Question Classification of CoQA (QCoC) Dataset

Abbas Saliimi Lokman, Mohamed Ariff Ameedeen, Ngahzaifa Ab. Ghani
Faculty of Computing,Universiti Malaysia Pahang
Pekan, Pahang

ABSTRACT:
This paper proposes a new dataset for question classification process. Named QCoC (Question Classification of CoQA), this dataset is created based on Stanford's CoQA (A Conversational Question Answering Challenge) dataset. The total of QCoC datapoint is 116630 (total of combined questionanswer pairs in CoQA training and evaluation dataset). Common question classification datasets are classifying question based on its paired answer's knowledge (the semantic of answer's context). For QCoC, classification is done differently that is per answer's feature (semantic and syntactic of answer's type). This paper discusses the question classification datasets, QA datasets, and justification of CoQA as selected base for QCoC. Then QCoC specification is discussed with class definition, classification method and result subsections. To the author's knowledge, such dataset is still nonexistent to date. This paper suggests that this type of dataset is useful in solving abstractive answers issue in Question-Answering (QA) system. While factual answers can be directly produced by regular QA system, abstractive answers need some additional components. Although it is a recognizable issue, lack of suitable dataset perhaps is the reason why such direction is not being pursued. With QCoC dataset made publicly available1, hopefully such direction is open for further exploration.

KEYWORDS:
Dataset; QA system; Natural language processing

REFERENCES

[1] A. S. Lokman, “Chatbot Development in Data Representation for Diabetes Education,” UMP, 2011.

[2] A. S. Lokman, M. A. Ameedeen, and N.A. Ghani, “A conceptual IR chatbot framework with automated keywords-based vector representation generation,” OP Conf Ser Mater Sci Eng 769(1):012020 IOP Publishing, 2020.

[3] A. S. Lokman, M. A. Ameedeen, and N.A. Ghani, “Modern chatbot systems: A technical review,” In Proceedings of the future technologies conference, pp. 1012-1023. Springer, Cham, 2018.

[4] A. S. Lokman, M. A. Ameedeen, and N.A. Ghani, “A Question-Answering System that Can Count,” Computational Science and Technology, Lecture Notes in Electrical Engineering, vol. 724, pp. 61–70, 2021.

[5] A. Singhal, J. Choi, D. Hindle, D. D. Lewis, and F. Pereira, “At&t at trec-7,” NIST SPECIAL PUBLICATION SP, pp. 239–252, 1999.