Evaluation of interactive question answering

Abstract

Interactive question answering(QA) systems are designed for users to take some control on the content and seek the answer in an interactive way with contextual environment, which can be considered as an intersection of QA and conversational agent [1]. Evaluation of interactive QA is of great importance in the systems, but developing robust evaluation methods remains an unresolved research issue. In this project, I proposed an evaluation method for interactive QA systems in terms of two aspects: capabilities and performance. Capabilities evaluation presents a roadmap to assess the capacity of salient features that are needed in interaction for a good interactive QA system by querying corresponding questions. Performance evaluation combines the evaluation techniques for QA system on Text Retrieval Conference(TREC) and the evaluation metrics for conversational agents using Mechanical Turk(MTurk) as platform. A series of experiments were conducted on MTurk using simulated conversational question answering data. The result indicates the success of the designed evaluation in providing appropriate evaluation metrics and reliable evaluation results for interactive QA systems.

Keyword: Evaluation, interactive question answering, Mechanical Turk, TREC

1.Introduction

QA systems are designed to provide automatic answers for human questions or clues rather than a list of documents that contain the answer [2]. Evaluation plays an important role to improve the performance of the system. Among all the evaluation techniques, TREC-QA track as a large-scale evaluation of independent-domain QA systems is a predominant evaluation platform to foster the research of QA [3]. However, one of the significant shortcomings for the QA systems is that the system can not provide an appropriate answer if a user's initial query is not enough. In order to take previous questions and answers into consideration and to create a kind of continuous interaction between the user and the system, a new type of QA systems, interactive QA systems, has been developed [4]. It’s difficult to evaluate an interactive QA because of the unpredictable evolvement of interaction. Thus it is necessary to have humans involved in the evaluation process which makes the process slow and difficult. To overcome this obstacle, this project introduced a method that combines the evaluation techniques of dialogue agents and QA systems using Amazon Mechanical Turk (MTurk) as a platform, which is a fast, feasible, cheap and reliable crowdsourcing method [5]. The workers on MTurk can assess whether the returned answer is correct as assessors of TREC and fill in questionnaire for the conversation quality evaluation as a user. Since the fundamental purpose for evaluation is to improve the performance of the system, this project also proposed to measure the abilities that are important to an interactive QA system, such as context processing abilities, which can provide detailed improvement in terms of specific features.

In this paper, I will investigate the existing literatures about evaluation of interactive QA in Section 2. In Section 3, I will introduce the state-of-art QA systems and the evaluation methods for QA from TREC. In Section 4, I’ll present conversational agents and the existing evaluation metrics. In Section 5, the designed evaluation method for interactive QA is shown. In Section 6, the experiment implemented on MTurk is described and the result is analyzed. Finally, the conclusions and future directions are provided in Section 7.

2.Related work

Most of the previous attempts at evaluating interactive QA have involved a human user questionnaire. Kelly[6] et al. described the usage of questionnaires to evaluate interactive QA systems. They proposed fifteen hypotheses about the types of functions that a good IQA system should provide, such as, they assumed a good interactive QA system should support information gathering with lower cognitive workload. Under the guide of hypotheses, they developed the methods and metrics using three questionnaires: Cognitive Workload Questionnaire, which was used to assess analysts' cognitive workloads as they completed scenarios, Task Questionnaire, which was used to assess dimensions of scenarios, such as their realism and difficulty, and System Questionnaires, which was used to assess analysts' experiences using this particular system to prepare a pseudo-report. The questionnaires were evaluated with four systems, seven analysts, and eight scenarios in the domain of chemical/biological weapons of mass destruction [7]. This method provides ways to evaluate systems from different angles, but involves a lot of preliminary manual work and time-consuming labor of human assessors [1]. Salime[9] et al. designed an interactive QA system which uses statistical techniques to extract information and a standard questionnaire was conducted to five users with the same level of knowledge to assess the quality of the system. Small [9] et al. presented an interactive QA named HITIQA, which allows the users to query complex questions, such as exploratory and analytical questions. The system was evaluated by two analysts on eleven targets and a questionnaire about user satisfactory was filled in by both analysts.

To indicate the performance of interactive QA systems in a comparative way, some literatures presented evaluation methods using both interactive QA system and non-interactive QA systems. Quarteroni [9] et al. described YourQA, which is an open-domain interactive QA system that provides answers to both fact-based and complex questions. They used series questions from TREC-QA and twelve users were invited to seek answers to the questions in non-interactive interface of the QA prototype and then the interactive version. The questionnaire result from the users showed the standard and interactive versions of the system offer different advantages but users tend to be generally more satisfied with the interactive QA than with the baseline version.

