Evaluating Variable-Length Multiple-Option Lists in Chatbots and Mobile Search

Pepa Atanasova
University of Copenhagen, Copenhagen, Denmark

Georgi Karadzhov, Yasen Kiprov
SiteGround Hosting EOOD, Sofia, Bulgaria

Preslav Nakov
Qatar Computing Research Institute, HBKU, Doha, Qatar

Fabrizio Sebastiani
Istituto di Scienza e Tecnologie dell'Informazione,
Consiglio Nazionale delle Ricerche, 56124 Pisa, Italy

ABSTRACT

In recent years, the proliferation of smart mobile devices has lead to the gradual integration of search functionality within mobile platforms. This has created an incentive to move away from the "ten blue links" metaphor, as mobile users are less likely to click on them, expecting to get the answer directly from the snippets. In turn, this has revived the interest in Question Answering. Then, along came chatbots, conversational systems, and messaging platforms, where the user needs could be better served with the system asking follow-up questions in order to better understand the user's intent. While typically a user would expect a single response at any utterance, a system could also return multiple options for the user to select from, based on different system understandings of the user's intent. However, this possibility should not be overused, as this practice could confuse and/or annoy the user. How to produce good variable-length lists, given the conflicting objectives of staying short while maximizing the likelihood of having a correct answer included in the list, is an underexplored problem. It is also unclear how to evaluate a system that tries to do that. Here we aim to bridge this gap. In particular, we define some necessary and some optional properties that an evaluation measure fit for this purpose should have. We further show that existing evaluation measures from the IR tradition are not entirely suitable for this setup, and we propose novel evaluation measures that address it satisfactorily.

KEYWORDS

Chatbots, Mobile Search, Evaluation Measures.

ACM Reference Format:
Pepa Atanasova, Georgi Karadzhov, Yasen Kiprov, Preslav Nakov, and Fabrizio Sebastiani. 2019. Evaluating Variable-Length Multiple-Option Lists in Chatbots and Mobile Search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331308

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '19, July 21–25, 2019, Paris, France
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331308

1 INTRODUCTION

Mobile devices have emerged as an essential and integral part of our lives. Yet, the limited size of mobile screens, and the consequently reduced amount of displayed information, have brought about new challenges in terms of user experience. The first step towards addressing them is to depart from the "ten blue links" metaphor and to actually understand the user's information needs [6, 18]. Moreover, chatbots have been introduced as a way to include a follow-up interaction with the user [12].

Most chatbots in actual use [7, 9, 11, 24] share a common characteristic: they are capable of handling different types of user needs, usually called *intents*. As the chatbot might engage in an elaborate series of actions and utterances to fulfill the user's intent, a high-accuracy intent detection module becomes a crucial component of such systems. Yet, as the number of intents being handled grows, the accuracy of the intent classifier tends to decrease, with reported F_1 values down to 0.73 and even to 0.52 [8, 11], depending on the domain and the number of intents considered.

In order to mitigate intent misclassification, personal assistants can use strategies such as continue with the most likely intent, ask for confirmation, return a list of possible intents, or repeat the question. Previous research has found that users prefer a list of the most likely intents, but also noted that this "complicates with clutter; unnatural; more reading" [4], i.e., the list should be concise.

To this end, toolkits such as the IBM Watson Assistant[1] and Oracle's Digital Assistant[2] provide functionality for defining confidence thresholds, which allow more candidate intents to be displayed when the model is not confident. This means less typing by the user and faster narrowing down the user's request [4].

The examples we present below show that going in the wrong direction might have very negative consequences. Instead, the system could present a list of highly likely options and then leave it to the user to select the correct one, e.g.,

User:	*My credit card is toast.*	**User:**	*I want to cancel it today*
Bot:	What do you want to do?	Bot:	What do you want to cancel?
Bot:	►Replace a broken card.	Bot:	► account
Bot:	►Report a stolen card.	Bot:	► card
		Bot:	► last transaction

The system should be careful, though, not to suggest too many options, as this could confuse and/or annoy the user [10]. Moreover, it should try to make sure the list contains a good suggestion. Depending on the presentation mode, the order in which the options are presented might or might not matter.

