Studying Online and Offline Evaluation Measures: A Case Study Based on the NTCIR-14 OpenLiveQ-2 Task

Piyush Arora(✉) and Gareth J. F. Jones

ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland
{Piyush.Arora,Gareth.Jones}@dcu.ie

Abstract. We describe our participation in the NTCIR-14 OpenLiveQ-2 task and our post-submission investigations. For a given query and a set of questions with their answers, participants in the OpenLiveQ task were required to return a ranked list of questions that potentially match and satisfy the user's query effectively. In this paper we focus on two main investigations: (i) Finding effective features which go beyond only-relevance for the task of ranking questions for a given query in Japanese language. (ii) Analyzing the nature and relationship of online and offline evaluation measures. We use the OpenLiveQ-2 dataset for our study. Our first investigation examines user log-based features (e.g number of views, question is solved) and content-based features (BM25 scores, LM scores). Overall, we find that log-based features reflecting the question's popularity, freshness, etc dominate question ranking, rather than content-based features measuring query and question similarity. Our second investigation finds that the offline measures highly correlate among themselves, but that the correlation between different offline and online measures is quite low. We find that the low correlation between online and offline measures is also reflected in discrepancies between the systems' rankings for the OpenLiveQ-2 task, although this depends on the nature and type of the evaluation measures.

Keywords: Learning To Rank models · Question-answer ranking · Online and offline testing · Correlation of online and offline measures

1 Introduction

Interactive websites for community based question answering (CQA) provide opportunities to search and ask questions ranging from critical topics related to health, education and finance to recreational queries for the purpose of fun and enjoyment. Yahoo Chiebukuro (YCH)[1] is a community question answering service which provides a question retrieval system in Japanese language managed by the Yahoo Japan Corporation. The NTCIR-14 OpenLiveQ-2 is a benchmark task which aims to provide an open live test environment using the Yahoo

[1] https://chiebukuro.yahoo.co.jp/.

M. P. Kato et al. (Eds.): NTCIR 2019, LNCS 11966, pp. 70–82, 2019.
https://doi.org/10.1007/978-3-030-36805-0_6

Chiebukuro engine where, given a query and a set of questions with their answers, task participants had to return a ranked list of questions. Final evaluation of the results was based on real user feedback. Involving real users in evaluation helps to incorporate the diversity of search intents and relevance criteria by utilising real queries and feedback from users who are engaged in real search tasks, which makes this task more interesting. The submitted systems were evaluated using offline measures such as NDCG@10, ERR@10 and Q-measures and online evaluation metrics using a pairwise preference multileaving approach (discussed later in Sect. 2). This paper describes our participation in the OpenLiveQ-2 task and our post-submission investigations.

Overview of System Submissions: This task focuses on modelling textual based information and click log based information to rank questions to handle the challenges of: (i) queries being ambiguous and having diverse intent, and (ii) modelling user behaviour effectively. A range of Learning To Rank (L2R) models have been investigated in the OpenLiveQ task held at NTCIR-13 and NTCIR-14, respectively [4,7–9]. These L2R models focus on selecting a diverse range of features with effective weights to improve the systems' performance as measured using offline and online evaluation measures. However, apart from [8], not much work has been done in analyzing the nature and type of good features to address the OpenLiveQ task of ranking question-answer pairs for a given query. This observation motivated us to study feature importance for question ranking. In [8], the authors trained a L2R model using a coordinate ascent algorithm for question ranking. To calculate feature importance they removed each feature one at a time, retrained their ranking model and analyzed the relative decrease in the overall scores of NDCG@10, ERR@10 compared to the ranking model learnt using all the features. If a feature is relatively important, its removal led to a greater decrease in the NDCG@10, ERR@10 scores.

As L2R approaches have shown to be quite successful for this task, we also explored L2R models to address the task of ranking question-answer pairs for a given query. We submitted 14 system runs (including the baseline) for the OpenLiveQ-2 task [9]. Our top submission systems were ranked 2 on NDCG@10, ranked 3 on ERR@10, and ranked 6 on Q-measure among the 65 system submissions made to the task. However, these systems which ranked quite high on the offline evaluation measures of NDCG@10, ERR@10 and Q-measure, had a rank below 35 among the 65 submissions made to the task on the online evaluation measure. This contrasting ranking of our system submissions between online and offline evaluation measures motivated us to pursue an investigation on the relationship of the online and offline evaluation measures used in the OpenLiveQ-2 task.