Besides the usage of questionnaire, objective metrics were also used for evaluation. Toney[10] et al. developed a texted-based interface to the RITEL system, which integrated a spoken language dialogue system and an open-domain information retrieval system. They invited each user to conduct eight conversations with the RITEL system, four speech-based and four text-based. The questionnaire they conducted for evaluation measures task ease, language generation, user expertise and expected behavior, system response time and future use. Objective methods, such as such as mean number of dialogue turns, mean duration per turn, mean number of words and word error rate were also measured. The results indicate the spoken and textual versions of the system differ significantly in objective metrics but the difference is not obvious in the perception of the users. Harabagiu[11] et al. developed FERRET, which is an interactive QA based on predictive questioning. The system returns an answer to the question that the user asks and proposes other three questions as suggestions of possible evolvement of the interaction. They proposed that the interactive QA could be evaluated in terms of efficiency, defined as the number of questions that the user must pose to find particular information, effectiveness, defined by the relevance of the answers returned, and user satisfaction. They conducted experiments with FERRET involving 8 scenarios and more than 30 users. The experiment result demonstrated that the performance of an interactive QA system can be significantly improved by predicting possible continuous questions on a given topic.

Other researchers present evaluation methods from different point of view into the filed of evaluations of interactive QA systems. Tsuneaki[12]et al. examined the capabilities and functions needed for interactive QA systems in information access interaction for writing a report through WoZ(Wizard of Oz) simulation. Four experts acted as a WoZ in the simulated QA systems and tried to answer questions from users on 20 topics. Then the utterances they collected were analyzed in terms of pragmatic phenomena, clarifications etc. They proposed the importance of providing cooperative and helpful responses as a new finding. Chai[13] et al. assumed that user's follow-up questions can provide feedback for the system to assess if the status of preceding answer is problematic. They used a set of features (e.g. target matching) in classification approaches (e.g. SVM) to decide whether the answer is classified as problematic based on the follow-up question and context. The result indicates the best performance can achieve 73.8% accuracy in identifying problematic situations. Their studies indicate that users issued distinctive utterances, such as rephrasing the question, when a problematic answer returned, which can provide useful cues for automated performance assessment in interactive QA.

3.Question Answering

3.1 Introduction

Question Answering(QA) system is designed to automatically answer questions posed by humans in a natural language using information retrieval and natural language processing techniques. There are substantial accessible QA platforms on the Web, such as START [14], YodaQA [15], True Knowledge [16], Wolfram Alpha [17] and so on. The pipeline of a QA system can be summarized as question reader, question analysis, answer production, answer analysis, answer merging and scoring and answer writer [18].

3.2 Evaluation

Since 1999, the TREC-QA track has fostered researches on QA systems steadily by expanding both the type and difficulty of the questions asked. There are three types of questions evaluated in TREC, factoid, list and other questions. Each of the three question types has its own response format and evaluation method. The series questions consist of the above three types of questions, which are about the same topic. For example,

Target:" Wolfgang Amadeus Mozart"
"Factoid": Where was Mozart born?
"Factoid": What year did Mozart die?
"List": List Mozart's operas.
"Factoid": Who was Mozart's rival?
"Other": Other

This chapter will introduce the guidance of the evaluation for different types of questions in the TREC-2007 QA track[3].

3.2.1 Factoid question evaluation

For factoid questions, the response is judged as "incorrect", "unsupported", "non-exact", "locally correct", or "globally correct" by two assessors. If the two assessors disagreed in their judgments, a third adjudicator made the final determination. The main evaluation metric for the factoid component was accuracy, which is defined as the fraction of questions judged to be globally correct.

3.2.2 List question evaluation

For list questions, systems return an unordered set of different instances of a particular type, for example, *Which airlines use Dulles Airport?* Each instance returned by the system was evaluated in the same way as the factoid questions. The score of the evaluation for list questions use instance recall (IR) and instance precision (IP) and compute F measure by combining recall and precision with equal weight:

$$F = \frac{2\times\mathrm{I}P\times\mathrm{I}R}{IP + \mathrm{I}R}$$

The list-score for a series is the mean of the F scores of the list questions in the series.

