table 3: Performance on the task of question-answer retrieval. CQ stands for clarification question. P@k represents the precision at the kth position of the ranked list. MRR represents the Mean Reciprocal Rank.

Even though there is a substantial decline in recall, the down-sampling procedure helps in increasing the overall precision.

3.1.2 Up-Sampling

Table 2 describes the performance of the classifier on the annotated test set during the up-sampling process. It can be clearly observed that recall of the classifier increases with each iteration, although the final recall (i.e at iteration 5) is lower than the recall obtained in the first iteration of the down-sampling process. Given that there are a large number of (p, q) tuples, a drop in recall will not hamper the quality nor the diversity of the dataset. At the end of the process, we also observe that there is only a marginal drop in precision. Thus, at the end of the last iteration we are able to obtain a classifier which has a high precision and a reasonable recall.

3.2 Downstream Utility

We evaluate the utility of the clarification question in ClarQ by using it for the task of reranking answers. We first randomly sample 1000 (p, q) tuples from 11 different domains (Apple, askubuntu, biology, cooking, english, gaming, money, puzzling, scifi, travel, unix). Corresponding to each tuple, we randomly sample a list of 99 answers (from the same domain as that of the post) and append the actual answer to this list. We first rerank the answers based on the post alone. Later, we rerank the answers by concatenating the post and the clarifying question. Based on the results from Table 3, we observe that concatenating the clarification question to the post does help in improving the performance. The success of this experiment depicts the usefulness of our created dataset.

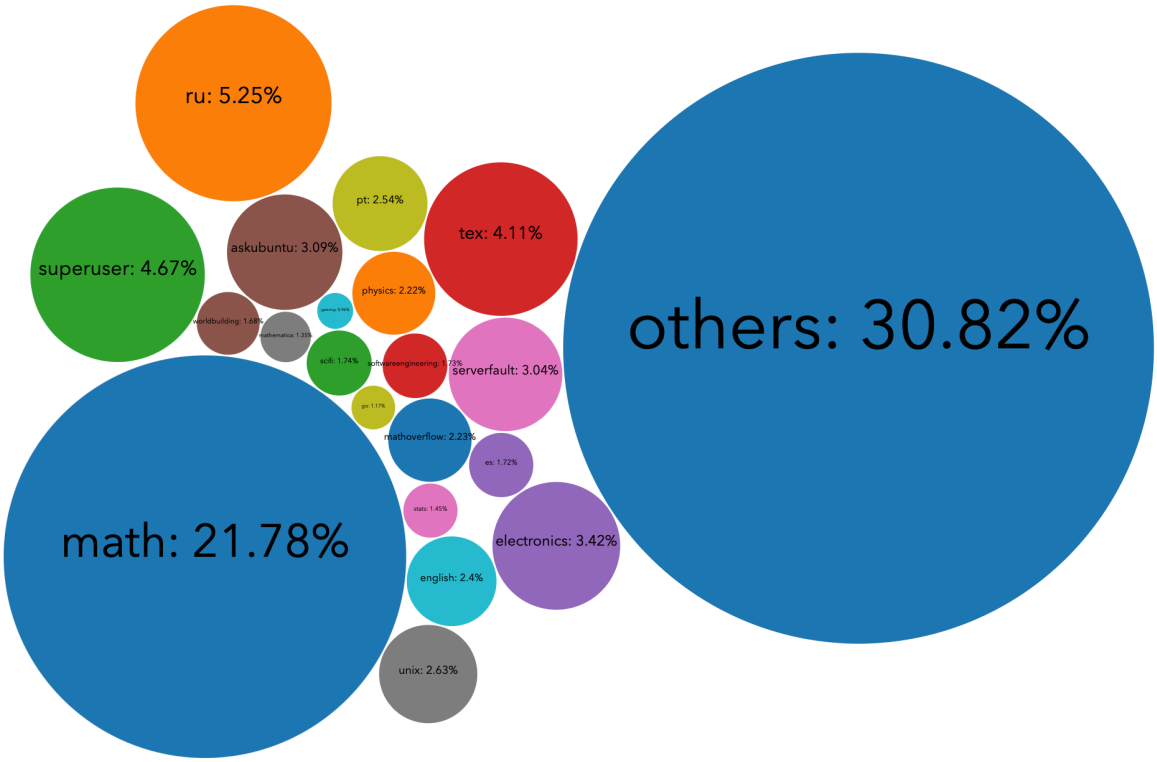

Figure 1: Distribution of the Clarifying Questions across different domains. The figure depicts the top 20 domains. Rest of the domains are clubbed at the end of the spectrum in "others".

4 Dataset Statistics

The classifier obtained at the end of iterative refinement procedure is used for classifying the initially collected (p, q) tuples of 6,186,934. The classifier predicts 2,079,300 tuples as actual clarification questions. As can be seen from Figure 1, these tuples are unequally distributed across 173 different domains. The top 20 domains account for 69.18% of the total (p, q) tuples in the dataset. The remaining 155 domains account for the remaining 30.82% of the total number of tuples.

It is noteworthy that our provided dataset also comprises of actual answers to each post. This would help researchers in evaluating the quality of the clarification questions in a standalone perspective and at the same time with respect to the downstream task of question-answering.

5 Conclusion and Future Work

In this paper, we present a diverse, large-scale dataset (**ClarQ**) for the task of clarification question generation. It is created by a two-step iterative bootstrapping framework based on self-supervision. ClarQ consists of ∼2M post-question tuples spanning 173 different domains. We hope that this dataset will encourage research into clarification question generation and, in the long run, enhance dialog and question-answering systems.

Acknowledgments

We would like to extend our sincere gratitude to Abhimanshu Mishra, Mrinal Dhar and Yash Kumar Lal for helping us understand the structure of the comments and their distribution across domains.

References

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and De-*

velopment in Information Retrieval, pages 475–484. ACM.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 2019. What makes a good conversation?: Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 475. ACM.

Marco De Boni and Suresh Manandhar. 2003. An analysis of clarification dialogue for question answering. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–55. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jiwei Li, Alexander H Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2016. Learning through dialogue interactions by asking questions. *arXiv preprint arXiv:1612.04936*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM.

Julian McAuley and Alex Yang. 2016. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. *arXiv preprint arXiv:1805.04655*.

Sudha Rao and Hal Daumé III. 2019. Answer-based adversarial training for generating clarification questions. *arXiv preprint arXiv:1904.02281*.

Yansen Wang, Chenyi Liu, Minlie Huang, and Liqiang Nie. 2018. Learning to ask questions in open-domain conversational systems with typed decoders. *arXiv preprint arXiv:1805.04843*.

Zhou Yu, Ziyu Xu, Alan W Black, and Alexander Rudnicky. 2016. Strategy and policy learning for non-task-oriented conversational systems. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, pages 404–412.