To address the above limitations of finding effective features for ranking questions and to study the nature of online and offline evaluation measures, we set out the following questions for our investigation:

- **RQ-1:** What are the effective features for the task of ranking question-answer pairs for the OpenLiveQ-2 task?

- **RQ-2:** What is the correlation between different online and offline evaluation measures used in the OpenLiveQ-2 task?

Our work seeks to understand the relationship between the online and offline evaluation measures. This topic is of emerging interest in the information retrieval (IR) research community to build better ranking models, and to improve user engagement and satisfaction. The main contributions of our work are:

1. Investigating effective features for the task of question ranking. We find that log-based features reflecting a question's popularity, freshness, etc. dominate the question's ranking over content-based features measuring query-question similarity. Our findings on the OpenLiveQ-2 dataset support the findings of previous work [8] carried out on the OpenLiveQ-1 dataset [4]. In [8] the authors removed one feature at a time to examine the importance of each feature by calculating the decrease in the offline measures (e.g NDCG score) as compared to the equivalent metric score of a combined model built using all the features. In our work we build a comprehensive single model, using all the features, using Gradient Boosting Trees (described later in Sect. 4). We find feature importance by calculating probability estimates of how much the feature contributes to reducing data misclassification. Our work also contributes confirming the claims and findings of previous research on this topic.
2. We study the relationship between different offline and online measures for the OpenLiveQ-2 task. We analyze the fine-grained results output of our 65 system submissions for the task to calculate Pearson correlation between the offline measure scores, such as NDCG, ERR at rank 5, 10, 20 and 50 and Q-measure and the online measure score (described later in Sect. 5). We find that the offline measures correlate highly amongst themselves, but that the correlation between different offline and online measures is quite low. The low correlation between online and offline measures is also reflected in the discrepancies between the systems' rankings for the OpenLiveQ-2 task.

We anticipate that our findings will encourage IR researchers to carefully examine the relationship and variation in system scores and rankings, while using alternative online and offline evaluation measures. The remainder of this paper is organised as follows: Sect. 2 introduces the dataset, tools used and the evaluation strategy of the OpenLiveQ-2 task, Sect. 3 describes an overview of our approach adopted in our participation in this task, Sect. 4 gives results and analysis of our submissions to the task, Sect. 5 descries the relative performance and ranking of the top-k system submissions and describes our investigation studying the relationship between the different evaluation measures used in this task, and finally Sect. 6 concludes.

2 Dataset, Tools and Evaluation

In this section we describe the dataset for OpenLiveQ-2 task, the tools used for this work and the evaluation strategy adopted for the task. As a part of the dataset for the OpenLiveQ-2 task, the organisers provided the query logs and for each query a corresponding set of questions with a best answer retrieved by the YCH engine. Table 1 presents information regarding the number of queries, questions in the training and the test sets. Since the data is in the Japanese language, so as to facilitate participation from diverse and non-native speaking teams in the development of effective systems, the task organisers provided a list of textual features indicating the scores of relevance models such as BestMatch (BM25) [14], Language Model (LM) [13] etc., for a query and a corresponding set of questions. Table 2 presents a list of the complete features which were provided by the task organisers comprising of textual and click-log based information. We refer interested readers to [4,7] for more details of these features and the dataset construction for this task.

Table 1. Dataset details

Training set	Size	Test set	Size
Number of Queries	1000	Number of Queries	1000
Number of Questions	986125	Number of Questions	985691
Number of click logs	288502	Number of click logs	148388

