Evaluating Group Fairness and Relevance in Conversational Search: An Alternative Formulation

Tetsuya Sakai tetsuyasakai@acm.org Waseda University/Naver Corporation Japan Sijie Tao tsjmailbox@ruri.waseda.jp Waseda University Japan

Young-In Song song.youngin@navercorp.com Naver Corporation Korea

Abstract

The Conversational Search (CS) subtask of the NTCIR-18 FairWeb-2 task used Sakai's GFRC (Group Fairness and Relevance for Conversations) measure for evaluating the participating systems. As the Relevance and Group Fairness components were not directly integrated in GFRC and the measure lacked a clear user model, the present pilot study discusses an alternative called GFRC2. By directly transferring the general idea of the GFR (Group Fairness and Relevance) framework for web search to the task of evaluating generated conversations, we formulate GFRC2 as a form of expected user experience for a population of users who go through the words within the conversation. This also lets us visualise the Relevance and Group Fairness component scores for each cluster of users who are assumed to abandon the conversation at a particular relevant nugget. We demonstrate the steps of computing GFRC2 using real runs from the FairWeb-2 CS subtask.

Keywords

conversational search, evaluation, group fairness, nuggets

1 Introduction

Normlised discounted cumulative gain (nDCG) was proposed in 2002 [12];¹ after twenty three years, it is still the most widely used ranked retrieval measure in information retrieval and related tasks. In contrast, even though the recent and rapid advent of large language models is forcing the IR community to shift from ranked document retrieval to *conversational search*, there is no *de facto* standard measure for *conversational search evaluation* as of to-day [3, 18, 20, 22, 33, 37]. Compared to ranked document retrieval, there are at least² three major challenges in conversational search evaluation: (a) The system responses may be multi-turn, and their nature depends on the previous turns; (b) There is no predefined *retrieval unit*; we need to deal with a series/trees of texts instead of a ranked list of items; (c) Relevance is *not* all you need.

To address Challenge (a), simulating user-system interaction (e.g., [15, 34, 38]) and collecting real user-system conversations should probably complement each other, as the latter is more realistic but generally does not scale. In the present study, we take a few examples from real user-system conversations obtained from the Conversational Search (CS) subtask of the NTCIR-18 FairWeb-2 task [35], some of which are multi-turn while others are single-turn.

To address Challenge (b), nugget-based evaluation is regaining popularity [1, 2, 6, 11, 24, 28]. Although nugget-based approaches have been explored for evaluating textual outputs well before the advent of the Large Language Models (LLMs) [5, 9, 14, 19, 29, 36], the LLMs themselves are making the nugget-based approaches far more feasible than ever before, by freeing researchers from the burden of fully manual nugget extraction. The present study also takes the nugget-based approach: our premise is that nuggets have already been identified within the system turns for evaluating conversational search; more specifically, we leverage the nuggets from the aforementioned CS subtask of the NTCIR-18 FairWeb-2 task, which represent relevant entities, e.g., researchers (for Researcher topics).³ Note that, while a nugget is often defined as an atomic factual claim, we define it more broadly as any part of a conversational text that is associated with a position (e.g., word count) within the conversation and helps us assess the quality of the conversation.⁴

Regarding (c), it goes without saying that researchers should make sure that the behaviours of conversational search systems (and "artificial intelligence" systems in general) align with human values [17]. Sakai [28] lists 20(+1) axes for evaluating conversational search systems based on nuggets. Following the FairWeb-2 task, we consider two axes from the list: the correctness of a nugget (which we equate with relevance or utility) and fair exposure (which we treat as group fairness).⁵ While both the TREC Fair Ranking Tracks [8] and the NTCIR FairWeb tasks [35] have evaluated document ranking based on relevance and group fairness, there have been very few similar efforts or discussions for conversational search [10, 13]. For example, although Liu et al. [16] have proposed to evaluate fair treatment [7, 28] (which they call user-oriented fairness) for conversational recommendation by taking the difference between the average score over an active group (i.e., people who actively interact with a recommendation system) and that over an inactive group, their work is not about fair exposure; neither can they handle more than two groups.

In the present study, we propose an alternative to the GFRC (Group Fairness and Relevance for Conversations) measure [27] used in the CS subtask of the NTCIR-18 FairWeb-2 task [35]: we propose *GFRC2*, by directly transferring the general idea of the GFR (Group Fairness and Relevance) framework for web search to the task of evaluating generated textual conversations. GFRC2 can be seen as a form of *expected user experience* for a population of users who go through the words within the conversation; this

¹Well, actually, nDCG is very similar the *Sliding Ratio* measure proposed in 1968 [23]; the main difference is that nDCG incorporates a logarithmic function.

²For example, *reusability* is another major challenge: evaluating generated text is very different from traditional IR where document IDs of previously identified relevant documents can often be useful for evaluating new systems.

 $^{^3{\}rm The}$ FairWeb-2 task also collected movie entities (for Movie topics) and YouTube video entities (for YouTube topics) [35].

⁴For example, in addition to factual claims and relevant entities, *dialogue acts* could also be considered to be nuggets [28].

