Evaluating the Cranfield Paradigm for Conversational Search Systems

Xiao Fu
University College London
London, UK
xiao.fu.20@ucl.ac.uk

Emine Yilmaz
University College London & Amazon
London, UK
emine.yilmaz@ucl.ac.uk

Aldo Lipani
University College London
London, UK
aldo.lipani@ucl.ac.uk

ABSTRACT

Due to the sequential and interactive nature of conversations, the application of traditional Information Retrieval (IR) methods like the Cranfield paradigm require stronger assumptions. When building a test collection for Ad Hoc search, it is fair to assume that the relevance judgments provided by an annotator correlate well with the relevance judgments perceived by an actual user of the search engine. However, when building a test collection for conversational search, we do not know if it is fair to assume that the relevance judgments provided by an annotator correlate well with the relevance judgments perceived by an actual user of the conversational search system. In this paper, we perform a crowdsourcing study to evaluate the applicability of the Cranfield paradigm to conversational search systems. Our main aim is to understand what is the agreement in terms of user satisfaction between the users performing a search task in a conversational search system (i.e., directly assessing the system) and the users observing the search task being performed (i.e., indirectly assessing the system). The result of this study is paramount because it underpins and guides 1) the development of more realistic user models and simulators, and 2) the design of more reliable and robust evaluation measures for conversational search systems. Our results show that there is a fair agreement between direct and indirect assessments in terms of user satisfaction and that these two kinds of assessments share similar conversational patterns. Indeed, by collecting relevance assessments for each system utterance, we tested several conversational patterns that show a promising ability to predict user satisfaction.

KEYWORDS

dialogue systems, evaluation, relevance, satisfaction

ACM Reference Format:
Xiao Fu, Emine Yilmaz, and Aldo Lipani. 2022. Evaluating the Cranfield Paradigm for Conversational Search Systems. In *Proceedings of the 2022 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '22), July 11–12, 2022, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3539813.3545126

ICTIR '22, July 11–12, 2022, Madrid, Spain

ACM ISBN 978-1-4503-9412-3/22/07…$15.00
https://doi.org/10.1145/3539813.3545126

1 INTRODUCTION

Traditionally search systems are evaluated offline by building test collections. Test collections consist of a collection of documents, a set of topics, and a set of relevance assessments indicating to which topic a document is relevant. Normally, in either industrial settings or in large evaluation campaigns like TREC, NTCIR, CLEF, etc., topics are defined as a sample of a search log, while relevance assessments are collected by assessors who, given a topic, assess the relevance of the documents returned by one or several search systems for the given topic. This assessment exercise is useful if two assumptions are met. The first assumption is that the relevance indicated by the assessors is fairly correlated to the relevance perceived by the user of the search system. The second assumption is that the score provided by a user-based evaluation measure applied to the assessed search results fairly correlates to user satisfaction. Several works in Information Retrieval (IR) have demonstrated that such assumptions are fair.

However, the offline evaluation of conversational search systems (CSSs) is more challenging [28]. Although we can decompose a conversation into turns and evaluate each turn independently, this assumption disregards the fact that a turn's relevance may be dependent on what happened in the previous turns. In other words, this assumption disregards the *context* of the conversation. Moreover, it is unclear whether the satisfaction perceived by a user interacting with a CSS and an assessor reading the user's conversation log correlates. Understanding how much these assumptions hold will allow the extension of the Cranfield paradigm to CSSs. This is important also in order to inform the development of better user models [27, 28] and user simulators [21], therefore better evaluation measures and training procedures.

In this paper, we perform a crowdsourcing study to compare the online and offline evaluation of a CSS. For the online evaluation, we reuse the conversation logs collected by Lipani et al. [28]. For the offline evaluation, we perform the crowdsourcing study and collect new assessments. The goal of this paper is to study the differences between the direct assessment (online evaluation) and the indirect assessment (offline evaluation) of a CSS. We then analyze the agreement between these two different types of assessment, which include both turn relevance and conversation satisfaction. Finally, we explore several conversational patterns to determine whether we can predict conversation satisfaction from turn relevance.

Our contributions can be summarized as follows:

- The design of a crowdsourcing study to collect direct and indirect assessments.
- An analysis of the agreement of conversation satisfaction between the direct and indirect assessments.