Table 2. All extracted features provided in the dataset

Title	Id	Snippet	Id	Question body	Id	Best answer	Id	Click logs	Id
tf_sum	F1	tf_sum	F18	tf_sum	F35	tf_sum	F52	answer_num	F69
log_tf_sum	F2	log_tf_sum	F19	log_tf_sum	F36	log_tf_sum	F53	log_answer_num	F70
norm_tf_sum	F3	norm_tf_sum	F20	norm_tf_sum	F37	norm_tf_sum	F54	view_num	F71
log_norm_tf_sum	F4	log_norm_tf_sum	F21	log_norm_tf_sum	F38	log_norm_tf_sum	F55	log_view_num	F72
idf_sum	F5	idf_sum	F22	idf_sum	F39	idf_sum	F56	is_open	F73
log_idf_sum	F6	log_idf_sum	F23	log_idf_sum	F40	log_idf_sum	F57	is_vote	F74
icf_sum	F7	icf_sum	F24	icf_sum	F41	icf_sum	F58	is_solved	F75
log_tfidf_sum	F8	log_tfidf_sum	F25	log_tfidf_sum	F42	log_tfidf_sum	F59	rank	F76
tfidf_sum	F9	tfidf_sum	F26	tfidf_sum	F43	tfidf_sum	F60	updated_at	F77
tf_in_idf_sum	F10	tf_in_idf_sum	F27	tf_in_idf_sum	F44	tf_in_idf_sum	F61		
bm25	F11	bm25	F28	bm25	F45	bm25	F62		
log_bm25	F12	log_bm25	F29	log_bm25	F46	log_bm25	F63		
lm_dir	F13	lm_dir	F30	lm_dir	F47	lm_dir	F64		
lm_jm	F14	lm_jm	F31	lm_jm	F48	lm_jm	F65		
lm_abs	F15	lm_abs	F32	lm_abs	F49	lm_abs	F66		
dlen	F16	dlen	F33	dlen	F50	dlen	F67		
log_dlen	F17	log_dlen	F34	log_dlen	F51	log_dlen	F68		

As outlined in Sect. 1, the OpenLiveQ-2 task had offline and online evaluation phases.

- **Offline evaluation phase**: system performance was measured using *NDCG* [3], *ERR* [1], and *Q-measure* [15,16].

- **Online evaluation phase**: a *pairwise preference multileaving (ppm)* approach was used [6,12] to measure system performance.

The evaluation methodology in OpenLiveQ-2 focused on a two phase online evaluation strategy. In the first phase all the systems were evaluated online to identify the top-k systems, these top-k systems were then compared in detail to ensure that the top systems could be statistically distinguished. For each of the submitted rankings of questions, a multileaving approach was used to form a new set of combined rankings and shown to the users as part of the YCH engine. For a given query each of the questions in the original ranked list that was clicked when presented to a user received a credit, these credit scores were aggregated over the ranked list, and are referred to as the cumulative gain (CG). This CG score was used to rank the systems in the online evaluation phase [4]. To find the top-k systems, the task organisers used a pairwise preference multileaving (PPM) approach which infers pairwise preferences between documents from clicks. The PPM model is based on the assumption that a clicked document is preferred to: (a) all of the unclicked documents above it; (b) the next unclicked document. These assumptions are commonly used in pairwise Learning To Rank models, for more details refer to [12].

3 System Development: Approach Used

In this section we present an overview of our approach to the OpenLiveQ-2 task. Submissions to the previous OpenLiveQ-1 task showed positive results using L2R models [8], thus as a part of our investigation we focused on exploring L2R models [10,11] to rank a set of question-answer pairs given an input query. In L2R models, a ranking function is created using the training data, such that the model can precisely predict the ranked lists in the training data. Given a new query, the ranking function is used to create a ranked list for the documents associated with the query. The focus of L2R technologies is to successfully leverage multiple features for ranking, and to learn automatically the optimal way to combine these features. In this work, we used the Lemur RankLib toolkit [2]. This toolkit provides an implementation of a range of L2R algorithms which have been shown to be successful in earlier work.

Table 3. Feature set

Type of features	Feature's id range	Information type
Title based textual features (Title set)	[F1–F17]	Content based information
Snippet based textual features (Snippet set)	[F18–F34]	Content based information
Question body based textual features (Body set)	[F35–F51]	Content based information
Body answer based textual features (Answer set)	[F52–F68]	Content based information
Click log features (Click set)	[F69–F77]	Logs based information

The main focus of this task was to effectively combine text-based features measuring the similarity of queries with a set of questions and click-based information captured through user logs. We investigated feature selection extensively to determine a good set of features to rank the questions effectively for a given set of test queries. A complete set of features is shown in Table 2. To select features and combine them effectively, we broadly categorised the set of 77 features into 5 main categories, as shown in Table 3. We have diverse feature sets capturing relevance of: (i) user query to question title (*Title set*), (ii) user query to question body (*Body set*), (iii) user query to question snippets (*Snippet set*), (iv) user query to the best answer (*Answer set*), and (v) click logs based information (*Click set*). We explored alternative combinations of these diverse features set.