⁵*Individual fairness* [7] is outside the scope of the present study.

lets us visualise the Relevance and Group Fairness component scores for each cluster of users who are assumed to abandon the conversation at a particular relevant nugget. We demonstrate the steps of computing GFRC2 using real runs from the FairWeb-2 CS subtask.

2 GFRC: The Approach used in the NTCIR-18 FairWeb-2 CS subtask

The NTCIR-18 FairWeb-2 CS subtask used Sakai's GFRC [27], which is described below. (This section is largely duplicated from Sakai [27] and is not our new contribution.)

2.1 GFRC: A Generic Formulation

We want to evaluate a *T*-round user-system textual conversation $C = (U_1, S_1, \ldots, U_T, S_T)$. Let n_{ij} denote the *j*-th *nugget* in the *i*-th system turn S_i ; we define n_{ij} to be any substring of S_i that represents a *relevant* piece of information in the context of the previous turn sequence (U_1, S_1, \ldots, U_i) as well as the previous nuggets in S_i , i.e., $(n_{i1}, \ldots, n_{i(j-1)})$. For convenience, n_{ij} is denoted as an *element* of *set* S_i . Furthermore, let $pw(n) (\in [0, 1])$ denote the position-based weight of nugget *n*, where the position is defined in the context of conversation *C*. Let $g(n) (\in [0, 1])$ denote the gain value of nugget *n*. The R (relevance) score of conversation *C* is defined as

$$R(C) = \frac{1}{\mathcal{N}} \sum_{i=1}^{T} \sum_{n_{ij} \in S_i} pw(n_{ij})g(n_{ij}), \qquad (1)$$

where N is a normalisation factor.

Meanwhile, given M attribute sets⁶ and a target distribution (i.e., a probability mass function) D_*^m for each $m(=1, \ldots, M)$, the same conversation C is evaluated in terms of group fairness as follows. For every system turn S that contains a relevant nugget, an *achieved distribution* $D^m(S)$ is computed based solely on the group memberships of relevant nuggets contained in S (the *independent distribution* method [27]). In the present study, we ignore system turns that do not contain any relevant nuggets.

Let PW(S) denote the position-based weight of *turn S*. The GF (group fairness) score of conversation *C* can be computed as:

$$GF(C) = \frac{1}{N'} \sum_{i=1}^{T} PW(S_i) \sum_{m=1}^{M} w_m DistrSim^m (D^m(S_i) \parallel D_*^m), \quad (2)$$

where \mathcal{N}' is a normalisation factor and w_m is a weight assigned to the *m*-th attribute set s.t. $\sum_{m=1}^{M} w_m = 1$. Following GFR for document ranking tasks [30], JSD (Jensen-Shannon Divergence) is used for computing the DistrSim (distribution similarity) function if the attribute set contains *nominal* groups, while either NMD (Normalised Match Distance) or RNOD (Root Normalised Order-aware Divergence) is used if the attribute set contains *ordinal* groups [30]. The definitions of the divergence measures can be found in Sakai [26]. Note that a DistrSim score is obtained as *one minus* a divergence score. As a "quick summary" measure for ranking systems, the following combined measure (GFRC) may be of some use:

$$GFR(C) = \alpha R(C) + (1 - \alpha)GF(C), \qquad (3)$$

where the α is a parameter. Sakai [27] recommends to report R and GF scores separately and to visualise the relationship between the R and GF scores.

2.2 GFRC: An Instantiation

This section describes the instance of GFRC that was proposed in Sakai [27] and implemented in the NTCIR-18 FairWeb-2 CS subtask.

Following *S-measure* [29], the position weight of each nugget (See Eq. 1) was instantiated as a linear function.

$$pw(n) = \max(0, 1 - \frac{wc(n) - 1}{L}),$$
 (4)

where wc(n) is the word count that represents the end position of nugget *n* within the conversation, and L = 1,000: assuming that the user's average reading speed is 250 words per minute, this *L* implies that the user has four minutes to gather information through a conversation. Thus, the first word within a conversation has the position weight of 1, while the 1,001-th word has the position weight of 0 and therefore considered worthless. The accompanying normalisation factor N (See Eq. 1) can then be given by:

$$\mathcal{N} = \sum_{l=1}^{L} \left(1 - \frac{l-1}{L} \right) = \frac{L+1}{2} .$$
 (5)

This is a "hard" normalisation factor, which represents a practically unattainable situation where every word in conversation C represents a relevant nugget. The instantiation of the R score should therefore be:

$$R(C) = \frac{2}{L+1} \sum_{i=1}^{T} \sum_{n_{ij} \in S_i} \max(0, 1 - \frac{wc(n_{ij}) - 1}{L})g(n_{ij}) .$$
(6)

However, due to a minor inconsistency in Sakai's paper [27, Eqs. 6-8], the R score implemented for the CS subtask was

$$R(C) = \frac{2}{L+1} \sum_{i=1}^{T} \sum_{n_{ij} \in S_i} \max(0, 1 - \frac{wc(n_{ij})}{L})g(n_{ij}) .$$
(7)

The normalisation factor for Eq. 7 should have been

$$\mathcal{N} = \sum_{l=1}^{L} \left(1 - \frac{l}{L} \right) = \frac{L-1}{2} , \qquad (8)$$

so that

$$R(C) = \frac{2}{L-1} \sum_{i=1}^{T} \sum_{n_{ij} \in S_i} \max(0, 1 - \frac{wc(n_{ij})}{L})g(n_{ij}) .$$
(9)

In the present study, for the purpose of comparison with our new GFRC2 measure, we use Eq. 6 to compute GFRC as this appears to be what Sakai originally intended.