[1] http://console.bluemix.net/docs/services/assistant/dialog-runtime.html
[2] https://bit.ly/2XbW1Do

Current research and development for chatbots moves in the direction of open-domain task-oriented systems, with an ever-increasing number of intents, which makes variable-length lists much more common. However, the evaluation of such systems remains an underexplored problem, and existing measures from the IR tradition do not fit this setup well enough. Therefore, we define a set of properties that an evaluation measure should satisfy in order to optimize two conflicting objectives simultaneously, i.e., (*i*) reduce the size of the list, and (*ii*) maximize the likelihood that a good option is indeed in the list. We further propose evaluation measures that satisfy all these desiderata. While here we focus on chatbots, most of the arguments we present apply to mobile search, too.

2 ASSUMPTIONS AND DESIDERATA

Our goal is to evaluate systems that, given a user question, try to understand the underlying intent and to answer with a suitable response. We assume that the question expresses a single user intent and therefore the system can return a single correct response. [3] We further assume that the system always returns a non-empty list of responses,[4] and that different responses correspond to different intents. We represent response lists as sequences of symbols from $\{c, w\}$, where c stands for a correct response and w stands for a wrong one. Finally, we assume that the position of the correct answer (if returned) in the list may or may not matter, depending on the context. In other words, we cater for the fact that in some applications the results should be considered a plain unordered *set*, while in some others they should form a *ranked list.*

Next, we define a set of properties that an evaluation measure M for variable-length lists should satisfy. Given two response lists $\mathbf{r}_1$ and $\mathbf{r}_2$ for the same user question, a property specifies which one M should give a higher score. Note that we take M to be a measure of accuracy, and not of error, i.e., higher values of M are better. We use $r_{ij} \in \{c, w\}$ to refer to the response item at the j-th position of $\mathbf{r}_i$. We further define a function $\#_s(\mathbf{r_i}) \equiv |\{r_{ij} \in \mathbf{r}_i | r_{ij} = s\}|$ that, given a response list $\mathbf{r}_i$, returns the number of responses of type s (where $s \in \{c, w\}$) that $\mathbf{r}_i$ contains. Moreover, we define a function $p(\mathbf{r_i}) \equiv \sum_{r_{ij}=c} \frac{1}{rank(r_{ij})}$ that, given a response list $\mathbf{r_i}$, returns the reciprocal rank of the correct response, or 0 if no correct response was returned. Finally, we define a re-scaling function $s(x, newMAX) \equiv x * newMAX$, which re-scales x from the range [0, 1] to the range [0, newMAX]. We use $|\mathbf{r}_i|$ for the *length* of $\mathbf{r}_i$, and the symbol $>$ to express preference between two lists.

Property 1. ***(Correctness)*** *If* $\#_c(\mathbf{r}_1) > \#_c(\mathbf{r}_2)$, *then* $M(\mathbf{r}_1) > M(\mathbf{r}_2)$. □

This property states that a response list that contains a correct response should be preferred to one that does not.

Property 2. ***(Confidence)*** *If* $\#_c(\mathbf{r}_1) = \#_c(\mathbf{r}_2)$ *and* $\#_w(\mathbf{r}_1) < \#_w(\mathbf{r}_2)$, *then* $M(\mathbf{r}_1) > M(\mathbf{r}_2)$. □

This property states that, if two lists contain the same number of correct responses (can be 0 or 1), the list with fewer wrong responses is preferable. The aim is to limit the length of the response list.

Property 3. ***(Priority)*** *If* $\#_c(\mathbf{r}_1) = \#_c(\mathbf{r}_2)$ *and* $\#_w(\mathbf{r}_1) = \#_w(\mathbf{r}_2)$ *and* $p(\mathbf{r}_1) \geq p(\mathbf{r}_2)$, *then* $M(\mathbf{r}_1) \geq M(\mathbf{r}_2)$. □

This property states that if two lists both contain a correct response, then the list where the correct response is ranked higher should be preferred.

We view ***Correctness*** and ***Confidence*** as mandatory properties for all our measures, and ***Priority*** as an optional one, depending on whether the results from the chatbot application are presented as an unordered set or as a ranked list.

3 EVALUATION MEASURES

In Table 1, we present an evaluation of existing information retrieval measures w.r.t. the introduced properties. The response lists are ranked according to the properties at hand, and together with the priority between the properties, we obtain a unique "gold ranking."

Next, we compare the evaluation scores and the resulting rank order of the various measures with respect to the "gold ranking." To this end, we estimate Kendall's Tau and Spearman's rank correlation between the two and we indicate the errors in the rank positions. We also provide information about the properties that each of the measures satisfies or violates.