- An analysis of the agreement of turn relevance between the direct and indirect assessments.
- An analysis of several conversational patterns to predict conversation satisfaction from turn relevance assessments.

2 RELATED WORK

The evaluation of CSSs is still an open problem [7, 25].

2.1 Conversational Search

CSSs are at the intersection of search systems, chatbots, and dialogue systems. CSSs share similar goals to such systems like the need to access information, interactivity, statefulness, and interaction naturalness [4]. Several studies focus on user interfaces of CSSs, to improve their user experience [3, 6]. Others take advantage from other areas, such as knowledge graphs [10, 15] and neural networks [4, 36, 37] to improve their performance.

Besides these topics, however, the evaluation of CSSs is still relatively undeveloped [7, 19, 28]. Thought CSSs include many functional extensions from existing information retrieval systems [4], some studies still use traditional metrics, such as MAP, nDCG and MRR to evaluate these systems [9, 19, 24]. Metrics from other domains, such as ROUGE and BLEU scores, are also popular choices [32, 35]. Recent studies suggest that however without real users' interaction, these metrics cannot reflect users' satisfaction [25, 29].

2.2 Online and Offline Evaluation

The offline evaluation of search systems consists of their testing on a test collection via a metric. These test collections normally consist of a set of documents, a set of queries, and a set of relevance judgments determining which document is and is not relevant to which query. There are several metrics for evaluating search systems [17]. The offline evaluation sometimes referred to as the Cranfield paradigm, faces many challenges. Most notably, normally queries and relevance judgments are performed by different individuals generating noisy labels. This happens for example when queries are sampled from search logs and expert assessors have no way to know the original users' contexts and intents [19]. Moreover, the complexity of users' queries and the connection between users' contexts are often ignored by classic metrics and require a more holistic assessment [13, 18]. This problem is in fact overcome by the online evaluation of search systems, which consists of exploiting the behavioral feedback of users when interacting with search systems to evaluate them. Fox et al. [11] discovered an association between user satisfaction and user interest. Then, Huffman and Hochster [16] found that user satisfaction can also be predicted by patterns in the search results, where relevant documents retrieved first are important predictors of user satisfaction. Like these previous researchers, several others also focused on the behavior that users have when interacting with search systems [1, 11, 12, 14, 16, 19].

2.3 Evaluating User Satisfaction

User satisfaction is a highly abstracted subjective attitude towards the experience of search and the interactions with a search system [7, 24]. It can be defined as the fulfillment of users who are pursuing their goals [20]. To understand this concept, many studies exist [2, 18, 22–25, 39]. Yilmaz et al. [39] provide many metrics that reflect user satisfaction. Some studies analyze user satisfaction with CSSs by breaking the conversation into many query-level satisfactions [22, 23]. However, some other studies claim that the overall user satisfaction cannot be considered as the sum of query-level satisfactions [18].

Järvelin et al. [18] concludes that query-level metrics do not capture the user information journey by not considering the dependency between queries. In fact, for a traditional search system, a wrong result may be lethal, but in a conversational search system, users can ask follow-up questions to seek better answers [2].

In CSSs, measuring user satisfaction is still an open problem. Many studies are performed by collecting the satisfaction feedback from users [24, 25]. This, however, requires the running of a system online, having a user base, and finding ways to encourage users to provide feedback. Another way to predict user satisfaction is by modeling it using machine learning [7, 21].

3 RESEARCH QUESTIONS

In this paper we aim to answer the following research questions:

RQ1. *What is the agreement between the direct and indirect assessments of conversation satisfaction?*

To answer this question, we need to quantify the agreement in terms of user satisfaction between a user performing the search task (direct assessment) and a user observing the search task (indirect assessment). This is fundamental in order to understand if we can rely on indirect assessments in order to evaluate simulated conversations.

RQ2. *How does the agreement between direct and indirect assessments change when we follow a standard Cranfield paradigm but in a conversational setting?*

In a standard Cranfield paradigm, indirect assessments are used to evaluate whether a document is relevant to a given topic. In a conversational setting, we can do the same by assessing the relevance of the returned document at each turn of the conversation. This is fundamental in order to understand if the relevance perceived by the users and assessors is congruent.