On the other hand, the instantiation of the GF score (See Eq. 2) was as follows. Let S be the set of system turns *excluding* those that do not contain any relevant nugget, and simply let $PW(S_i) = 1$ iff $S_i \in S$, with $\mathcal{N}' = |S|$.

⁶An attribute set is a set of groups prepared for the purpose of evaluating group fairness. For example, the NTCIR-18 FairWeb-2 task considered the PRONOUN and HINDEX attribute sets for Researcher topics to encourage gender diversity in the conversation and to cover researchers at different career stages. The PRONOUN attribute set contained three *nominal* groups: he, she, and other. On the other hand, the HINDEX attribute set contained four *ordinal* groups based on the Google Scholar h-index values [35].

$$GF^{m}(C) = \begin{cases} 0 & \text{if } S = \phi \\ \frac{1}{|S|} \sum_{S_{i} \in S} DistrSim^{m}(D^{m}(S_{i}) \parallel D_{*}^{m}) & \text{otherwise.} \end{cases}$$
(10)

That is, the DistrSim's over relevant system turns are simply averaged.

For the FairWeb-2 CS subtask, final GFRC score (Eq. 3) was computed by simply averaging the R and GF scores.

3 GFRC2: The Proposed Alternative

According to the above formulation of GFRC, the R score and GF score are somewhat disconnected, in that (a) the position weight only applies to the R score; and (b) while the R score is nugget-based, the GF score is turn-based; and (c) the final GFRC score is just an average of these different scores. Here, we propose a simple alternative that has an easily interpretable user model, by transferring the idea behind the GFR measure for document ranking to conversational search evaluation.

The Sakai/Robertson Normalised Cumulative Utility (NCU) [31, 32] and GFR frameworks model a population of users (with the same information need) that examine the ranked list from the top and then abandon the list at different document ranks.⁷ A group of users who abandon the list at a particular rank r (typically due to satisfaction) constitutes one *user cluster*; for this cluster, the *utility* of the SERP is a function of the quality of the top r documents of the SERP [31, 32]; more generally, the total *user experience* for this cluster can be estimated as a linear combination of the *utility* and the *group fairness* of the top r documents [30]. By taking into account the users' abandoning probability distribution over the ranks, NCU is computed as the *expected utility* over the user clusters, while GFR is computed as the *expected experience* over theses clusters.

Now, let us consider a similar user model for conversational search. Figure 1 depicts our idea. Imagine a population of users who examine the same user-system conversation. While a tree of user-system turns (i.e., conversations that branch out according toprevious user/system turns) is a possibility, for simplicity we assume that we have a sequence of alternating user-system turns as in the FairWeb-2 CS subtask.⁸ Just as we envision different users abandoning a ranked list of documents at different ranks, let us assume that conversational search users also abandon conversations at different positions within the conversation. Thus, a group of users who abandon a conversation at a particular position constitutes a user cluster, and we can assume that there is a probability distribution over these positions. For a user cluster uc that abandons the conversation at word count wc(uc), the total user experience can be computed as a function of the *utility* of the conversation from the beginning down to position wc(uc), and the overall group fairness of the same portion of the conversation. Thus, we can formulate GFRC2 as the expected total experience over the user clusters at different abandoning positions within the same conversation.

Formally,

$$GFRC2 = \sum_{uc} Prob(uc) Experience(uc) , \qquad (11)$$

⁸Even if we have a tree of turns, note that the different paths within the tree can still be treated as a sequence of user-system turns.

$$Experience(uc) = w_0 Utility(uc) + \sum_{m=1}^{M} w_m GF^m(uc) , \qquad (12)$$

where Prob(uc) is the probability that user cluster uc abandons the conversation at that position, Utility(uc) is the utility of the conversation for that user cluster, and $GF^m(uc)$ is the group fairness experienced by that user cluster with respect to the *m*-th attribute set, and $w_m(m = 0, ..., M)$ is a weight for each component score such that $\sum_{m=0}^{M} w_m = 1$.

In the present study, we consider a particular instantiation of GFRC2 by adopting word count-based notations as described below. As illustrated in Figure 1, note that the raw user-system conversation is first broken down into a list of words, as was done at the FairWeb-2 CS subtask.