Existing Measures for Unranked Retrieval. Considering evaluation of unordered sets, one obvious candidate is F_1. We can see in Table 1 that F_1 is successful at rewarding the presence of the correct answer (due to the recall component) and, usually, at minimizing the length of the response list (due to the precision component). Unfortunately, its value is always 0 when there is no correct response, Thus, it fails to satisfy ***Confidence***.

We try to solve the problem by smoothing F_1 (denoted as F_1^s), which we obtain by appending an extra correct response at the end of each list. The resulting measure does not suffer from the above problems of F_1, but fails to distinguish between a list consisting of a single wrong response and lists with one correct and four wrong responses, thus failing to satisfy ***Correctness***.

Existing Measures for Ranked Retrieval. A natural candidate measure for ranked retrieval is *Average Precision* (*AP*). It computes precision after each relevant response, thus satisfying both ***Correctness*** and ***Priority***. However, it disregards the number of the returned irrelevant responses and even stops computing at the last relevant response, ignoring all the subsequent irrelevant responses, and thus it fails to satisfy ***Confidence***.

In order to allow the length of the response list to influence the evaluation score, it has been proposed [14] to append a terminal response **t** at the end of each response list, which is called the AP^L measure. AP^L manages to alleviate some of the ranking problems observed in *AP*, but still fails to satisfy ***Confidence*** in some cases.

Another way to make *AP* penalize wrong responses at the end of the list is to use smoothing (denoted as AP^s). Now, the more wrong responses we return at the end of the list, the lower the precision at the last recall level will get. As a result, AP^s manages to reduce further the number of errors in the ranking with respect to the gold order. Nevertheless, AP^s still does not satisfy ***Confidence*** in cases when the result lists have different numbers of wrong results and different positions of the correct responses, as in *wcw* and *cwww*. This is due to AP^s failing to apply the properties in the correct order, i.e., applying ***Priority*** before ***Confidence***.

[3] We leave the case of multiple possible correct answers for future work.
[4] We assume a special *default* intent with a default answer to cover the case when the system cannot understand the intent.

	Unranked Retrieval				Ranked Retrieval									
Result list	Gold	F_1	F_1^s	*LAR*	Gold	*AP*	AP^L	AP^s	*RR*	*nDCG*	$nDCG^L$	*RBP*	RBP^L	*OLAR*
c	1	1.00	1.00	1.00	1	1.00	1.00	1.00	1.00	1.00	1.00	0.50	1.00	1.000
cw	2	0.67	0.80	0.75	2	(*) 1.00	0.83	0.83	(*) 1.00	(*) 1.00	0.92	(*) 0.50	0.75	0.756
wc	2	0.67	0.80	0.75	3	0.50	0.58	0.58	0.50	0.63	0.69	0.25	0.50	0.744
cww	4	0.50	0.67	0.67	4	(*) 1.00	(*) 0.75	(*) 0.75	(*) 1.00	(*) 1.00	(*) 0.88	(*) 0.50	(*) 0.63	0.675
wcw	4	0.50	0.67	0.67	5	(*) 0.50	0.50	0.50	(*) 0.50	(*) 0.63	0.65	(*) 0.25	0.38	0.663
wwc	4	0.50	0.67	0.67	6	0.33	0.42	0.42	0.33	0.50	0.57	0.13	0.25	0.659
cwww	7	0.40	0.57	0.63	7	(*) 1.00	(*) 0.70	(*) 0.70	(*) 1.00	(*) 1.00	(*) 0.85	(*) 0.50	(*) 0.56	0.634
wcww	7	0.40	0.57	0.63	8	(*) 0.50	(*) 0.45	(*) 0.45	(*) 0.50	(*) 0.63	(*) 0.62	(*) 0.25	(*) 0.31	0.622
wwcw	7	0.40	0.57	0.63	9	(*) 0.33	0.37	0.37	(*) 0.33	(*) 0.50	0.54	(*) 0.13	0.19	0.618
wwwc	7	0.40	0.57	0.63	10	0.25	0.33	0.33	0.25	0.43	0.50	0.06	0.13	0.616
cwwww	11	0.33	0.50	0.60	11	(*) 1.00	(*) 0.67	(*) 0.67	(*) 1.00	(*) 1.00	(*) 0.83	(*) 0.50	(*) 0.53	0.610
wcwww	11	0.33	0.50	0.60	12	(*) 0.50	(*) 0.42	(*) 0.42	(*) 0.50	(*) 0.63	(*) 0.61	(*) 0.25	(*) 0.28	0.598
wwcww	11	0.33	0.50	0.60	13	(*) 0.33	(*) 0.33	(*) 0.33	(*) 0.33	(*) 0.50	(*) 0.52	(*) 0.13	(*) 0.16	0.594
wwwcw	11	0.33	0.50	0.60	14	(*) 0.25	0.29	0.29	(*) 0.25	(*) 0.43	0.48	(*) 0.06	0.09	0.591
wwwwc	11	0.33	0.50	0.60	15	0.20	0.27	0.27	0.20	0.39	0.46	0.03	0.06	0.590
w	16	0.00	(△) 0.50	0.50	16	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.488
ww	17	(*) 0.00	0.40	0.25	17	(*) 0.00	(*) 0.00	0.17	(*) 0.00	(*) 0.00	(*) 0.00	(*) 0.00	(*) 0.00	0.244
www	18	(*) 0.00	0.33	0.17	18	(*) 0.00	(*) 0.00	0.13	(*) 0.00	(*) 0.00	(*) 0.00	(*) 0.00	(*) 0.00	0.163
wwww	19	(*) 0.00	0.29	0.13	19	(*) 0.00	(*) 0.00	0.10	(*) 0.00	(*) 0.00	(*) 0.00	(*) 0.00	(*) 0.00	0.122
wwwww	20	(*) 0.00	0.25	0.10	20	(*) 0.00	(*) 0.00	0.08	(*) 0.00	(*) 0.00	(*) 0.00	(*) 0.00	(*) 0.00	0.098
Correctness		Yes	No	Yes		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Confidence		No	Yes	Yes		No	No	No	No	No	No	No	No	Yes
Priority		No	No	No		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Kendall's Tau		0.970	0.985	1		0.746	0.827	0.857	0.746	0.746	0.811	0.746	0.811	1
Spearman correlation		0.992	0.994	1		0.855	0.926	0.934	0.855	0.855	0.918	0.855	0.918	1