3.2.3 Other questions

The answer for the other question is the information that is desirable about the topic and not contained in the previous questions or answers in the series. The evaluation for the other questions are derived from the evaluation methods for the definition questions on TREC 2003, which were the only type of question in TREC-QAs for which the evaluation was not stable [19]. The response was evaluated in two steps. In the first step, the assessor constructed a list of desirable information nuggets about the target using the returned responses from systems and searches. Once the nugget list was created for a target, the assessor decided which were vital, which is defined as information must be returned for a response to be good and which were non-vital, which is defined as acceptable but not necessary information. Each nugget will be assigned a weight with the number of assessors who marked it to be vital and nugget weights will then be normalized. In the second step, the assessor marked and counted nuggets that appeared in the response from each system. The score for the "other" question used nugget recall (NR), which was the ratio of the sum of matched nuggets to the sum of all nuggets in the list and an approximation to nugget precision (NP) based on length. In particular, NP is computed according to the following formula, where allowance is defined as 100 times number of nuggets returned, length is the total number of characters in answer.

$$NP = 1 - \frac{\text{length}-\text{allowance}}{\text{length}}$$

If the length is smaller than the allowance, NP is 1. The final score for an Other question was an F-score, with NR weighted heavily three times than NP($\beta = 3$):

$$F = \frac{(\beta^2 + 1)\times P\times R}{\beta^2\times P + R}$$

3.2.4 Combined Score

The score(S) for per series is a combined weighted score, which is computed as the average of the three scores for three questions types in the series:

$$S=\frac{1}{3}\times(\text{factoid-score} + \text{list-score} + \text{other-score}).$$

The final score for a system will be the mean of scores for per series.

4. **Conversational agents**

4.1 Introduction

The term “Conversational agent” can be interpreted in different ways by different researchers. However the essence of the conversational agent(CA) is that it engages in interaction with other human or other computer participant in human language and can take several turns[20]. CAs are playing significant roles in different applications, for instance, in e-commerce, such as Ask Anna for IKEA, for disembodied pocket assistant in device, such as Apple Siri and Google Now, and for education, such as Oscar CITS [21]. Some of the successful online conversational agents, for instance, ALICE [22] , CLEVERBOT [23] , etc. can interact with humans on some topics but the shortcoming of these system is that they cannot adequately answer all of the queries given to them.

4.2 Evaluation

Previous attempts at evaluating conversational agents are mainly done either by constructing a questionnaire to the users trying to the reveal their subjective assessment of using the conversational agents or by studying the resulting dialogue[28]. PARADISE [24], is one prominent evaluation framework that combines user satisfaction, task success and dialogue cost into a performance function. Task success is computed as the percentage of right responses. Dialogue cost can be decomposed into efficiency and quality cost, where efficiency cost corresponds to resource consumption to complete a task and quality cost is corresponding to actual conversation content, such as naturalness and friendliness. Semeraro[25] et al. conducted a questionnaire for their bookstore conversational agent, where 7 metrics were assessed: impression, command, effectiveness, navigability, ability to learn, ability to aid, and comprehension. Shawar[26] et al. proposed evaluation in terms of 3 metrics: dialogue efficiency, quality metric and users' satisfaction and demonstrated that the proper assessment should be determined by how successfully the system completes the user's goals. Rzepka[27] et al. and Hung[28] et al. focused on methodology to determine the effectiveness and naturalness of a dialog system.

5.Interactive QA Evaluation

This chapter will present a novel framework and roadmap for evaluating interactive QA systems in two aspects, abilities and performance evaluation. Abilities evaluation is to evaluate the different capabilities that are needed for interactive QA in specific interactive circumstances and the performance evaluation is to evaluate the system in terms of quality of QA, which is to assess how well answers returned by a system meet the specific information requirements of a single question using TREC methods and interaction quality, which corresponds to the dialogue cost and user satisfaction as the metrics for conversational agents.

5.1 Abilities evaluation

The abilities evaluation can be conducted by analyst or developers of the system to assess the capacity of handling different functions in interaction by querying corresponding questions to see if the systems can return the expected response. The salient features that a good interactive QA should support and the testing question examples are summarized as Table1.

5.1.1 Context processing ability

Context processing ability is that the interactive QA systems can process and understand the questions by reformulating or requesting clarify according to the contextual environment. Context processing ability consists of ability of handling elliptic questions, ability of handling anaphoric questions, clarity ability.

(1) Ability of handling elliptic questions

Elliptic questions are the questions that some part of the question is omitted. But the system can understand the elliptic questions by taking previous questions and answers into consideration [9]. For example:

USR：When was Wolfgang Amadeus Mozart born?

SYS: January 27, 1756

USR: Where?