Run Submissions: As described above, we used the RankLib toolkit for our experiments. Models were trained on the training dataset comprising of about 1M questions (data points) and among which about 300k questions (data points) had information about user interactions. The models were optimised based on the ERR@10 metric. We submitted 14 systems as a part of this investigation. For more details on our approach and different runs that were submitted for this task kindly refer to our system submission paper [9].

Table 4. Offline evaluation scores and system rankings for our submissions for the OpenLiveQ-2 task. The best scores are in boldface.

		System scores			System ranking		
Systems	System - id	NDCG@10	ERR@10	Q-Measure	NDCG-Rank	ERR-Rank	Q-Rank
Best scores		0.333	0.209	0.502	1	1	1
Average scores		0.204	0.128	0.436	NA	NA	NA
System-1	99	0.074	0.044	0.382	59	59	56
System-2	106	0.237	0.171	0.454	32	24	26
System-3	110	0.239	0.138	0.444	31	33	29
System-4	112	0.188	0.137	0.370	36	35	64
System-5	118	0.117	0.106	0.340	45	38	65
System-6	123	0.326	0.202	**0.495**	5	5	**6**
System-7	126	0.204	0.138	0.438	34	34	32
System-8	128	0.285	0.191	0.459	22	20	24
System-9	130	**0.331**	**0.203**	0.464	**2**	**3**	21
System-10	133	0.227	0.148	0.449	33	32	27
System-11	143	0.302	0.189	0.445	19	21	28
System-12	147	0.287	0.179	0.466	21	23	20
System-13	150	0.295	0.181	0.464	20	22	22
System-14	152	0.258	0.154	0.491	25	31	17

4 Results and Analysis

In this section we give results and analysis of our submissions to the OpenLiveQ-2 task. Tables 4 and 5 present the results of our submitted systems for the offline and online evaluation measures respectively. As a part of the official metrics, the organisers reported and compared the ranks and scores of the systems across all three measures NDCG (normalized discounted cumulative gain), ERR (expected reciprocal rank), and Q-measure. As shown in Table 4, we can see some quite distinct variations across the three scores (NDCG@10, ERR@10 and Q-scores) for the system submissions, indicating that these three evaluation metrics do not show consistent trends. For example, System-14 shows Q-scores similar to the best scores of System-6, however the NDCG@10 and ERR@10 scores are quite low compared to the highest scores of System-9.

Table 5. Online evaluation scores and system rankings for our submissions for the OpenLiveQ-2 task. The best two systems are in boldface.

Systems	System - id	Phase-1 online evaluation		Phase-2 online evaluation	
		Cumulative gain	Rank	Cumulative gain	Rank
Best scores		2633.202	1	1867.440	1
Average scores		−4.92	NA	−13.94	NA
System-1	99	−1420.907	61	NA	NA
System-2	**106**	**1843.815**	**7**	**1002.855**	**7**
System-3	110	190.395	40	NA	NA
System-4	112	1721.431	8	428.370	10
System-5	**118**	**2006.333**	**4**	**1129.577**	**6**
System-6	123	70.797	43	NA	NA
System-7	126	1326.385	14	241.362	12
System-8	128	−83.103	46	NA	NA
System-9	130	282.391	38	NA	NA
System-10	133	−40.834	44	NA	NA
System-11	143	171.302	41	NA	NA
System-12	147	452.896	29	−418.210	23
System-13	150	276.774	39	NA	NA
System-14	152	369.791	35	NA	NA

As described in Sect. 2, the online evaluation was conducted in two phases. In the first phase all 61 distinct system submissions were compared in an online setting using a pairwise preference multileaving approach to select the top 30 submissions which were then compared extensively. Table 5 presents results of both the online evaluation phases. In the first phase of online evaluation only two of our submissions (System: ID-128 and ID-133) scored below the average score, the remaining 11 systems performed better than the average score, and five of our thirteen systems were selected to be compared in the final phase of online evaluation. In the final phase of online evaluation only one of our five systems scored below the average score. Our best systems in the online phase

System-5 (ID: 118) and System-2 (ID: 106) were ranked “6” and “7”, among the top 30 systems.

