AITOK at the NTCIR-14 OpenLiveQ-2 Task

Hiroki Tanioka

Center for Administration of Information Technology
Tokushima University, Japan
tanioka.hiroki@tokushima-u.ac.jp

Abstract. The AITOK team participated in the NTCIR-14 OpenLiveQ-2 Task. This report describes our approach to ranking question-answering lists of Yahoo! Chiebukuro search results for the queries of Query IDs, and discusses our results of test results. Our approach intends to make sure of the degree of catchy to question-answering, thereby integrates two strategies. The first strategy is statistics based approach with questions and clickthrough data. The second strategy is natural language based approach with question-answer pairs. In the offline test, we struggled increasing Q-measure using these two strategies day by day. Additionally, we employed manually dynamic programming approach for optimization of Q-measure. Although the approach is very simple, a result sorted by a score mainly based on page view is the best in all results. However, the best result in the offline test is not enough in the online test. Instead, the other result sorted by a score based on clickthrough with last-updated time in descending order is ranked in the top group.

Team Name. AITOK

Subtasks. OpenLiveQ-2

Keywords: Question-answering · Information Retrieval · Freshness

1 Introduction

Question-answering system (QA system) is expected to reply appropriate responses to a user question. However, it is difficult to appropriately respond to un-responsible question. Thus, we struggled detecting the response possibilities to the question [8], and challenged how to respond to the user from the viewpoint of common grounding [7]. Meanwhile, commercial QA systems adopts a search engine as knowledge database. The search engine contains past question-answer pairs, and replies some past question-answer pairs related to the user question. However, we have an unsettled issue that how to order is the best for result of question-answer pairs on the QA system. This issue was discussed at the NTCIR-13 Open Live Test for Question Retrieval (OpenLiveQ) Task [4] [1] [2]. Hence, the AITOK team participated in the NTCIR-14 OpenLiveQ-2 task [3], in order to clear how to rank the QA results. This report describes our approach to ranking question-answering lists of Yahoo! Chiebukuro search results for the queries of Query IDs, and discusses our results of test results.

Table 1. A list of statistical fields.

Field Name	Description
Title	Title of the question
Snippet	Snippet of the question in a search result
Body	Body of the question

Table 2. A list of natural language field.

Field Name	Description
Page view	Page view of the question
Number of answers	Number of answers for the question
Update	Last update time of the question
Clickthrough rate	Clickthrough rate

2 Approach

The organizer's report [4] describes a reversal phenomenon between the offline test and the online test. For instance, even though a submitted run is the best on the offline test, the run is the lower place on the online test. This phenomenon is occurred upon all the submitted runs. The OKSAT team reports [5] that the top-ranked run result is ranked with the training dataset. On the offline test, the run results seemed to be submitted and analyzed based on the feedbacked evaluation results. Therefore, we also use the training dataset for ranking our run results and analyze based on the feedbacked evaluation results. The AITOK team considers that the most often clicked question-answer pair is the catchiest question-answer pair. Hence, the catchy score is defined in some ranking scores with statistics information (Table 1) and natural language information (Table 2). Then, the ranking scores are composed of various field information and improved on the feedback evaluation result (Q-measure).

3 Methodologies

The NTCIR-14 OpenLiveQ-2 task provides keywords as queries. A query keyword which is separated into some words is used for calculating a catchy score to question-answer pairs. Some statistics are also added to the score of separated words. Furthermore, we consider which field or score is effective for ranking question-answer pairs, comparing between the results of the offline test and the online test.

3.1 NLP

For calculating the catchy score, Cabocha [10] is used as a morphological analyzer and a dependency parser, and word2vec [9] is used for word embeddings. Separated words are weighted as 1-gram or 2-grams TF-IDF on the Bag-of-Words

Table 3. A run list on offline test.

No.	Date	Q-measure	Rank	Desc
101	8/24	0.38194	56	This result is only for uploading test from AITOK.
102	8/25	0.39724	45	1-gram & TF-IDF with click through rate with cutoff
103	8/26	0.39852	44	1-gram TF-IDF+ with click through rate with cutoff
105	8/27	0.40479	43	2-gram TF-IDF+ with click through rate with cutoff
107	8/28	0.42008	40	2-gram TF-IDF+ with click through rate with cutoff without rank
109	8/29	0.41748	42	Dependent 2-gram TF-IDF with click through rate with cutoff without rank
111	8/30	0.4391	31	2-gram TF-IDF+ with click with cutoff and view without rank
115	8/31	0.42676	36	2-gram TF-IDF+ with click and view with cutoff without rank
117	9/1	0.43231	33	cutoff and view
120	9/2	0.49363	14	view count
122	9/3	0.49319	16	click through and view count
124	9/4	0.49347	15	view count sorted with click, updated, answers, order, rank and cutoff
125	9/5	0.49393	13	view count sorted with answers, cutoff, click, updated, order and rank
127	9/6	0.499	4	view count sorted with answers × tf-idf weighted by query
129	9/7	0.5	3	view count + answers × 2-gram tf-idf weighted by query
131	9/8	0.50152	1	view count + answers × snippet 2-gram tf-idf weighted by query
134	9/9	0.49838	5	view count + answers × snippet 2-gram tf-idf weighted by query
137	9/10	0.50028	2	view count + answers × snippet 2-gram tf-idf double-weighted by norm query
139	9/11	0.49483	7	view count + answers × snippet word2vec double-weighted by norm query
141	9/12	0.49427	9	view count + answers × snippet word2vec double-weighted by norm query v2
145	9/13	0.49412	12	view count + answers × snippet L1 word2vec double-weighted by norm query
149	9/14	0.49437	8	view count + answers × snippet cos word2vec double-weighted by norm query