(2) Ability of handling anaphoric questions:

In anaphoric questions, there is a pronoun or other kind of word which refers to the previous question or answer, such as “he/she/it/they/his/her/its/their”[12].For example,

USR: What was Alfred Hitchcock’s first movie?

SYS: Rebecca

USR: How many Oscar awards did he win?

(3) Ability of clarity:

The ability of clarity is that the interactive QA system can issue a request for missing constraint or generate two or more candidates for confirmation when the user queries an ambiguous referent, vague question subject or missing constraint. For example,

USR: Where is Sully showing in Santa Cruz?

SYS: Can you specify to which day your question refers, please?’

USR: “When did Richard Nixon meet Enlai Zhou in China?”

SYS: “In 1972.”,

USR: “How old was he?”.

The system can either ask “Which person do you mean, Richard Nixon or Enlai Zhou?” or “Do you mean how old was Richard Nixon?” for confirmation.

5.1.2 Complex question processing ability

The fundamental function of the interactive QA is to seek the answer

for users’ questions. Because different evaluation measures are needed for complex questions’ answers than for those supporting factoid QA, the processing of some representative complex questions can be considered as one of the abilities that are needed to assess. The types of the questions include analytical questions, questions involved relation and comparison and double questions etc.

(1) Exploratory and analytical questions

For example, *How has air pollution in Beijing affected economy?*

(2) Questions concerning comparison and relation

For example, What is the largest bird?

(3) Double questions linked by conjunctions

For example, *In what province and country were Yeti vocalizations recorded?*

5.1.3 Complete structure ability

Conversational agents usually have an opening, a body and a closing as complete structure. The complete structure ability can be easily assessed by starting a conversation with different greeting and ending a conversation with different closing utterances, such as “Hello”, “Goodbye”.

5.1.4 Control ability

The user should take the control in the interaction to seek the information they want. While the system can initiate a question in order to confirm given information, clarify the situation, or constrain user responses, the user can barge in the systems and shift focus. This ability can be evaluated by changing subjects of the conversation.

5.1.5 Ability to provide complementary information

It’s helpful and cooperative and helpful to provide complementary additional information [12]. For example,

USR: What year was Barack Obama born?

SYS: He was born in 1961 in Honolulu, Hawaii.

5.1.6 Ability to correct wrong presupposition or spelling

When a user posed a question with wrong presupposition or incorrect spelling, a good interactive QA can recognize the mistake the correct it. For example,

USR: What is the movie Gone Girl directed by James Cameron about?

SYS: I don’t know any movie named Gone Girl directed by James Cameron. Are you interested in Gone Girl directed by David Fincher or in movies by James Cameron?

USR: Who is the author of Harry Portter?

SYS: Do you mean who is the author of Harry Potter?

Table1. abilities evaluation objectives and examples

Ability	Query or response type	Example
Context processing	Elliptic questions	*USR: “When did Richard Nixon meet Enlai Zhou in China?”* *SYS: “In 1972.”,* *USR: “How old was he?”.* *SYS: “Which person do you mean, Richard Nixon or Enlai Zhou? ”*
	Anaphoric question	
	Ambiguous question	
Complex questions handling	Exploratory and analytical question	*USR: How has air pollution in Beijing affected economy?*
	Comparison and relation question	
	Double linked questions	
Complete structure	Opening and closing	*SYS: “How can I help you?”*
Control ability	Barge in or changing subject	*SYS: typing...* *USR: “Can you tell me…”*

Provide complementary information	Exemplification, or detailed response	*USR: What year was Barack Obama born?* *SYS: He was born in 1961 in Honolulu, Hawaii.*
Correct wrong presupposition and spelling	Wrong presupposition	*USR: Who is the author of Harry Portter?* *SYS: Do you mean who is the author of Harry Potter?*
	Wrong spelling	

5.2 Performance Evaluation

5.2.1 Evaluation steps

The evaluation will be processed in two steps. A series of experiments will be conducted on MTurk in the first step. The workers will be asked to interact with the interactive QA system given the questions that they need to seek the answers. The workers can formulate the question order and question format in their opinion. After the interaction, the same worker will be asked to evaluate if the answer returned by the system is correct given the suggested answer for each question. Then a simple questionnaire was given to users to provide their feedback about their experience with the system. In the second step, the dialogue result will be exacted for quantitative analysis by coding and then result including the QA success, the efficiency cost and quality cost metrics will be shown.