Table 1: Comparison of evaluation measures. The 1st column indicates all response lists with up to 5 responses. "Gold" columns contain the ideal ranking of these lists according to our properties, while the other columns contain the evaluation scores from the corresponding measures. We designate inconsistencies in the rank order w.r.t. the gold order with an asterix (*) when due to violation of *Confidence*, and with a triangle (△) when due to violation of *Correctness*. The compliance of the measures with the properties is indicated at the bottom of the table, where the correlation of the ranking with the gold ranking is also shown.

Reciprocal Rank (*RR*) is a popular measure for ranked retrieval, which accounts for the position of the 1st correct response, disregarding any following irrelevant responses. *AP* is equivalent to *RR* in the case of a single correct response (and so is *DCG*).

Furthermore, although *normalized Discounted Cumulative Gain* (*nDCG*) is designed specifically for the evaluation of different relevance scores, we study its performance on our "gold ranking." However, it does not penalize wrong responses, violating ***Confidence***. Using the technique proposed in [14], we end up with $nDCG^L$, which manages to reduce the number of ranking errors, but still violates ***Confidence*** in some cases.

[16] proposed *Rank-Biased Precision (RBP)*, which models a user that decides to continue reading the next item in the response list with probability p. As in [14], we set p=0.5. *RBP* struggles with the same problems as *RR* and violates ***Confidence***. Even if we apply the technique from [14], the ranking produced by RBP^L contains a lot of errors compared to the "gold ranking."

New Measures. The existing measures we have discussed are suitable for optimizing the number of correct responses and their positions. Most of them fulfill ***Correctness*** and ***Priority***, but struggle with ***Confidence***. This is not surprising as the IR tradition (except for recent work like [1, 14]) is concerned with the rank positions of the relevant documents, not with the length of the response list, which is conceptually infinite. Before [1, 14], this length had never been a parameter in any of the proposed measures.

Smoothing lists by appending an additional correct response is beneficial but not enough to achieve perfect correlation with the "gold ranking," as the Kendall's Tau and the Spearman rank correlation scores indicate. In order to bridge this gap, we introduce a new measure, *Length-Aware Recall* (*LAR*), which operates on a list of responses $\mathbf{r}_n$ and gives preference to lists with fewer negatives:

$$LAR = \frac{R(\mathbf{r}_n) + \frac{1}{|\mathbf{r}_n|}}{2} \quad (1)$$

The new measure first computes the recall $R(\mathbf{r}_n)$ of the returned list. Then, it includes the confidence of the system about the correct result expressed by the reciprocal value of the list's length, i.e., the confidence decreases when the number of returned responses increases. Taking the mean of the two scores, we create an intuitive score of both recall and confidence. The possibility of having zero values for recall makes arithmetic mean preferable to harmonic mean, also used to combine evaluation criteria.

