Fairness-Aware Question Answering for Intelligent Assistants

Sachin Pathiyan Cherumanal
RMIT University
Melbourne, Australia
sachin.pathiyan.cherumanal@student.rmit.edu.au

ABSTRACT

Conversational intelligent assistants, such as Amazon Alexa, Google Assistant, and Apple Siri, are a form of voice-only Question Answering (QA) system and have the potential to address complex information needs. However, at the moment they are mostly limited to answering with facts expressed in a few words. For example, when a user asks Google Assistant if coffee is good for their health, it responds by justifying why it is good for their health without shedding any light on the side effects coffee consumption might have [1]. Such limited exposure to multiple perspectives can lead to change in perceptions, preferences, and attitude of users, as well as to the creation and reinforcement of undesired cognitive biases. Getting such QA systems to provide a fair exposure to complex answers – including those with opposing perspectives – is an open research problem.

In my PhD work, I aim to the address the following research questions:

RQ1: How can we quantify the fair exposure of multiple perspectives? RQ1 Investigates techniques to quantify the fair exposure of multiple perspectives in a search result. We adapted existing ranking fairness metrics proposed by Yang and Stoyanovich [5] to evaluate top-heavy QA systems. As a preliminary study, we also explored a popular diversity metric, α-nDCG, since the idea of exposing different perspectives can be seen as a Search Results Diversification (SRD) problem. Besides being able to quantify fairness, our results also showed that, while fairness and diversity metrics might show similar characteristics, they measure different dimensions [2]. Furthermore, metrics proposed in RQ1 along with other fairness metrics [4], shall be used for evaluation in future research questions. While RQ1 focuses on quantifying the bias in exposure, RQ2 investigates which existing (or perhaps a novel) optimisation technique would help provide fairer results. In RQ2, I am particularly interested in multi-attribute fairness where I aim to achieve fairer results along multiple definitions of perspectives i.e., Stance (PRO/CON) and Sub-Topics (Healthcare, Tax, Poverty, and so on).

RQ2: How can we jointly optimize relevance and multi-attribute fairness in QA systems? While emphasizing on fairness, it is imperative that these systems also return relevant results to the users [1], especially since past studies have shown that the performance of a search systems can drop while trying to provide fairer results. SRD can be seen as a mechanism to strike a balance between diversity and relevance in a ranked list. So as a preliminary study, I had investigated the applicability of various SRD techniques in a fairness-aware ranking problem and attempted to optimize rankings for both fairness and relevance by using the evaluation framework proposed at the TREC 2021 Fair Ranking Track [3]. The results accentuated the fact that SRD would work for fairness optimization only in the case where proportionality was maintained. The future work includes developing novel framework to optimize fairness and relevance, with a focus on top-heavy QA systems. RQ1 and RQ2 looks at the problem from a ranked list perspective. However, the challenge of fairly exposing answers in a voice-only conversation still remains an open research problem. This leads to the RQ3.

RQ3: How can we fairly present multiple perspectives to the user in a multi-turn conversation without compromising user satisfaction? Current QA systems tend to stick to single-turn responses, which restricts the possibility of exposing different perspectives. Consequently, I intend to focus on multi-turn conversations. I shall use the findings from RQ1 and RQ2 to design a controlled laboratory user study that will help identify different strategies to fairly present multiple perspectives in a multi-turn voice-only conversation. I intend to use the Wizard of Oz technique, where a user interacts with a voice-only search system and an intermediary performs the search like a conversational search system would.

To summarize, in this research, I aim to address the problem of fairly exposing multiple perspectives and relevant answers to users in a multi-turn conversation without negatively impacting user satisfaction.

REFERENCES

[1] Ruoyuan Gao and Chirag Shah. 2020. Toward Creating a Fairer Ranking in Search Engine Results. *Information Processing & Management* 57, 1 (2020), 102138. https://doi.org/10.1016/j.ipm.2019.102138

[2] Sachin Pathiyan Cherumanal, Damiano Spina, Falk Scholer, and W. Bruce Croft. 2021. Evaluating Fairness in Argument Retrieval. In *Proc. CIKM*. 3363–3367. https://doi.org/10.1145/3459637.3482099

[3] Sachin Payhiyan Cherumnala, Damiano Spina, Falk Scholer, and W. Bruce Croft. 2022. RMIT at TREC 2021 Fair Ranking Track. In *Proc. TREC*. https://trec.nist.gov/pubs/trec30/papers/RMIT-IR-F.pdf

[4] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists. In *Proc. WWW*. 553–562. https://doi.org/10.1145/3308560.3317595

[5] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proc. SSDBM*. Article 22, 6 pages. https://doi.org/10.1145/3085504.3085526