5.2.2 Evaluation Metrics

The performance of interactive QA system can be evaluated in terms of question answering success metrics, inspired by TREC and dialogue costs metrics, derived from the PARADISE framework[27]. Question answering success will be evaluated by workers on MTurk, which will be represented by the combined score from score of factoid questions, list questions and other questions. Dialogue cost metrics contain efficiency cost and quality cost. The efficiency cost will evaluate the resources that consumed, such as elapsed time and turns in interaction. The quality cost can be evaluated in forms of questionnaire, which will ask the worker to fill in on MTurk on a scale from 1 to 5 after the completion of their interaction. The questions in the questionnaire and their objectives are shown in Table 3. Table 2 depicts the evaluation metrics to evaluate the performance of system.

Table 2.Evaluation Metrics

Metric	Type	Data collection method
Factoid-score	Question answering score	MTurkers
List-score	Question answering score	MTurkers
Other-score	Question answering score	MTurkers
Combined-score	Question answering score	Quantitative Analysis
Total time duration	Efficiency cost	Quantitative Analysis
Total number of user/system turns	Efficiency cost	Quantitative Analysis
Total number of system turns	Efficiency cost	Quantitative Analysis
Average number of turns per question	Efficiency cost	Quantitative Analysis
Average elapsed time per question	Efficiency cost	Quantitative Analysis
Ease of usage	Quality cost	Questionnaire
Naturalness	Quality cost	Questionnaire
Friendliness	Quality cost	Questionnaire
Robustness regarding misunderstandings	Quality cost	Questionnaire
Willingness to use system again	Quality cost	Questionnaire

Table 3. Questionnaire

Questions	Objective	Evaluation instruction

Do you think a person unfamiliar with the system could use the system easily?	Ease of usage	Please finish each question in the questionnaire by giving a score on a scale from 1= "Not at all" to 5= "Yes, Absolutely".
Were answers provided quickly enough?	Naturalness	
Did the system understand your requests the first time?	Robustness	
Is the system able to provide additional useful information?	Friendliness	
Do you think the system can maintain a natural conversation flow when interacting with the user?	Naturalness	
Do you think you would use this system again?	Willingness to use system again	
Did you get all the information you wanted using the system?	Robustness	
Overall, are you satisfied with the system?	User satisfactory	

6.Experiment

There are no actual working interactive QA systems out there on the Web that are accessible, though some systems were presented by literatures. It's not realistic to evaluate a real system using the designed evaluation method. For testing purpose, the evaluation experiment is conducted given the simulated conversation data, which can be considered as the result of the interaction. Mechanical Turk is a platform for the requester to assign a human intelligence task(HIT) for the workers. A HIT is a question that needs an answer, which represents a single, self-contained task that a worker can work on, submit an answer, and collect a reward for completing. The worker on MTurk can act as both the role of the user of the system and the role of TREC assessors. Instead of rating the result using incorrect, inexact, locally correct and globally correct, the evaluation value on MTurk are set up to binary evaluation with correct and wrong because there is no specific data collection provided as TREC. A standard questionnaire will also be given in MTurk platform to evaluate the subjective factors, such as user satisfactory. The result form MTurk will be filtered and analyzed.

6.1 Data collection

It's a common usage for users to collect information using an interactive QA system for writing a report on a given topic. For this reason, series questions that related to a specific target are collected for experiment. I selected 18 question series from the TREC-QA 2007 campaign (i.e. series 216, 217, 211,225, 226, 229, 230, 238, 240, 244, 245, 258, 261, 262, 272, 277, 279, 280, 285). There are 77 factoid questions in total, which included persons, events, organizations, plants, animals and health. Most of answers to those questions are named entities, including date expressions and places values. Most questions except the first one of each series have some anaphoric expressions, which can also be considered as testing questions for the context processing capability. Though interaction conversations are simulated, the returned answer in the conversation is actually returned by real QA systems that are submitted by the competitors on TREC and judged by the assessors. And the suggested answer is obtained by the answers which the assessor on TREC judged as correct and online searches during the question development.

6.2 Experiment design

It's important to design the experiment carefully, which will affect the result quality and the speed of completion of the experiment. First additional qualifications workers must meet are required to control the result quality: HIT approval rate for all requesters' HITs greater than 95%. Because it's rude for workers to limit too few time but too much time will result in distraction, duration has been limited as 20 minutes. Reward is $0.15 per HIT, since too low price will cause slow progress and too high price will attract cheaters. The conversations and suggested answers are uploaded through a csv file, each row in the csv file is corresponded to one HIT. Because the line break character is not supported in MTurk, each sentence of the conversations and answers are put in one column. Each HIT was assigned to 5 unique workers.