Table 6. Top two systems, L2R models were trained using coordinate ascent algorithm with default parameters: tolerance = 0.001, iterations = 25 and random restarts = 5

Systems	System - id	Features combined	Feature set
System-2	106	All 77 Features (F1: F77)	Title+Snippet+Body+Answer+Click
System-5	118	Feature F69:F77 (Click features)	Only click set

To find effective features for the task of ranking question-answer pairs for a given query, we inspect our two best system submissions, System-5 and System-2, in detail. We select these two systems as they perform best in the online evaluation and are the only systems which capture all nine user log based features, as shown in Table 6. We calculated feature importance to find effective features for ranking question-answer pairs for a given query for both System-2 and System-5. We learnt a gradient boosting classification algorithm on the training data using scikit-learn[2]. A gradient boosting algorithm builds a decision tree model using a cross entropy loss function, where node of the trees are the features in the training model. The decision tree model splits the tree node by calculating the gini impurity over all the features. Gini impurity is a measurement of the likelihood of an incorrect classification, it gives a probability estimate of how well the feature splits the data to minimize the data misclassification [17]. The importance of each feature is calculated based on its contribution to splitting the data effectively to perform better classification over the training corpus.

Table 7. Feature rankings representing important features for System-5.

Rank	Feature-id	Feature-name	Important value
1	Feature-71	Number of views	0.246
2	Feature-72	Log of number of views	0.239
3	Feature-77	Updated date	0.226
4	Feature-76	Best rank	0.173
5	Feature-69	Number of answers	0.052
6	Feature-70	Log of number of answers	0.051
7	Feature-75	Is solved	0.010
8	Feature-74	Is voted	0.004
9	Feature-73	Is open	0.000

Table 7 presents features in descending order of importance for predicting effective ranking of questions for System-5. Among the user log based information, features such as number of views, updated date and best rank are relatively

[2] https://scikit-learn.org/stable/.

important, indicating that the online results ranking and clicks are influenced by the questions' popularity, freshness, and the relative position in the ranked list, also called position bias [5].

Table 8. Feature rankings representing important features for System-2. Only features greater than 0.01 importance value are shown.

Rank	Feature-id	Feature-name	Value
1	F-71	Number of views	0.116
2	F-70	Log of number of answers	0.114
3	F-75	Is solved	0.058
4	F-76	Best rank	0.043
5	F-6	Log of Idf sum with title	0.034
6	F-40	Log of Idf sum with question body	0.031
7	F-5	Idf sum with Title	0.030
8	F-69	Number of answer	0.026
9	F-68	Log of best answer length	0.025
10	F-41	ICF sum of question body	0.024
11	F-4	Log of norm of TF sum	0.023
12	F-43	Tf-Idf sum of question body	0.023
13	F-39	Idf sum of question body	0.023
14	F-8	Log of Tf-Idf sum of title	0.022
15	F-7	ICF sum of title	0.021
16	F-42	Log of TF-Idf sum of question body	0.021
17	F-35	TF sum of question body	0.019
18	F-56	Idf sum of best answer	0.017
19	F-38	Log of norm of Tf sum of question body	0.015
20	F-57	Log of Idf sum of best answer	0.014
21	F-58	Icf sum of best answer	0.013
22	F-46	Log of BM25 of question body	0.011
23	F-12	Log of BM25 of question title	0.010
24	F-55	Log of norm of TF sum of best answer	0.010

Table 8 presents features in descending order of importance for predicting effective ranking for question-answer pairs for System-2. Most dominant features are "number of views" and "log of number of answer", indicating the popularity of the question. Important features corresponding to content-based information are "query" and "question title" matching followed by "query" and "question body" matching. It is peculiar to see BM-25 scores are ranked 22 and 23, and thus are not as effective relatively in ranking questions for System-2. Similar findings were observed in [8], where the authors found that the top 2 features for question ranking are “log of number of views” and “log of number of answers” indicating the popularity of a question for the NTCIR-13 OpenLiveQ dataset [4]. They found that BM25 scores ranked 15 and 20 in terms of the feature’s ranking. However, in their work they found that snippet based features are more effective, while in our work we found that feature matching “query” with “question title” and “question body”, respectively is more effective than matching “query” with “snippet”.