First, we define the relevance indicator at each word count as:

$$I(wc) = \begin{cases} 1 & \text{if } wc \text{ is the end position of an } l\text{-relevant nugget} \\ & (l \in \{1, \dots, l_{\max}\}) \\ 0 & \text{otherwise} . \end{cases}$$
(13)

We then instantiate the abandoning probability at wc as:

$$Prob(wc) = \begin{cases} 1/L & \text{if } wc \le L \text{ and } I(wc) = 1\\ 0 & \text{otherwise} \end{cases}$$
(14)

The assumptions behind the above are that there are at most L relevant nuggets in a conversation and that the user distribution is uniform over these L nuggets (even if the actual conversation contains far fewer relevant nuggets). This is analogous to the user model behind *Average Precision* (AP) and *Q-measure* for ranked retrieval evaluation [25], except that we use the parameter L instead of the recall base (R) used in AP and Q.⁹

As for utility, we compute *Graded Nugget Precision* (GNP) at the end of each relevant nugget:

$$GNP(wc) = \frac{GWCrel(wc)}{WCnonrel(wc) + GWCrel(wc)},$$
(15)

$$GWCrel(wc) = \sum_{l=1}^{l_{max}} l * WCrel_l(wc) , \qquad (16)$$

where $WCrel_l(wc)$ denotes the number of words that correspond to a *l*-relevant nugget within the first *wc* words of the conversation, and *WCnonrel(wc)* denotes the number of words that do not correspond to any relevant nugget within those first *wc* words.

As for group fairness at the end of each relevant nugget, let

$$GF^{m}(wc) = DistrSim^{m}(D^{m}(wc)||D_{*}^{m}), \qquad (17)$$

where $D^m(wc)$ is the achieved probability distribution, defined as follows. Let rel(wc) denote the set of relevant nuggets found within the first wc words of the conversation, let N(m) denote the number of groups in the *m*-the attribute set, and let $(p_1^m(n), \ldots, p_{N(m)}^m(n))$ denote the group membership probability vector for relevant nugget *n*. Then the membership probability for Group $i(=1, \ldots, N(m))$ in $D^m(wc)$ is computed as:

$$\bar{p}_i(D^m(wc)) = \frac{1}{|rel(wc)|} \sum_{n \in rel(wc)} p_i^m(n) .$$
(18)

⁹In generative text evaluation, it is usually not possible to define recall.



Figure 1: An illustration of the user model behind GFRC2.

Finally, as a quick summary measure, we simply average the utility and GF scores to define the total experience, by letting $w_m = 1/(M + 1)$ for m = 0, ..., M in Eq. 12. Hence, Eq. 11 reduces to:

$$GFRC2 = \frac{1}{L} \sum_{wc} I(wc) \frac{GNP(wc) + \sum_{m=1}^{M} GF^{m}(wc)}{M+1} .$$
 (19)

We can also average GNP and GF scores individually: for example, if we want to focus on GNP (i.e., *utility*), letting $w_0 = 1$ and $w_m = 0$ for m = 1, ..., M in Eq. 12 will give us the *Expected GNP*

$$EGNP = \frac{1}{L} \sum_{wc} I(wc) GNP(wc) .$$
 (20)

Similarly, *Expected GF* can be computed for each attribute set, by letting $w_m = 1$ and $w_{m'} = 0$ for $m' \neq m$.

$$EGF^{m} = \frac{1}{L} \sum_{wc} I(wc) GF^{m}(wc) .$$
⁽²¹⁾

However, we argue that it is more important to visualise and observe the change in *utility* (quantified by GNP scores) and *group fairness* (quantified by GF scores) across the user clusters, as we shall demonstrate in the next section.

Note that, just like GFRC [27] and S-measure [29] (both of which use a linear decay function given by Eq. 4), GFRC2 values relevant nuggets found earlier in the conversation higher than those found later. To see this, suppose that a relevant nugget moves to an earlier position within the conversation while keeping the other relevant nuggets unchanged. This would not affect the group membership of the nugget, but would affect the GNP score in Eq. 19. If we further go back to Eq. 15, we can see that the GWCrel would be unaffected (since the set of relevant nuggets we are looking at is still the same after the above change). On the other hand, the WCnonrel in Eq. 19 would either remain the same or *decrease*, since the above change means that the *wc* we are considering is now smaller. That is, GNP would either remain the same or *increase*: the relevant nugget that has moved "up" generally *raises* the GFRC2 score.

4 GFRC in Action: Revisiting the Case Study from the FairWeb-2 CS subtask

By revisiting a case study from the NTCIR-18 FairWeb-2 overview paper, this section demonstrates how GFRC2 can be computed. Figures 2 and 3 have been duplicated from Tao et al. [35] (except that we have added the start position of each relevant nugget for the purpose of calculating GNP): these are the CS subtask conversations submitted by Team COPWA and the organisers for Topic R112 ("*Researchers who have published at least one paper at the ICLR conference.*").¹⁰

Tables 1 and 2 demonstrate the steps for computing the GFRC2 scores for these two runs; each line corresponds to a relevant nugget shown in Figures 2 and 3. For example, the first line of Table 1 for the first relevant nugget returned by COPWA at wc = 33 is obtained as follows.