The workers are shown the purpose of the HIT and the answer format that they need to provide. The evaluation value of the retrieval result in the conversation for each question used 1= "Right" and 0="Wrong", and the assessment the performance is on a scale from 1= "Not at all" to 5= "Yes, Absolutely".

6.3 HIT instruction

In TREC, the assessors who did QA judging received training developed specifically for the task[28]. However, on MTurk there is no guarantee that each worker will follow the instruction and read the training content carefully especially it's allowed that different workers have different opinions about whether the returned answer is correct for each question. Using some standard rules can narrow the uncertainty of the evaluation and improve the comparability between different systems. The following rules are shown for evaluation guidance to users in the instruction and the instruction highlights the emphasis in bold.

1. Certain units and unit conversion were also required. “88” as not acceptable when correct answer is “88 miles”. “2 miles” was a correct substitute for answer “3.2 kilometers”.

2. If answer contained multiple entities including the correct entity, the response was judged as incorrect. For example,

USR: Who is the author of “Harry Potter”?

SYS: J. K. Rowling, Paula Hawkins, etc.

Answer: J.K.Rowling

Your evaluation for this answer would be "0".

3. For the answer that contains place entities, for example,

USR: Where was Harry Truman born?

Answer: Lamar, Missouri, USA

Missouri can also be conserved as a correct answer, but USA is definitely a wrong answer.

4. Please note that the answer strings returned by system would contain words that were not necessarily exactly the same as the suggested answer and answers that look similar may differ in their meanings. Please search the information online (through Google.com, Bing.com etc.) about either the question or the answer to confirm that your evaluation is correct.

USR: What city is Disney in?

SYS: Paris

Suggested Answer: Tokyo

Your evaluation should be “1”

6.4 Result analysis

6.4.1 Result filter

Though it's hard to control the quality of the final result for the crowdsourcing low-pay method and especially each person has different opinions for each answer[], there are some results that can be filtered to improve the accuracy of the evaluation. First those answers in wrong format will be filtered, for example, the result with "yes" instead of number "1". And if the answer provided is not complete or redundant, for example, there are 5 questions to be evaluated but the evaluation of the worker provides less than 5 numbers or more than 5 numbers, will be rejected and the hit will be assigned to other workers. For further filtering, according to the statistics provided by MTurk, each worker's approval rates in last 7 days and last 30 days are showing in the result file, which can also be a clue to filter out the results from the workers with low approval rate.

6.4.2 Result and coding

The result from the MTurk is a csv format file, which contains the hit content, the worker's statistics and the evaluation result from the workers. The experiment has been completed in two days. 90 approved results with 18 HITs and 5 assignments for each HIT are obtained for further analysis. Then the result is analyzed for quantitative computation by coding. When the workers don't generate the evaluation agreement for an answer returned by the system, a final judgment is required based on the results from the workers. 5 metrics with different credits distribution are used to decide the final evaluation for each answer when the disagreement appears and the evaluation result for the simulated QA system using the 5 metrics is shown in Table 4. Metric1 is the majority vote, which means the final evaluation is determined by the results with more than 2 workers' agreement. Metric2 assesses the evaluation would be "1" only when there are more than 3 workers believe that the answer is correct. Metric3 judged the final assessment would be "1" only when all the workers believe the answer is correct, otherwise the final evaluation would be "0". Metric4 gives more confidence on the workers who think the answer is correct, which gives "1" when more than 1 worker judge the answer is correct. Metric5 set the result to be "1" as long as there is some worker believe the answer is correct.

Table4 Metrics and accuracy for the system using the metric

	Metric1	Metric2	Metric3	Metric4	Metric5
Number of evaluation "1" from workers for final judgment "1"	>2	>3	>4	>1	>0
Accuracy	74%	62%	35%	81%	93%

Besides the question answering success evaluation for the system, the efficiency cost and the quality cost evaluation metrics and the corresponding score for the simulated system (rank from 1 to 5) are also shown as Table5 by coding.

Table5 Efficiency cost and quality cost evaluation

Total number of user/system turns	157
Total number of system turns	79
Average number of user/system turns per task	8

Total elapse time	60 minutes
Average elapsed time per task	3.33 minutes
Get all the wanted information	3.47
Clarity	3.96
Willingness to use system again	4.08
Naturalness	3.52
User satisfactory	3.81

6.4.3 Evaluation of the evaluation from MTurk

To evaluate the reliability and feasibility, the correct evaluation set is developed from evaluation result by the TREC assessors and online searches. The accuracy of the evaluation using each metric is shown as Table 6, which shows the majority vote metric achieved 92.2%(71 out of 77 evaluations are correct) is the most accurate evaluation method evaluation from the MTurk. However, 6 of the answers returned by the system are not correctly evaluated using the majority vote algorithm, shown as Table 7. The questions and answers returned by the system, the suggested answer, the evaluation from MTurk and the correct evaluation are described in the table.