As *LAR* satisfies both ***Correctness*** and ***Confidence***, it can be used to evaluate variable-length lists by modeling the true positive rate and the optimal response length jointly. Moreover, it is perfectly correlated with the "gold ranking" in the unordered scenario.

However, *LAR* does not satisfy the ***Priority*** property, which makes it unfit for scenarios where order does matter. In order to fix this issue, we propose an extension of *LAR* that includes an additional third term for the rank of the correct response.

This fix gives rise to *Ordered Length-Aware Recall* (OLAR):

$$OLAR = \frac{R(\mathbf{r}_n) + \frac{1}{|\mathbf{r}_n|} + s(p(\mathbf{r}_n), \mu)}{2 + \mu} \quad (2)$$

The third term in the above equation accounts for ***Priority***, and it is larger when the rank of the correct item moves lower in the list. We rescale it because of the priority order that we have defined for the properties – it should not contribute to the score more than the ***Confidence*** term. Given that $(\frac{1}{max(|r_i|)-1} - \frac{1}{max(|r_i|)}) = 0.05$ is the smallest difference between two ***Priority*** scores of lists with length up to 5, we re-scale it in $[0, \mu]$, where $\mu = 0.05 - \lambda$ and $\lambda = 0.001$ is an insignificantly small number.

To sum up, we introduced two measures, which are in a perfect correlation with the two "gold rankings." The measures consist of separate terms, accounting for different properties, which makes them easily interpretable and extensible for specific needs.

4 RELATED WORK

Related properties of evaluation measures. As we define a set of properties that need to be satisfied by an evaluation measure for variable-length output, our work is closely related to the properties of truncated evaluation measures introduced in [1]. Their *Relevance Monotonicity* property is similar to our *Correctness*, except that *Correctness* imposes strict monotonicity. The *Irrelevance Monotonicity* property is similar to our *Confidence*, but it discounts for irrelevant documents at the end of the list only.

[15] defined seven properties for effectiveness measures. The relevant properties, which our new evaluation measures also satisfy are *Completeness*, *Realisability*, *Localisation* and *Boundedness*. However, their *Monotonicity* property contradicts our *Confidence* property because it states that adding documents, which are not relevant, to the end of the list increases the score.

[3, 21] also conducted an "axiomatic" analysis of IR-related evaluation measures. Our properties *Confidence* and *Priority* are akin to the ones discussed in [2], but the latter are used in a different setup, i.e., for document retrieval, clustering, and filtering.

Related evaluation measures. In Section 3, we discussed and evaluated the most relevant evaluation measures for both unranked (precision, recall, F_1, and smoothed F_1) and ranked retrieval (*nDCG* [13], *RR*, *MAP*, smoothed *MAP*, *RBP* [14]). We found that they were unable to penalize the wrong responses according to the "gold order," thus violating *Confidence*.

Apart from these measures, [17] introduced the *c@1* measure, a modification of accuracy, suited for systems that may not return responses. However, their approach still does not penalize the number of returned wrong responses at the end of the list. Furthermore, [17, 19, 23] assumed that a system can return an empty result list, which tackles the problem when the request does not have a correct response. However, we assume that the system should return at least one result, even if it is a default fallback intent.

Another relevant field of research analyzes the likelihood that the user will continue exploring the response list based on different signals - time-biased gain [22], length of the snippet [20], and information foraging [5]. However, in mobile search and chatbot platforms, the presented information is already minimized, and thus we aim to reduce the length of the returned results instead.

5 CONCLUSION AND FUTURE WORK

We have studied the problem of evaluating variable-length response lists for chatbots, given the conflicting objectives of keeping these lists short while maximizing the likelihood of having a correct response in the list. In particular, we argued for three properties that such an evaluation measure should have. We further showed that existing evaluation measures from the IR tradition do not satisfy all of them. Then, we proposed novel measures that are especially tailored for the described scenarios, and satisfy all properties.

We plan to extend this work to the context in which more than one correct answer might exist, since a long and complex input question may contain multiple intents [25]. This would also be of interest for mobile retrieval in general, where results may need to be truncated due to limited screen space.