- The relevant nugget (with a relevance level of l = 2) contains two words,¹¹ and hence *GWCrel* = 2 * 2 = 4 (Eq. 16) and *WCnonrel* = 33 2 = 31. Therefore, *GNP* = 4/(31 + 4) = 0.114286 (Eq. 15).
- Regarding the PRONOUN-based group fairness, the first relevant nugget is a "he" and the group membership vector is (1,0,0);¹² By comparing this achieved distribution with the gold uniform distribution in terms of JSD [26], *DistrSim* = 1 0.459148 = 0.540852.
- Regarding the HINDEX-based group fairness, the first relevant nugget belongs to Group 4 (i.e., researchers with a very high h-index); if we compare the group membership

¹⁰ https://waseda.box.com/fairweb2topics

¹¹Although it actually contains *three* words (*Yoshua, Bengio*, and the URL), the official word count data from the CS subtask does not include the word *Yoshua* as part of the nugget. This is because, as the first red line in Figure 2 shows, *Yoshua* is treated as part of *S:Yoshua*, as only white spaces were used as word separators in the CS subtask. We have opted to use the data verbatim in this study, but recomputing GFRC2 by treating this nugget as a three-word nugget is trivial.

¹²Note that the he/she/other annotations were obtained at the FairWeb-2 task solely based on whether "he" or "she" was found within the researchers biography text, or not.



Figure 2: The output of COPWA-CS-QD-MN-2 for Topic R112, with additional information provided in baloons (Duplicated from Tao et al. [35]).

<R112> U:Please list researchers who have published at least one paper at the ICLR conference. Format: researcher name, google scholar URL in plain text S:Here are a few researchers who have published papers at the ICLR conferen WC=37-39, L2, h=241 (G4), he (G1 Yoshua Bengio https://scholar.google.com/citations?user=kukA0LcAAAAJ Geoffrey Hinton https://scholar.google.com/citations?user=JicYPdAAAAAJ WC=40-42, L2, h=187 (G4), he (G1) Ian Goodfellow https://scholar.google.com/citations?user=iYN86KEAAAAJ WC=43-45, L2, h=94 (G4), he (G1) Sergey Levine https://scholar.google.com/citations?user=9zK1SCkAAAAJ Chelsea Finn https://scholar.google.com/citations?user=7xVZ2RIAAAAJ These researchers have made significant contributions to the field of machine learning and have been active participants in the ICLR community. If you need more names or specific papers, feel free to ask! U:More please. S:Sure! Here are more researchers who have published papers at the ICLR conference: Pieter Abbeel https://scholar.google.com/citations?user=vtwH6uAAAAJ WC=103-105, L2, h=150 (G4), he (G1) Andrew Ng https://scholar.google.com/citations?user=mG4imMEAAAAJ WC=106-108, L2, h=158 (G4), she (G2) Fei-Fei Li https://scholar.google.com/citations?user=rDfyQnIAAAAJ Jürgen Schmidhuber https://scholar.google.com/citations?user=VTZqWIAAAAAJ Daphne Koller https://scholar.google.com/citations?user=5L5s7cwAAAAJ If you need even more names or specific details about their contributions, let me know! </R112>

Figure 3: The output of ORG-CS-D-MN-1 for Topic R112, with additional information provided in baloons (Duplicated from Tao et al. [35]).

vector (0,0,0,1) with the gold uniform distribution in terms of RNOD [26], *DistrSim* = 1 - 0.595119 = 0.404881;

• The *Experience* is simply the average of the GNP and the two DistrSim scores: (0.114286 + 0.540852 + 0.404881)/3 = 0.353340.

Table 3 shows the GFRC2 scores obtained as the sum of *Experience*'s divided by L = 1,000, as well as the EGNP and EGF scores obtained as the sum of GNP (or DistrSim) scores divided by L = 1,000. It can be observed that COPWA outperforms ORG in terms of every measure for this topic.

Figures 4 and 5 visualise how the GNP and GF scores change as the word count increases in each run. We believe this visualisation approach is useful, as we can see how the *utility* and *group fairness* changes as the conversation progresses. For example, from Figure 4, it is easy to see that, for COPWA,

- The *utility* improves at the third user cluster and then remains more or less constant after that;
- The group fairness for PRONOUN improves at the fourth user cluster and then remains more or less constant after that;
- The *group fairness* for HINDEX is unchanged throughout the conversation.

Moreover, by comparing Figures 4 and 5, it can be observed that the ORG run returns its fifth relevant nugget (which triggers the fifth user cluster to abandon the conversation) later than COPWA's sixth relevant nugget: the word count of the ORG nugget is 108 while that for the COPWA nugget is 87. The PRONOUN-based group fairness of the ORG run finally improves at the fifth user cluster.

For completeness, Tables 4 and 5 show how the original GFRC (but using Eq. 6 instead of Eq. 7 for the R scores) are computed. The

				GF (PRONOUN, JSD)		GF (HINDEX, RNOD)		
wc	GWCrel	WCnonrel	GNP	D	DistrSim	D	DistrSim	Experience
33	4	31	0.114286	(1,0,0)	0.540852	(0,0,0,1)	0.404881	0.353340
36	10	31	0.243902	(1,0,0)	0.540852	(0,0,0,1)	0.404881	0.396545
39	16	31	0.340426	(1,0,0)	0.540852	(0,0,0,1)	0.404881	0.428720
54	22	43	0.338462	(3/4,1/4,0)	0.769708	(0,0,0,1)	0.404881	0.504350
63	28	49	0.363636	(4/5, 1/5, 0)	0.749772	(0,0,0,1)	0.404881	0.506096
87	34	70	0.326923	(5/6,1/6,0)	0.733061	(0,0,0,1)	0.404881	0.488288

Table 1: Steps for computing GFRC2 for COPWA-CS-QD-MN-2.