Table 6 accuracy of the evaluation metrics

	Metric1	Metric2	Metric3	Metric4	Metric5
Accuracy of the evaluation	92.2%	85.7%	66.2%	87.0%	75.3%

Table7 Question and answers with wrong evaluation

ID	Conversation between user and system	Suggested answer	Evaluation from MTurkers	Correct evaluation
1	*USR: What holiday are kumquat trees associated with?* *SYS: Lunar New Year*	Chinese New year	Wrong(3 of 5)	Correct
2	*USR: What is the largest meat-eating marsupial?* *SYS: Sarcophilus harrisii*	Tasmanian devil	Wrong(4 of 5)	Correct
3	*USR: What is his academic specialty?* *Answer: Economist*	Economy	Correct(3 of 5)	Wrong
4	*USR: What is the non-stop distance record established by the Boeing 777?* *SYS:11,664 miles*	11,664 nautical miles	Correct(4 of 5)	Wrong
5	*USR: In what state was the Sago mine?* *SYS: Virginia*	West Virginia	Correct(3 of 5)	Wrong
6	*USR: At which university does he teach?* *Answer: MIT*	Krugman was previously a professor of economics at MIT, and later at Princeton University. He retired from MIT on 2015.	Correct(4 of 5)	Wrong

From Table 7, the first two responses that are not correctly evaluated are involved synonym concepts. “Chinese New year” and “Lunar New Year”, “Tasmanian devil” and “Sarcophilus harrisii” refer to the same entities respectively, which can only be known by some information searches online for some workers. The third answer by the system is similar in both format and content with the suggested answer. However, economist belongs to a vocation type instead of a specialty, which should be judged as wrong response. Though the fourth and the fifth answer is partly same with the suggested answer, the internal meanings differ a lot. And the last conversation queries bout the present tense and the answer indicates that he retired, so the answer can not be treated as right. The analysis for the incorrect evaluations indicates that the workers on MTurk are likely to generate wrong answers when it involves synonym concepts, similar meaning with wrong format, partly same but different entity, and different tenses.

However, the result shows the reliability of the results from MTurk with 92.2% accuracy. It indicates that workers can judge most of the responses correctly with careful reading of the instructions and online searches even for some tricky answers. For example,

USR: "which newspaper does Paul Krugman write?"
SYS: "the times"
Suggested Answer: "the New York times"
Evaluation from MTruk: "0"

Even though "the times" and "the New York times" seem to be related and similar to some degree, they are two different magazines. The workers can also provide the correct evaluations as the guidance shown in the instructions about the unit and the place entity. For example,

USR: "How much did a poster of the winning candidate cost?"
SYS: "280"
Suggested answer: "EGP280"
Evaluation from MTurk:"0"
USR: "Where is the Boeing 777 built?"
SYS: "U.S"

Suggested Answer: "Everett, Wash., USA."

Evaluation from MTurk: "0"

The workers can also judge correctly when the system return an answer with omission or abbreviation than the suggested answer, for instance, Toyota (Toyota Motor Corp.), MSHA(Mine Safety and Health Administration). They can also identify similar meanings with different expressions, for example, "attacked by an animal" and "mauled by a dog". The result shows that the evaluation using MTurk can be a realistic method to evaluate the performance of the system, since the result shows that the workers can extract the answers from the conversation and give a reasonable assessment for the response through careful experiment design and specific guidance.

7. Conclusion and future work

This project surveyed the evaluation methods for interactive QA systems and presents the systems should be evaluated in terms of abilities examination and performance assessment. It demonstrates capabilities and functions that are needed for interactive QA to be evaluated with testing questions. In addition, the performance evaluation, which measures question answering success and dialogue cost in quantitative scores, combines the evaluation techniques from the TREC and the evaluation metrics for conversational agents. The designed evaluation method uses MTurk as platform to ask workers to interact with the system for given questions and employ evaluation tasks after the completion of their interaction. A series of experiments were conducted on the MTurk using simulated data, and the evaluation of the workers' assessment is analyzed. The result indicates the evaluation from MTurk could be a reliable and appropriate method, but it also offers limitations, such as they tend to generally evaluate the synonyms incorrectly.