In summary, we inspected the top features for question ranking by analyzing our top 2 systems for the OpenliveQ-2 task. Some of our findings on the relative importance of features concur with the previous findings reported in [8], thus adding to the reproducibility of the claims with respect to feature importance for question ranking. As most of the traditional IR models work on optimizing relevance to improve over the offline metrics, it becomes necessary to model other aspects such as popularity, diversity and freshness as they tend to perform relatively better on the online metrics. We anticipate that the findings from the feature analysis for the task of ranking questions will encourage more work on understanding how different features correspond to online user behaviour. We have tried to bridge the gap between understanding important features for question ranking and hope that this work will lead to more investigation on the interaction and relationship across these different features.

5 Evaluation Metrics Correlation

In this section we investigate the relationship between the online and offline evaluation metrics. We study how well the online and offline evaluation measures correlate with each other. We use Pearson correlation (r) which is a measure of the linear correlation between two variables x and y as shown in Eq. 1 using scipy library[3].

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{1}$$

where n is the sample size, x_i, y_i are the individual sample points indexed with i, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

As indicated in Sect. 1, there were 65 system submissions made for the OpenLiveQ-2 task. We calculated fine grained evaluation scores for the relevance measures such as NDCG and ERR at different ranks 5, 10, 20 and 50 and Q-Measure for all the 65 systems. We also had the online cumulative gain scores for the online evaluation phase-1 for all 65 systems. We used these 65 data points to find correlation values across different offline and online evaluation measures.

Table 9 presents Pearson correlation results between diverse set of NDCG, ERR at rank 5, 10, 20 and 50 values and for Q-measure and online cumulative gain metrics. The results indicate that the correlation coefficient of the

[3] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html.

Table 9. Pearson correlation of all the reported evaluation measures used for the OpenLiveQ-2 task. $*$ and $**$ indicates that the p-value is more than 0.05 and 0.01 respectively. For all other correlation values p-value is less than 0.01. ≡ indicates that it is a symmetrical relationship. N stands for NDCG, E stands for ERR and CG stands for cumulative gain measures.

Pearson	N@5	N@10	N@20	N@50	E@5	E@10	E@20	E@50	Q	CG
N@5	–	≡	≡	≡	≡	≡	≡	≡	≡	≡
N@10	0.997	–	≡	≡	≡	≡	≡	≡	≡	≡
N@20	0.989	0.997	–	≡	≡	≡	≡	≡	≡	≡
N@50	0.972	0.986	0.995	–	≡	≡	≡	≡	≡	≡
E@5	0.995	0.987	0.976	0.953	–	≡	≡	≡	≡	≡
E@10	0.996	0.992	0.982	0.966	0.999	–	1.00	≡	≡	≡
E@20	0.997	0.994	0.986	0.968	0.997	1.00	–	≡	≡	≡
E@50	0.996	0.994	0.986	0.970	0.996	0.999	1.00	–	≡	≡
Q	0.917	0.939	0.954	0.972	0.892	0.904	0.91	0.912	–	≡
CG	0.333	0.325	0.301**	0.278**	0.368	0.375	0.372	0.374	0.225*	–

offline evaluation metrics is quite high ($r >= 0.9$). However there are some noticeable differences, NDCG@k and ERR@k measures have higher correlation as compared to between NDCG@k and Q-measure and between ERR@k and Q-measure, respectively. Overall, the online evaluation metric CG shows low correlation with the offline evaluation metrics ($r \in [0.225 - 0.375]$). The CG evaluation measure shows higher correlation with ERR@k values as compared with NDCG@k and Q-measures.[4]

The low correlation values between the online and offline evaluation measures explains why the system rankings are quite varied depending on the choice of evaluation metric, as shown in Tables 4 and 5. The online and offline metrics do not go hand in hand and focus on optimization of different aspects and lead to a difference in system ranking. The trained models are tuned and optimized on metrics including NDCG@10 and ERR@10. Thus, for the test queries, question rankings perform quite well when measured using NDCG@10, ERR@10, but evaluating the systems using online metrics, such as cumulative gain, produces low results. For future tasks, involving online and offline evaluation, we recommend the exploration of alternative offline measures for model training and system evaluation that correlates well with the online metrics.