Table 2: Steps for computing GFRC2 for ORG-CS-D-MN-1.

				GF (PRONOUN, JSD)			GF (HINDEX, RNOD)	
wc	GWCrel	WCnonrel	GNP	D	DistrSim	D	DistrSim	Experience
39	6	36	0.142857	(1,0,0)	0.540852	(0,0,0,1)	0.404881	0.362863
42	12	36	0.250000	(1,0,0)	0.540852	(0,0,0,1)	0.404881	0.398578
45	18	36	0.333333	(1,0,0)	0.540852	(0,0,0,1)	0.404881	0.426355
105	24	93	0.205128	(1,0,0)	0.540852	(0,0,0,1)	0.404881	0.383620
108	30	93	0.243902	(4/5,1/5,0)	0.749772	(0,0,0,1)	0.404881	0.466185

Table 3: GFRC2: EGNP and EGF scores of COPWA-CS-QD-MN-2 and ORG-CS-D-MN-1 for Topic R112.

	EGNP	EGF (PRONOUN, JSD)	EGF (HINDEX, RNOD)	GFRC2
COPWA-CS-QD-MN-2	0.001728	0.003875	0.002429	0.002677
ORG-CS-D-MN-1	0.001175	0.002913	0.002024	0.002038

pw's in our table are larger than those in the FairWeb-2 overview paper by only 0.001 for the reason discussed in Section 2.2. As every relevant nugget is L2-relevant, the gain value is always $g = (2^2 - 1)/l_{\text{max}} = 3/4 = 0.75$. Compared to Table 3, this table and the similar table from the FairWeb-2 overview paper share a minor inconvenience when averaging to compute GFRC: the R scores are far smaller than the GF scores and therefore the R scores have very little impact on the final GFRC scores. Our GFRC2 framework does not have this problem thanks to the way we formulate the measure as the expected total user experience.

5 Conclusions and Future Work

By directly transferring the general idea of the GFR (Group Fairness and Relevance) framework for web search [30] to the task of evaluating generated conversations, we formulated GFRC2 as a form of expected user experience for a population of users who go through the words within the conversation. Using real runs from the CS subtask of the NTCIR-18 FairWeb-2 task [35], we demonstrated that this formulation also lets us visualise the Relevance and Group Fairness component scores for each cluster of users who are assumed to abandon the conversation at a particular relevant nugget. In addition, GFRC2 is free from the minor scaling issue of GFRC (i.e., R scores are generally far smaller than the GF scores).

Applying GFRC2 to a large scale (simulated or real) conversation evaluation is left for future work.¹³ The idea of defining the total user experience based on *utility* and *group fairness* could be extended to other axes for conversational search (e.g., *harmlessness* and *modesty*) [28].

Acknowledgement

We thank the NTCIR-18 FairWeb-2 task organisers and participants for enabling us to conduct a pilot study with their data.

References

- Zahra Abbasiantaeb, Simon Lupart, Leif Azzopardi, Jeffery Dalton, and Mohammad Aliannejadi. 2025. Conversational Gold: Evaluating Personalized Conversational Search System using Gold Nuggets. arXiv:2503.09902 [cs.IR] https://arxiv.org/abs/2503.09902
- [2] Mohammad Aliannejadi, Zahra Abbasiantaeb, Simon Lupart, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. 2025. TREC iKAT 2024: The Interactive Knowledge Assistance Track Overview. In *Proceedings of TREC 2024*. https: //trec.nist.gov/pubs/trec33/papers/Overview_ikat.pdf
- [3] Negar Arabzadeh and Charles L. A. Clarke. 2024. A Comparison of Methods for Evaluating Generative IR. arXiv:2404.04044 [cs.IR] https://arxiv.org/abs/2404. 04044
- [4] Nuo Chen, Jiqun Liu, and Tetsuya Sakai. 2023. A Reference-Dependent Model for Web Search Evaluation: Understanding and Measuring the Experience of Boundedly Rational Users. In Proceedings of the ACM Web Conference 2023 (Austin, TX, USA) (WWW '23). Association for Computing Machinery, New York, NY, USA, 3396–3405. https://doi.org/10.1145/3543507.3583551
- [5] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation (SIGIR '08). Association for Computing Machinery, New York, NY, USA, 659–666. https://doi.org/10.1145/1390334.1390446
- [6] Laura Dietz. 2024. A Workbench for Autograding Retrieve/Generate Systems. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (Washington DC, USA) (SIGIR '24). Association for Computing Machinery, New York, NY, USA, 1963–1972. https://doi.org/10.1145/3626772.3657871
- [7] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. Fairness in Information Access Systems. *Found. Trends Inf. Retr.* 16, 1–2 (July 2022), 1–177. https://doi.org/10.1561/1500000079
- [8] Michael D Ekstrand, Graham McDonald, and Amifa Raj. 2022. Overview of the TREC 2021 Fair Ranking Track. In *The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings*. NIST. https://trec.nist.gov/pubs/trec30/papers/Overview-F.pdf
- [9] Matthew Ekstrand-Abueg, Virgil Pavlu, Makoto Kato, Tetsuya Sakai, Takehiro Yamamoto, and Mayu Iwata. 2013. Exploring semi-automatic nugget extraction for Japanese one click access evaluation. In *Proceedings of the 36th International*

¹³At the NTCIR-18 FairWeb-2 task, only one team participated in the CS subtask (besides the organiser team), and the task will be discontinued, unfortunately.