As future work, special techniques and better algorithms for quality control, for example, more robust final evaluation determination method when there is disagreement by workers, can improve the robustness of the evaluation. In addition, the experiment of the evaluation for complex scenarios-based questions can be conducted in the future research.

8.Reference

[1] Konstantinova, Natalia, and Constantin Orasan. "Interactive question answering." Emerging Applications of Natural Language Processing: Concepts and New Research (2012): 149-169.

[2] Ferrucci, David, et al. "Building Watson: An overview of the DeepQA project."*AI magazine* 31.3 (2010): 59-79.

[3] Dang, Hoa Trang, Diane Kelly, and Jimmy J. Lin. "Overview of the TREC 2007 Question Answering Track." *TREC*. Vol. 7. 2007.

[4] Salime Sadat Shahraini and Morteza Zahedi "A Language-Independent Interactive Question Answering System" Int. J. Rev. Life. Sci., 5(10), 2015, 961-965

[5] Callison-Burch, Chris. "Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk." *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 2009.

[6] Kelly, Diane, et al. "Questionnaires for eliciting evaluation data from users of interactive question answering systems." *Natural Language Engineering*15.01 (2009): 119-141.

[7] Kelly, Diane, et al. "User-centered evaluation of interactive question answering systems." *Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006*. Association for Computational Linguistics, 2006.

[8] Small, Sharon, et al. "HITIQA: an interactive question answering system a preliminary report." *Proceedings of the ACL 2003 workshop on Multilingual summarization and question* answering-Volume 12. Association for Computational Linguistics, 2003.

[9] Quarteroni, Silvia, and Suresh Manandhar. "Designing an interactive open-domain question answering system." Natural Language Engineering 15.01 (2009): 73-95.

[10] Toney, Dave, et al. "An Evaluation of Spoken and Textual Interaction in the RITEL Interactive Question Answering System." *LREC*. 2008.

[11] Harabagiu, S., Hickl, A., Lehmann, J., & Moldovan, D. (2005). Experiments with interactive question-answering. In Proceedings of the 43rd annual meeting on association for computational linguistics (p. 205-214). Ann Arbor, Michigan.

[12] Kato, Tsuneaki, et al. "Woz simulation of interactive question answering."*Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006*. Association for Computational Linguistics, 2006.

[13] Chai, Joyce Y., Tyler Baldwin, and Chen Zhang. "Automated performance assessment in interactive QA." *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2006.

[14] http://start.csail.mit.edu/index.php

[15] http://live.ailao.eu/

[16] https://www.evi.com/

[17] https://www.wolframalpha.com/

[18] Baudiš, Petr. "YodaQA: a modular question answering system pipeline."*POSTER 2015-19th International Student Conference on Electrical Engineering*. 2015.

[19] Voorhees, E. (2003). Overview of the TREC 2003 Question Answering Track. In Voorhees, E. M. and Buckland, L. P., editors, Proceedings of the Twelfth Text REtrieval Conference (TREC), pages 54–68, Department of Commerce, National Institute of Standards and Technology.

[20] O'Shea, J., Z. Bandar, and K. Crockett, Systems Engineering and Conversational Agents, in Intelligence-Based Systems Engineering, A. Tolk and L. Jain, Editors. 2011, Springer Berlin Heidelberg. p. 201-232.

[21] Latham, A., Crockett, K. & Mclean, D. (2014) An adaptation algorithm for an intelligent natural language tutoring system. Computers & Education, 71, 97-110.

[22] http://www.alicebot.org/

[23] http://www.cleverbot.com/

[24] Walker, M. A., Litman, D. J., Kamm, C. A. & Abella, A. (1997) PARADISE: A framework for evaluating spoken dialogue agents. Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. 271-280.

[25] G. Semeraro, H. H. K. Andersen, V. Andersen, P. Lops, and F. Abbattista, "Evaluation and validation of a conversational agent embodied in a bookstore," Universal Access: Theoretical Perspectives,Practice and Experience. Lecture Notes in Computer Science, 2615, 2003, pp. 360-371.

[26] B. A. Shawar, and E. Atwell, "Different measurements metrics to evaluate a chatbot system," Proc. of the 2nd Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing, 2007.

[27] R. Rzepka, Y. Ge, and K. Araki, "Naturalness of an utterance based on the automatically retrieved commonsense," Proc. of Nineteenth IJCAI, 2005.

[28] Hung, Victor, et al. "Towards a method for evaluating naturalness in conversational dialog systems." *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*. IEEE, 2009.

[29] Voorhees, Ellen M., and Dawn M. Tice. "Building a question answering test collection." *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2000.