RQ3. *Can we predict conversation satisfaction from conversational patterns?*

In this analysis, we aim to analyze several conversational patterns to predict conversation satisfaction. For example, are users satisfied when the results are somewhat relevant across the whole conversation? Or, users are more satisfied if the last few turns are relevant? These patterns are fundamental in order to guide the design of more reliable metrics.

4 THE DIRECT AND INDIRECT ASSESSMENTS DATA

In this study, we use the dataset collected by Lipani et al. [28]. This dataset consists of conversation logs over 11 topics defined based on the SQuAD dataset [33]. The documents from SQuAD are divided into paragraphs, and each paragraph is labeled with subtopics. Participants were given an interface to query and read the returned paragraphs. They were asked to mark if the returned paragraph was relevant to their query and which subtopic it belonged to at each turn. When the participants wanted to end the conversation, they needed to also assess whether they had been satisfied or not with

Table 1: Statistics about the two datasets.

Dataset	Users	Conv.	Topics	$\overline{\text{Turns}}$	$P(rel)$	$P(sat)$
DI	133	160	11	5.4	0.71	0.74
IN	119	353	11	6.6	0.70	0.86

the conversation. These conversation logs represent the dataset we use for the direct assessments.

In order to study the agreement between direct and indirect assessments, we performed a crowdsourcing task in order to collect the latter. Crowdworkers were asked to read the conversation logs and mark for each turn whether the returned paragraph was relevant or not to the query and determine at the end whether the user would have been satisfied or not. For quality control, for each requested feedback, we also required the provision of a supporting claim. In this study, the indirect assessors are the crowdworkers of this study while the direct assessors are the participants in the crowdsourcing task designed by Lipani et al. [28].

This crowdsourcing task was performed on Amazon Mechanical Turk, where 119 crowdworkers were recruited. Crowdworkers, to be selected, had to satisfy the following 3 conditions: 1) They had performed more than 500 previous tasks; 2) They had more than 90% accepting rate, and; 3) They were English speakers. From the 160 conversation logs collected by Lipani et al. [28], we let at least 3 crowdworkers annotate each log. This task consisted of the following parts:

- **Introduction**: A brief describing the task – "For each turn, mark each system's response as relevant or not relevant to the user's query, and finally determine if the user would have been satisfied with the conversation."
- **Context**: The topic that users (direct assessors) used to generate the conversation log.
- **Conversation**: The conversation log, with an input form for each turn where annotators can provide the relevance of the returned response to the user query and supporting claims.
- **Satisfaction**: An input form at the end of the conversation to let the annotators determine whether the conversation would have been satisfactory or not to the original users.

We manually inspected each conversation to determine if the workers did the task correctly or not. The conversations that did not pass a condition set based on the time the task took to perform it and a condition based on the quality of the supporting claims, were filtered. This generated 358 indirect assessments for a cost of around 160 $. In Table 1, we show the statistics of the two datasets. Here we observe that the estimated probability of a system response being relevant is similar across the two datasets, while the estimated probability of a conversation being satisfactory is higher for the indirect assessments.

5 EXPERIMENTS

To answer the first two research questions, we use three coefficients of agreement. These coefficients are Krippendorff's alpha (α), Randolph's kappa (κ_r), and Cohen's kappa (κ_c). The Krippendorff's alpha [26] is an inter-rater reliability measure that calculates disagreement as opposed to an agreement like for the case of Randolph's kappa and Cohen's kappa. Krippendorff's alpha ranges from 0 to 1, indicating perfect agreement when 1 and perfect disagreement when 0. Cohen's kappa [8] is another inter-rater reliability measure that calculates the agreement relative to an agreement achieved by chance using the observed marginal probabilities of the categories to be annotated. However, due to the distribution imbalance observed in our dataset, this metric tends to produce low agreement scores. The Randolph's kappa [34] is a similar inter-rater reliability measure that alleviates the imbalance issue by assuming a uniform distribution across the categories to be annotated [38]. Cohen's and Randolph's kappas range from -1 to 1 indicating perfect agreement when 1, no agreement based on the one assumed by the distribution set when 0, and worse than this last agreement when negative.

Table 2: Satisfaction agreement coefficients: α refers to the Krippendorff's alpha, κ_r refers to the Randolph's kappa, and κ_c refers to the Cohen's kappa.