Figure 4: Visualising the effectiveness of COPWA-CS-QD-MN-2 based on the GFRC2 framework.



Figure 5: Visualising the effectiveness of ORG-CS-D-MN-1 based on the GFRC2 framework.

ACM SIGIR Conference on Research and Development in Information Retrieval (Dublin, Ireland) (SIGIR '13). Association for Computing Machinery, New York, NY, USA, 749–752. https://doi.org/10.1145/2484028.2484153

- [10] Emma J. Gerritse, Faegheh Hasibi, and Arjen P. de Vries. 2020. Bias in Conversational Search: The Double-Edged Sword of the Personalized Knowledge Graph. In Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval (Virtual Event, Norway) (ICTIR '20). Association for Computing Machinery, New York, NY, USA, 133–136. https://doi.org/10.1145/3409256.3409834
- [11] Lukas Gienapp, Harrisen Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, Matthias Hagen, and Martin Potthast. 2024. Evaluating Generative Ad Hoc Information Retrieval. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (Washington DC, USA) (SIGIR '24). Association for Computing Machinery, New York, NY, USA, 1916–1929. https://doi.org/10.1145/3626772.3657849
- [12] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. 20, 4 (Oct. 2002), 422–446. https://doi. org/10.1145/582415.582418
- [13] Johannes Kiesel, Damiano Spina, Henning Wachsmuth, and Benno Stein. 2021. The Meant, the Said, and the Understood: Conversational Argument Search and Cognitive Biases. In Proceedings of the 3rd Conference on Conversational User Interfaces (Bibao (online), Spain) (CUI '21). Association for Computing Machinery, New York, NY, USA, Article 20, 5 pages. https://doi.org/10.1145/3469595.3469615
- [14] Jimmy Lin and Dina Demner-Fushman. 2006. Will Pyramids Built of Nuggets Topple Over?. In Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, Robert C. Moore, Jeff Bilmes, Jennifer Chu-Carroll, and Mark Sanderson (Eds.). Association for Computational Linguistics, New York City, USA, 383–390. https://aclanthology.org/N06-1049/
- [15] Aldo Lipani, Ben Carterette, and Emine Yilmaz. 2021. How Am I Doing?: Evaluating Conversational Search Systems Offline. ACM Trans. Inf. Syst. 39, 4, Article 51 (Aug. 2021), 22 pages. https://doi.org/10.1145/3451160
- [16] Qin Liu, Xuan Feng, Tianlong Gu, and Xiaoli Liu. 2024. FairCRS: Towards Useroriented Fairness in Conversational Recommendation Systems. In Proceedings

of the 18th ACM Conference on Recommender Systems (Bari, Italy) (RecSys '24). Association for Computing Machinery, New York, NY, USA, 126–136. https://doi.org/10.1145/3640457.3688150

- [17] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2024. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. arXiv:2308.05374 [cs.AI] https://arxiv.org/abs/2308.05374
- [18] Zeyang Liu, Ke Zhou, and Max L. Wilson. 2021. Meta-evaluation of Conversational Search Evaluation Metrics. ACM Trans. Inf. Syst. 39, 4, Article 52 (Sept. 2021), 42 pages. https://doi.org/10.1145/3445029
- [19] Teruko Mitamura, Hideki Shima, Tetsuya Sakai, Noriko Kando, Tatsunori Mori, Koichi Takeda, Chin-Yew Lin, Ruihua Song, Chuan-Jie Lin, and Cheng-Wei Lee. 2010. Overview of the NTCIR-8 ACLIA Tasks: Advanced Cross-Lingual Information Access. In *Proceedings of NTCIR-8*. 15–24.
- [20] Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. 2024. A Survey of Conversational Search. arXiv:2410.15576 [cs.CL] https://arxiv.org/abs/2410.15576
- [21] Alistair Moffat, Joel Mackenzie, Paul Thomas, and Leif Azzopardi. 2022. A Flexible Framework for Offline Effectiveness Metrics. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (Madrid, Spain) (SIGIR '22). Association for Computing Machinery, New York, NY, USA, 578–587. https://doi.org/10.1145/3477495.3531924
- [22] Gustavo Penha and Claudia Hauff. 2020. Challenges in the Evaluation of Conversational Search Systems. In Proceedings of the KDD 2020 Workshop on Conversational Systems Towards Mainstream Adoption co-located with the 26TH ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD 2020). https://ceur-ws.org/Vol-2666/KDD_Converse20_paper_5.pdf
- [23] Stephen M. Pollock. 1968. Measures for the comparison of information retrieval systems. *American Documentation* 19, 4 (1968), 387–397. https://doi.org/10.1002/asi.5090190406 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.5090190406
- [24] Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Initial Nugget Evaluation Results for the TREC 2024 RAG Track with the AutoNuggetizer Framework. arXiv:2411.09607 [cs.IR]

Table 4: GFRC: Steps for computing the R and GF scores are computed for COPWA-CS-QD-MN-2 and ORG-CS-D-MN-1 for Topic R112. This similar to a table from Tao et al. [35], except that Section (a) of our table shows the results based on Eq. 6 instead of Eq. 7.