REFERENCES

[1] A. Albahem, D. Spina, F. Scholer, A. Moffat, and L. Cavedon. 2018. Desirable Properties for Diversity and Truncated Effectiveness Metrics. In *Proc. of ADCS.*
[2] E. Amigó, J. Gonzalo, and F. Verdejo. 2013. A General Evaluation Measure for Document Organization Tasks. In *Proceedings of SIGIR.*
[3] E. Amigó, D. Spina, and J. Carrillo de Albornoz. 2018. An Axiomatic Analysis of Diversity Evaluation Metrics: Introducing the Rank-Biased Utility Metric. In *Proceedings of SIGIR.*
[4] Z. Ashktorab, M. Jain, Q. V. Liao, and J. D Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of CHI.*
[5] L. Azzopardi, P. Thomas, and N. Craswell. 2018. Measuring the Utility of Search Engine Result Pages: An Information Foraging Based Measure. In *Proc. of SIGIR.*
[6] R. Baeza-Yates, A. Z. Broder, and Y. Maarek. 2011. The New Frontier of Web Search Technology: Seven Challenges. In *Search Computing.* Springer, 3–9.
[7] T. Bocklisch, J. Faulker, N. Pawlowski, and A. Nichol. 2017. Rasa: Open Source Language Understanding and Dialogue Management. (2017). arXiv:1712.05181.
[8] D. Braun, A. Hernandez-Mendez, F. Matthes, and M. Langen. 2017. Evaluating Natural Language Understanding Services for Conversational Question Answering Systems. In *Proceedings of SIGDIAL.*
[9] M. Burtsev and other 19 authors. 2018. DeepPavlov: Open-Source Library for Dialogue Systems. In *Proceedings of ACL: System Demonstrations.*
[10] A. P. Chaves and M. A. Gerosa. 2019. How Should my Chatbot Interact? A Survey on Human-Chatbot Interaction Design. (2019). arXiv:1904.02743 [cs.HC].
[11] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril, M. Primet, and J. Dureau. 2018. Snips Voice Platform: An Embedded Spoken Language Understanding System for Private-by-Design Voice Interfaces. (2018). arXiv:1805.10190 [cs.CL].
[12] A. Følstad and P. Brandtzæg. 2017. Chatbots and the New World of HCI. *Interactions* 24, 4 (2017), 38–42.
[13] K. Järvelin and J. Kekäläinen. 2000. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proceedings of SIGIR.*
[14] F. Liu, A. Moffat, T. Baldwin, and X. Zhang. 2016. Quit While Ahead: Evaluating Truncated Rankings. In *Proceedings of SIGIR.*
[15] A. Moffat. 2013. Seven Numeric Properties of Effectiveness Metrics. In *Proceedings of AIRS.*
[16] A. Moffat and J. Zobel. 2008. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems* 27, 1 (2008), 2.
[17] A. Peñas and A. Rodrigo. 2011. A Simple Measure to Assess Non-Response. In *Proceedings of ACL.*
[18] T. Russell-Rose and T. Tate. 2012. *Designing the Search Experience: The Information Architecture of Discovery.* Newnes.
[19] T. Sakai. 2004. New Performance Metrics Based on Multigrade Relevance: Their Application to Question Answering. In *Proceedings of NTCIR.*
[20] T. Sakai and Z. Dou. 2013. Summaries, Ranked Retrieval and Sessions: A Unified Framework for Information Access Evaluation. In *Proceedings of SIGIR.*
[21] F. Sebastiani. 2015. An Axiomatically Derived Measure for the Evaluation of Classification Algorithms. In *Proceedings of ICTIR.*
[22] M. D. Smucker and C. Clarke. 2012. Modeling User Variance in Time-Biased Gain. In *Proceedings of CHIIR.*
[23] E. M. Voorhees. 2001. Overview of the TREC 2001 Question Answering Track. In *Proceedings of TREC.*
[24] J. D. Williams, E. Kamal, M. Ashour, H. Amr, J. Miller, and G. Zweig. 2015. Fast and Easy Language Understanding for Dialog Systems with Microsoft Language Understanding Intelligent Service (LUIS). In *Proceedings of SIGDIAL.*
[25] P. Xu and R. Sarikaya. 2013. Exploiting Shared Information for Multi-Intent Natural Language Sentence Classification. In *Proceedings of INTERSPEECH.*