Type	Grouping	α	κ_r	κ_c
Overall	intra-DI	0.371	0.259	0.034
	intra-IN	0.672	0.522	0.021
	DI vs. IN	0.400	0.269	0.079
Per Topic	intra-DI	0.317	0.155	0.021
	intra-IN	0.671	0.496	0.000
	DI vs. IN	0.450	0.278	0.062

To answer the last research question, we use three performance evaluation measures: Precision, Recall, and the F1-score, and three correlation coefficients: Spearman's rank (ρ_s), Pearson rho (ρ), and Kendall's tau (τ). These correlation coefficients are used to assess the strength of the relationship between two variables. They range from -1 and 1 indicating a strong positive (or negative) relationship when 1 (or -1), and no relationship when 0.

5.1 User Satisfaction (RQ1)

We calculate the agreement on satisfaction between the raters performing the direct assessments (intra-DI), the raters performing the indirect assessment (intra-IN), and across direct and indirect raters (DI vs. IN). This agreement is calculated in two ways, *overall* and *per topic*. The former calculates the agreements by first grouping all the ratings in a pool, while the latter calculates the agreement by grouping per topic and then computing the average across topics.

In Table 2, we show the agreement on conversation satisfaction. The results show that there is a certain agreement among direct assessors and a stronger agreement among indirect assessors. This indicates that users experiencing the CSS agree much less than users observing the system being used in a conversation log. We also observe that the agreement between direct assessors and indirect assessors is higher than the direct assessors alone. This indicates that the indirect assessment procedure is capable of capturing fair satisfaction scores. We also observe that there is no difference in the trends observed between the two types of agreement computation, overall and per topic, and the 3 coefficients.

Table 3: Relevance agreement coefficients. α refers to the Krippendorff's alpha, κ_r refers to the Randolph's kappa, and κ_c refers to the Cohen's kappa.

Type	Grouping	α	κ_r	κ_c
Overall	intra-IN	0.167	0.205	0.116
	DI vs. IN	0.204	0.298	0.192
Per Topic	intra-IN	0.309	0.238	0.052
	DI vs. IN	0.318	0.327	0.139

5.2 Relevance (RQ2)

We calculate the agreement on relevance following the same protocol presented in the previous section. However, in this case, we cannot compute the agreement among the raters performing the direct assessments since their assessments are unique.

In Table 3, we show the agreement on turn relevance. The results show a certain agreement among indirect assessors and a stronger agreement when comparing them to the direct assessors. This trend is observed in both types, overall and per topic, and all 3 coefficients. This behavior is somehow inverted with respect to the one observed for conversation satisfaction (in Table 2), where the agreement among the indirect assessors was stronger than when comparing direct and indirect assessors. This indicates a difference between judging relevance and satisfaction, which depends on how the CSS is being experienced, directly or indirectly. Moreover, comparing these two tables, we also observe that the degree of agreement for conversation satisfaction is higher than the one for turn relevance, indicating that assessors agree more about conversation satisfaction than turn relevance.

5.3 Predicting Satisfaction (RQ3)

To answer **RQ3**, we will analyze several predictors of conversation satisfaction. These predictors will be based on a combination of statistics and turn relevance scores. We will first analyze their correlation individually, then we will train a classifier based on logistic regression to study the importance of each conversational predictor. The analyzed predictors are:

(1) The **number of turns** (ℓ) of a conversation. This predictor is indicative of the effort the user made to complete the task or is indicative of the user's patience when the system failed.
(2) The relevance of the **last reply** (j_ℓ). This predictor may capture the *recency effect*, which is a cognitive bias that makes remembering more clearly the last replies.
(3) The position starting from the end of the **last non-relevant reply** (i). This predictor is similar to the one above but focuses on when the negative experience happened.
(4) Whether the conversation had **at least two non-relevant replies** in sequence (ii). This predictor may capture a sign of user frustration in the conversation.
(5) The **ratio of non relevant replies** ($P(\overline{rel})$). This predictor may capture another sign of user frustration in the conversation.
(6) A normalized version of Rank-Biased Precision (**nRBP**) [5, 30, 31]. This predictor weights more the relevant replies happening at the beginning of the conversation. This predictor may capture the *anchoring effect*, which is a cognitive bias that influences the assessments of replies based on the assessments previously made. The p of nRBP is 0.8.
(7) The inverse normalized version of Rank-Biased Precision (**nRBP**$^{-1}$). This is like the previous predictor but computed from the last turn. This predictor weights more the relevant replies happening at the end of the conversation. This is another predictor that may capture the recency effect.