	COPWA-CS-QD-MN-2			ORG-CS-D-MN-1		
(a) R	pw	9	pw * g	pw	9	pw * g
	0.968	0.75	0.726000	0.962	0.75	0.721500
	0.965	0.75	0.723750	0.959	0.75	0.719250
	0.962	0.75	0.721500	0.956	0.75	0.717000
	0.947	0.75	0.710250	0.896	0.75	0.672000
	0.938	0.75	0.703500	0.893	0.75	0.669750
	0.914	0.75	0.685500			
(b) GF ^{JSD} (PRONOUN)	turn #	D	DistrSim	turn #	D	DistrSim
	1	(5/6, 1/6, 0)	0.733061	1	(1, 0, 0)	0.540852
				2	(1/2, 1/2, 0)	0.809126
(c) GF ^{RNOD} (HINDEX)	turn #	D	DistrSim	turn #	D	DistrSim
	1	(0, 0, 0, 1)	0.404881	1	(0, 0, 0, 1)	0.404881
				2	(0, 0, 0, 1)	0.404881

Table 5: GFRC: R and GF scores of COPWA-CS-QD-MN-2 and ORG-CS-D-MN-1 for Topic R112. This is similar to a table from Tao et al. [35], except that Section (a) of our table shows the results based on Eq. 6 instead of Eq. 7, and the final GFRC scores are presented in addition.

	R	GF (PRONOUN, JSD)	GF (HINDEX, RNOD)	GFRC
COPWA-CS-QD-MN-2	0.008532	0.733061	0.404881	0.382158
ORG-CS-D-MN-1	0.006992	0.674989	0.404881	0.362287

https://arxiv.org/abs/2411.09607

- [25] Tetsuya Sakai. 2014. Metrics, Statistics, Tests. Springer Berlin Heidelberg, Berlin, Heidelberg, 116–163. https://doi.org/10.1007/978-3-642-54798-0_6
- [26] Tetsuya Sakai. 2021. Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 2759–2769. https://doi.org/10.18653/v1/2021.acl-long.214
- [27] Tetsuya Sakai. 2023. Fairness-based Evaluation of Conversational Search: A Pilot Study. In Proceedings of EVIA 2023. 5–13.
- [28] Tetsuya Sakai. 2023. SWAN: A Generic Framework for Auditing Textual Conversational Systems. arXiv:2305.08290 [cs.IR] https://arxiv.org/abs/2305.08290
- [29] Tetsuya Sakai, Makoto P. Kato, and Young-In Song. 2011. Click the search button and be happy: evaluating direct and immediate information access (CIKM '11). Association for Computing Machinery, New York, NY, USA, 621–630. https: //doi.org/10.1145/2063376.2063669
- [30] Tetsuya Sakai, Jin Young Kim, and Inho Kang. 2023. A Versatile Framework for Evaluating Ranked Lists in Terms of Group Fairness and Relevance. ACM TOIS 42, 1, Article 11 (Aug. 2023), 36 pages. https://doi.org/10.1145/3589763
- [31] Tetsuya Sakai and Stephen Robertson. 2008. Modelling A User Population for Designing Information Retrieval Metrics. In Proceedings of EVIA 2008. 30–41.
- [32] Tetsuya Šakai and Zhaohao Zeng. 2021. Retrieval Evaluation Measures that Agree with Users' SERP Preferences: Traditional, Preference-based, and Diversity Measures. ACM Trans. Inf. Syst. 39, 2, Article 14 (Dec. 2021), 35 pages. https: //doi.org/10.1145/3431813
- [33] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2022. Evaluating Mixed-initiative Conversational Search Systems via User Simulation. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (Virtual Event, AZ, USA) (WSDM '22). Association for Computing Machinery, New York, NY, USA, 888–896. https://doi.org/10.1145/3488560.3498440
- [34] Ivan Sekulić, Mohammad Alinannejadi, and Fabio Crestani. 2024. Analysing Utterances in LLM-Based User Simulation for Conversational Search. 15, 3, Article 62 (May 2024), 22 pages. https://doi.org/10.1145/3650041
- [35] Sijie Tao, Tetsuya Sakai, Junjie Wang, Hanpei Fang, Yuxiang Zhang, Haitao Li, Yiteng Tu, Nuo Chen, and Maria Maistro. 2025. Overview of the NTCIR-18 FairWeb-2 Task. In Proceedings of NTCIR-18. to appear.
- [36] Ellen M. Voorhees. 2004. Overview of the TREC 2003 Question Answering Track. In Proceedings of TREC 2003.

- [37] Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10052–10065. https://doi.org/10.18653/v1/2023.emnlp-main.621
- [38] Shuo Zhang and Krisztian Balog. 2020. Evaluating Conversational Recommender Systems via User Simulation. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (Virtual Event, CA, USA) (KDD '20). Association for Computing Machinery, New York, NY, USA, 1512–1520. https://doi.org/10.1145/3394486.3403202