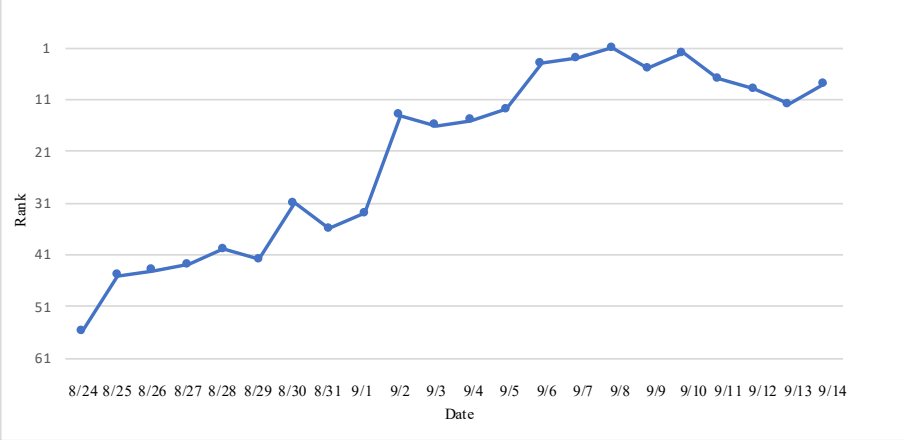

Fig. 1. The history of submitted runs on the offline test.

model. In case of dependency parser, a phrase (2-grams) is composed of separated words under the restriction of modification relation. In case of using word embeddings, a separated word is transformed to 200-demensions, and compared with word embeddings of question-answer pairs using L1 or cosine similarity.

3.2 Statistics

Catchy score increases if a question-answer pair includes words and phrases that are included in question-answer pair with high clickthrough rate and large amount of page view. Thereby, the number of answers and updated date are adopted to the catchy score.

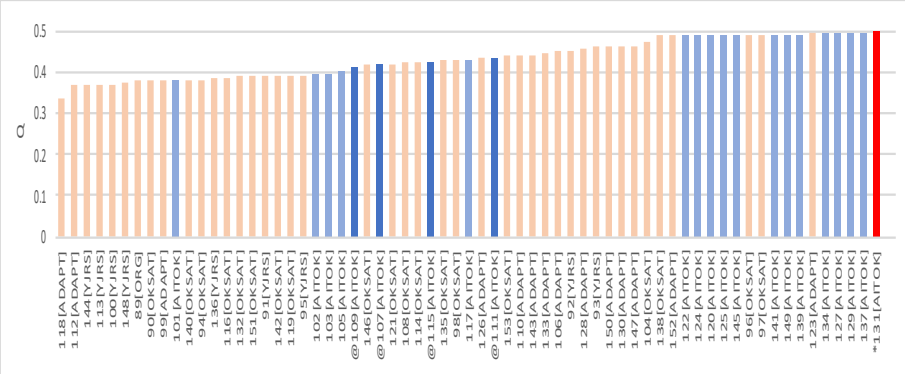

Fig. 2. The result of offline test. (Q-measure)

4 Experiments

Table 3 describes our all runs in the offline test. Fig. 1 shows the ranking history. The runs No.101 to No.115 are try and error phase. In the phase, we found out that page view is a key information. The runs No.107 to No.127 are sorted by page view with some other information. The runs No.129 to No.137 are tried to top ranked score with page view and number of answers weighted by TF-IDF. The runs No.139 to No.149 are optional challenges using word embeddings. After the offline test, all the runs of participated teams are uncontrollable during four months on the online test.

5 Results

5.1 Offline Test

Fig. 2 shows our good run results in Q-measure on the offline test. The run No.131 is the best. In addition, the run result is the best in almost all evaluation metrics (nDCG@5, nDCG@10, nDCG@20, nDCG@50, ERR@5, ERR@10, ERR@20, and ERR@50). The run No.131 focuses page view and number of answers with TF-IDF score.

5.2 Online Test

Fig. 3 and Fig. 4 show the results of the online test. The online test uses Multileaved Comparisons [6] for the evaluation. In these results, the run No.131 is overtaken by many runs including the runs No.115, No.107, No.109, and No.111 in our try and error phase. The run No.117 is just in the middle of runs in the Top 30 online test, which is sorted by page view with freshness (last-updated time). The run No.111 is sorted by TF-IDF score weighted with clickthrough, page view, and freshness.