In Table 4, we show the correlation results of each predictor against the conversation satisfaction. We observe that the strongest predictors are: the ratio of non-relevant replies, the inverse nRBP, the presence of at least two non-relevant replies, and the nRBP. The first and third predictors affect user satisfaction negatively, while the second and fourth predictors affect user satisfaction positively. Moreover, we observe that the patterns are almost all consistent across the two types of assessments, direct and indirect, except length, which seems to play a more critical role in the direct assessments.

In Table 5, we show the performance of two logistic regression models trained using the predictors above. The first classifier is trained using 80% of direct assessments (DI) and tested on the remaining 20% and the indirect assessments (IN). The second classifier is trained using 80% of IN and tested on the remaining 20% and the DI. The results show that all models achieve acceptable performance and that the direct assessments are better than the indirect assessments in predicting satisfaction. Moreover, we observe that the performance of the model trained and tested on IN (i.e., last row) is higher than when the model is trained and tested on DI, (i.e., first row) on every metric. This may be due to the previously observed larger agreement on conversation satisfaction across IN than DI. In Table 6, we show the weights of these models. Here, we observe that the most important feature is the length of the conversation. However, this shows an opposite behavior across the two models, indicating that the length oppositely influences assessors. Direct (indirect) assessors are negatively (positively) affected. The other most important features are the ratio of non-relevant documents, nRBP, and the presence of at least two non-relevant replies. These behave similarly across the two models.

6 DISCUSSION AND CONCLUSION

In this paper, we designed a crowdsourcing study to collect offline assessments of existing conversations assessed online. We then studied them and analyzed the difference between indirect and direct assessments in CCSs. In this study, we found that there is a stronger agreement when evaluating conversation satisfaction rather than turn relevance. However, we achieve a fair agreement across the two types of assessments in both cases. Moreover, analyzing the conversational patterns to predict conversation satisfaction, we found that the length of the conversation has a contrasting behavior across the two types of assessments, negatively influencing direct assessors while positively influencing indirect ones. Also, we observe that, as expected, conversation satisfaction depends on the ratio of non-relevant replies returned by the system, the presence of subsequent non-relevant replies, and the initial performance of the system (high nRBP). Indicating the presence of a degree of recency

Table 4: Correlation coefficient of the predictors against the satisfaction assessments for both direct (DI) and indirect (IN). ρ_s refers to the Spearman's rank; ρ refers to the Pearson's rho; and τ refers to the Kendall's tau. In brackets, we report the rank of the degree of correlation in absolute value.

	DI						IN					
	ρ_s		ρ		τ		ρ_s		ρ		τ	
ℓ	-0.374	(7)	-0.413	(5)	-0.321	(7)	0.071	(7)	0.066	(7)	0.061	(7)
i	0.382	(6)	0.328	(7)	0.339	(6)	0.263	(5)	0.227	(6)	0.231	(6)
j_ℓ	0.407	(5)	0.407	(6)	0.407	(5)	0.251	(6)	0.251	(5)	0.251	(5)
ii	-0.463	(4)	-0.463	(4)	-0.463	(1)	-0.360	(1)	-0.360	(4)	-0.360	(1)
$P(\overline{rel})$	-0.526	(2)	-0.556	(1)	-0.457	(3)	-0.346	(2)	-0.389	(1)	-0.299	(2)
nRBP	0.495	(3)	0.521	(3)	0.424	(4)	0.333	(4)	0.386	(2)	0.285	(4)
nRBP^{-1}	0.535	(1)	0.556	(1)	0.458	(2)	0.336	(3)	0.374	(3)	0.288	(3)

Table 5: Performance of the two logistic regression models. The first is trained on 80% of direct assessments (DI) and tested on the remaining 20% of DI and 100% of the indirect assessments (IN). The second is trained on 80% of IN and tested on the 20% of IN and 100% of DI.