6 Conclusions

In this study we examined the features that are important for question ranking for the OpenLiveQ-2 task. We explored different features to find those that

[4] Similar pattern of results were observed using Spearman's and Kendall's Tau correlation metrics during our investigation, results have been omitted because of the space constraints.

contribute effectively for the task of question ranking. We found that features indicating the popularity, freshness and relative position of a question are among the top features for question ranking. Some of these results concur with previous findings on the earlier OpenLiveQ-1 task. Most of IR approaches focus on improving relevance and optimizing models on NDCG, ERR, but we find that in an online setting, there are more diverse features which are important, thus there is a need to incorporate features beyond relevance that capture information effectively. We anticipate the findings in this work will lead to more investigation of the interaction between different features used for ranking questions.

We studied the relationship between online and offline evaluation measures. We calculated Pearson correlation between different offline evaluation measures such as NDCG, ERR at rank 5, 10, 20 and 50 and Q-measure and the online evaluation metric measured using cumulative gain. We found that all the offline evaluation measures correlate well with each other, however the correlation of the offline and online measures is quite low. The low correlation between the online and offline evaluation metrics lead to variation in the ranking of systems depending on the choice of evaluation metric. We anticipate the findings in this work will draw attention from the community, and lead to more work in understanding the relationship between online and offline evaluation measures.

Acknowledgement. This research is supported by Science Foundation Ireland (SFI) as a part of the ADAPT Centre at Dublin City University (Grant No: 12/CE/I2267).

References

1. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM, pp. 621–630 (2009)
2. Dang, V.: The Lemur Project-Wiki-Ranklib (2013). http://sourceforge.net/p/lemur/wiki/RankLib
3. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. (TOIS) **20**(4), 422–446 (2002)
4. Kato, M.P., Liu, Y.: Overview of NTCIR-13. In: Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies (2017)
5. Joachims, T., Granka, L.A., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 154–161. SIGIR (2005)
6. Kato, M.P., Manabe, T., Fujita, S., Nishida, A., Yamamoto, T.: Challenges of multileaved comparison in practice: lessons from NTCIR-13 OpenLiveQ Task. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM, pp. 1515–1518 (2018)
7. Kato, M.P., Nishida, A., Manabe, T., Fujita, S., Yamamoto, T.: Overview of the NTCIR-14 OpenLiveQ-2 task. In: Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies (2019)
8. Manabe, T., Nishida, A., Fujita, S.: YJRS at the NTCIR-13 OpenLiveQ task. In: Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies (2017)

9. Arora, P., Jones, G.J.F.: DCU at the NTCIR-14 OpenLiveQ-2 task. In: Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies (2019)
10. Qin, T., Liu, T.Y., Xu, J., Li, H.: LETOR: a benchmark collection for research on learning to rank for information retrieval. J. Inf. Retrieval **13**(4), 346–374 (2010)
11. Metzler, D., Croft, W.B.: Linear feature-based models for information retrieval. J. Inf. Retrieval **10**(3), 257–274 (2007)
12. Oosterhuis, H., de Rijke, M.: Sensitive and scalable online evaluation with theoretical guarantees. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM, pp. 77–86 (2017)
13. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 275–281. SIGIR (1998)
14. Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: NIST Special Publication, no. 500225, pp. 109–123 (1995)
15. Sakai, T.: Evaluating evaluation metrics based on the bootstrap. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 525–532. SIGIR (2006)
16. Sakai, T.: On the reliability of information retrieval metrics based on graded relevance. J. Inf. Process. Manag. **43**(2), 531–548 (2007)
17. Breiman, L.: Some properties of splitting criteria. J. Mach. Learn. **24**(1), 41–47 (1996)