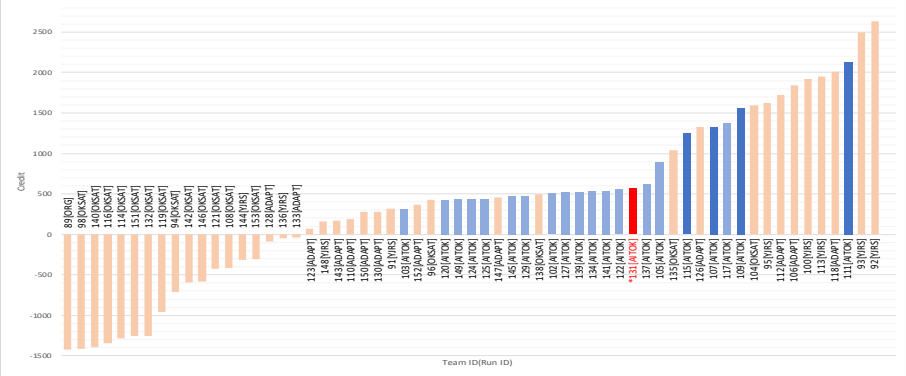

Fig. 3. The comparison of Top 60 online test results.

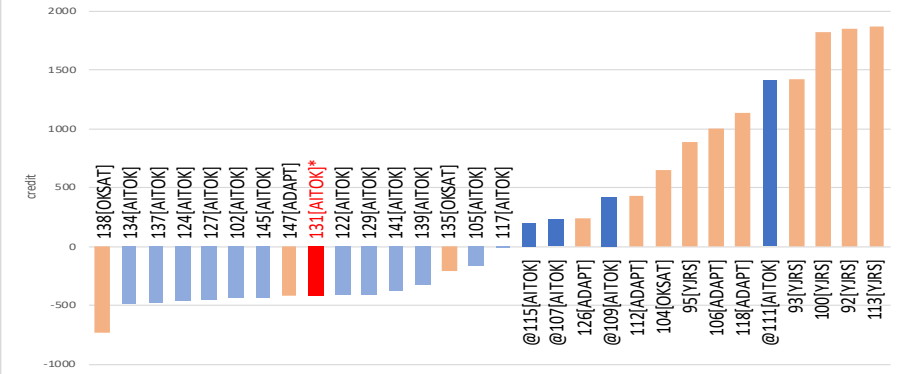

Fig. 4. The comparison of Top 30 online test results.

6 Conclusions

This report describes the AITOK team's results on the NTCIR-14 OpenLiveQ-2 task. According to the comparison between the offline test and the online test, the online test evaluation results are clearly different from the offline test results. Our team run results showed that TF-IDF and page view were important on the offline test. However, clickthrough rate and freshness (last-updated time) were more important on the online test. Ideally, the offline test trend should be same as the online test. Because the online test takes a high cost and is not replicable due to using a real service and users. Hence, we will try to refine the evaluation metrics and dataset of the offline task.

References

1. Kato, M., Yamamoto, T., Manabe, T., Nishida, A., Fujita, S.: A Large-scale Live Evaluation of Question-Answering Systems on a Community cQA Site [Translated

from Japanese.]. In: DEIM. G2-2 (2018)
2. 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. pp. 1515–1518. CIKM '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3269206.3269318, http://doi.acm.org/10.1145/3269206.3269318
3. Kato, M.P., Nishida, A., Manabe, T., , Fujita, S., Yamamoto, T.: Overview of the NTCIR-14 OpenLiveQ-2 Task. In: NTCIR-14 Conference (2019)
4. Kato, M.P., Yamamoto, T., Manabe, T., Nishida, A., Fujita, S.: Overview of the NTCIR-13 OpenLiveQ Task. In: NTCIR-13 Conference (2017)
5. Sato, T.: OKSAT at NTCIR-13 OpenLiveQ Task. In: NTCIR-13 Conference (2017)
6. Schuth, A., Sietsma, F., Whiteson, S., Lefortier, D., de Rijke, M.: Multileaved comparisons for fast online evaluation. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. pp. 71–80. CIKM '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2661829.2661952, http://doi.acm.org/10.1145/2661829.2661952
7. Tanioka, H.: Response Generation for Grounding in Communication at NTCIR-13 STC Japanese Subtask. In: NTCIR-13 Conference (2017)
8. Tanioka, H., Nakatani, R., Suzuko, Y., Uchida, Y.: Grounding of Question-Answering System and Detection of Response Possibility [Translated from Japanese.]. In: The 24th Annual Meeting of the Association for Natural Language Processing. The Association for Natural Language Processing (2018)
9. word2vec : Tool for computing continuous distributed representations of words. https://code.google.com/p/word2vec, [Online; accessed 11-September-2018]
10. CaboCha: Yet Another Japanese Dependency Structure Analyzer. http://www.chasen.org/~taku/software/cabocha/, [Online; accessed 16-September-2018]