Training	Testing	Precision	Recall	F1	AUC
DI	DI	0.889	0.923	0.905	0.983
	IN	0.932	0.813	0.868	0.932
IN	DI	0.748	1.000	0.856	0.894
	IN	0.928	0.984	0.955	0.994

Table 6: Weights and ranks of the two logistic regression models. One trained on direct assessments (DI) and one trained on indirect assessments (IN).

	DI-model		IN-model	
ℓ	-1.202	(1)	1.69	(1)
j_ℓ	0.264	(7)	0.274	(7)
i	0.435	(6)	0.483	(5)
ii	-1.047	(3)	-0.53	(4)
$P(\overline{rel})$	-0.65	(5)	-1.118	(2)
nRBP	1.199	(2)	0.681	(3)
nRBP^{-1}	0.886	(4)	0.403	(6)
1 (bias)	0.138		1.085	

effect: users' satisfaction scores are influenced negatively by the presence of non-relevant results, and a degree of anchoring effect: users' satisfaction scores are influenced by their initial relevant results.

In this study, we identified two limitations, whose investigation is left to future work. The first limitation is about the effect of user effort on conversation satisfaction. In our logs, we observed that indirect assessors spent less time than direct assessors in completing the crowdsourcing task. The average duration of an indirect assessment is 248s, while the average duration of a direct assessment is 399s. Kiseleva et al. [25] showed that more effort significantly negatively impacts user satisfaction and the difference in average duration may have impacted our results. This hypothesis is also corroborated by our initial observation (in Table 1) where we noted a difference in the probability of raters assessing conversation satisfaction. The second limitation is about the way we presented the conversation logs to the indirect assessors. In our crowdsourcing task, we presented the conversation as a whole. This together with the difference between how direct and indirect assessments are generated, that is, direct assessors can decide when to end the conversation while indirect assessors cannot, may have caused the large observed discrepancy in how the length of the conversation is perceived by the two kinds of assessors. In future work, we aim to investigate other ways of generating indirect assessments in order to mitigate the influence of this length bias on conversation satisfaction.

ACKNOWLEDGEMENTS

This project was funded by the EPSRC Fellowship titled "Task Based Information Retrieval" and grant reference number EP/P024289/1.

REFERENCES

[1] Mikhail Ageev, Qi Guo, Dmitry Lagun, and Eugene Agichtein. 2011. Find It If You Can: A Game for Modeling Different Types of Web Search Success Using Interaction Data. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2009916.2009965

[2] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 2007. The Relationship between IR Effectiveness Measures and User Satisfaction. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Amsterdam, The Netherlands) *(SIGIR '07)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1277741.1277902

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3290605.3300233

[4] Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Conversational Search (Dagstuhl Seminar 19461). *Dagstuhl Reports* 9, 11 (2020). https://doi.org/10.4230/DagRep.9.11.34

[5] Ben Carterette. 2011. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2009916.2010037

[6] Justine Cassell. 2001. Embodied Conversational Agents: Representation and Intelligence in User Interfaces. *AI Magazine* 22, 4 (2001). https://doi.org/10.1609/aimag.v22i4.1593

[7] Jason Ingyu Choi, Ali Ahmadvand, and Eugene Agichtein. 2019. Offline and Online Satisfaction Prediction in Open-Domain Conversational Systems. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3357384.3358047

[8] Jacob Cohen;. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960). https://doi.org/10.1177/001316446002000104

[9] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The Conversational Assistance Track Overview.

[10] Yue Feng, Aldo Lipani, Fanghua Ye, Qiang Zhang, and Emine Yilmaz. 2022. Dynamic Schema Graph Fusion Network for Multi-Domain Dialogue State Tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland. https://doi.org/10.18653/v1/2022.acl-long.10

[11] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating Implicit Measures to Improve Web Search. *ACM Trans. Inf. Syst.* 23, 2 (2005). https://doi.org/10.1145/1059981.1059982

[12] Ahmed Hassan. 2012. A Semi-Supervised Approach to Modeling Web Search Satisfaction. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Portland, Oregon, USA) *(SIGIR '12)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2348283.2348323

[13] Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. 2010. Beyond DCG: User Behavior as a Predictor of a Successful Search. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (New York, New York, USA) *(WSDM '10)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1718487.1718515

[14] Ahmed Hassan, Yang Song, and Li-wei He. 2011. A Task Level Metric for Measuring Web Search Satisfaction and Its Application on Improving Relevance Estimation. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (Glasgow, Scotland, UK) *(CIKM '11)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2063576.2063599

[15] Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. 2015. Learning Knowledge Graphs for Question Answering through Conversational Dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado. https://doi.org/10.3115/v1/N15-1086

[16] Scott B. Huffman and Michael Hochster. 2007. How Well Does Result Relevance Predict Session Satisfaction?. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Amsterdam, The Netherlands) *(SIGIR '07)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1277741.1277839

[17] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002). https://doi.org/10.1145/582415.582418

[18] Kalervo Järvelin, Susan L. Price, Lois M. L. Delcambre, and Marianne Lykke Nielsen. 2008. Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions. In *Advances in Information Retrieval*, Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg.

[19] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic Online Evaluation of Intelligent Assistants. In *Proceedings of the 24th International Conference on World Wide Web* (Florence, Italy) *(WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE. https://doi.org/10.1145/2736277.2741669

[20] Diane Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009). https://doi.org/10.1561/1500000012

[21] To Eun Kim and Aldo Lipani. 2022. A Multi-Task Based Neural Model to Simulate Users in Goal-Oriented Dialogue Systems. 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. https://doi.org/10.1145/3477495.3531814

[22] Julia Kiseleva, Eric Crestan, Riccardo Brigo, and Roland Dittel. 2014. Modelling and Detecting Changes in User Satisfaction. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (Shanghai, China) *(CIKM '14)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2661829.2661960

[23] Julia Kiseleva, Jaap Kamps, Vadim Nikulin, and Nikita Makarov. 2015. Behavioral Dynamics from the SERP's Perspective: What Are Failed SERPs and How to Fix Them?. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2806416.2806483

[24] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Predicting User Satisfaction with Intelligent Assistants. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2911451.2911521

[25] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding User Satisfaction with Intelligent Assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (Carrboro, North Carolina, USA) *(CHIIR '16)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2854946.2854961

[26] Klaus Krippendorff. 2013. Commentary: A Dissenting View on So-Called Paradoxes of Reliability Coefficients. *Annals of the International Communication Association* 36, 1 (2013). https://doi.org/10.1080/23808985.2013.11679143

[27] Aldo Lipani, Ben Carterette, and Emine Yilmaz. 2019. From a User Model for Query Sessions to Session Rank Biased Precision (SRBP). In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval* (Santa Clara, CA, USA) *(ICTIR '19)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3341981.3344216

[28] 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 (2021). https://doi.org/10.1145/3451160

[29] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas. https://doi.org/10.18653/v1/D16-1230

[30] Alistair Moffat, Falk Scholer, and Paul Thomas. 2012. Models and Metrics: IR Evaluation as a User Process.

[31] Alistair Moffat and Justin Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27, 1, Article 2 (2008). https://doi.org/10.1145/1416950.1416952

[32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA. https://doi.org/10.3115/1073083.1073135

[33] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas. https://doi.org/10.18653/v1/D16-1264

[34] Justus J Randolph. 2005. Free-Marginal Multirater Kappa (multirater K [free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. *Online submission* (2005).

[35] Ehud Reiter. 2018. A Structured Review of the Validity of BLEU. *Computational Linguistics* 44, 3 (2018). https://doi.org/10.1162/coli_a_00322

[36] Zhengxiang Shi, Yue Feng, and Aldo Lipani. 2022. Learning to Execute Actions or Ask Clarification Questions. In *Findings of NAACL* (2022-01-01) *(NAACL '22)*.

[37] Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. StepGame: A New Benchmark for Robust Multi-Hop Spatial Reasoning in Texts. In *Proceedings of the Association for the Advancement of Artificial Intelligence* (2022-01-01) *(AAAI '22)*.

[38] Matthijs J Warrens. 2010. Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification* 4, 4 (2010). https://doi.org/10.1007/s11634-010-0073-4

[39] Emine Yilmaz, Manisha Verma, Nick Craswell, Filip Radlinski, and Peter Bailey. 2014. Relevance and Effort: An Analysis of Document Utility. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (Shanghai, China) *(CIKM '14)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2661829.2661953