World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:4, No:6, 2010

A Black-box Approach for Response Quality Evaluation of Conversational Agent Systems

Ong Sing Goh, Cemal Ardil, Wilson Wong, Chun Che Fung

***Abstract*—** The evaluation of conversational agents or chatterbots question answering systems is a major research area that needs much attention. Before the rise of domain-oriented conversational agents based on natural language understanding and reasoning, evaluation is never a problem as information retrieval-based metrics are readily available for use. However, when chatterbots began to become more domain specific, evaluation becomes a real issue. This is especially true when understanding and reasoning is required to cater for a wider variety of questions and at the same time to achieve high quality responses. This paper discusses the inappropriateness of the existing measures for response quality evaluation and the call for new standard measures and related considerations are brought forward. As a short-term solution for evaluating response quality of conversational agents, and to demonstrate the challenges in evaluating systems of different nature, this research proposes a black-box approach using observation, classification scheme and a scoring mechanism to assess and rank three example systems, AnswerBus, START and AINI.

***Keywords*—**Evaluation, conversational agents, Response Quality, chatterbots

I. INTRODUCTION

THE common expectation in conversational agent systems is to be able to provide responses to questions in natural language interface by finding the correct answer from some sources (e.g. web pages, plain texts, knowledge bases), or by generating explanations in the cases of failure to locate relevant or accurate answers. Unlike information retrieval applications such as web search engines, the goal is to find a specific answer [1], rather than flooding the users with documents or best-matching passages as in most of the current information retrieval systems. With the increase in the number of online information seekers, strong demands for automated question answering systems have risen accordingly.

The problem of question answering can be approached from different dimensions [2]. Generally, conversational agent systems on question answering can be categorized into two groups based on their approaches. The first is question answering based on simple natural language processing and information retrieval. The second approach is question answering based on natural language understanding and reasoning. Table I summarizes the characteristics of the two approaches with respect to different dimensions. Some of the well known systems from the first approach are Webclopedia [3], AnswerBus [4] and MULDER [5]; while examples of question answering systems from the second approach are the work in biomedicine[6], system for weather forecast [7] WEBCOOP[8, 9] in tourism, AINI[10, 11] in medicine and legal domains, and START[12, 13] in multimedia information system.

TABLE I
CHARACTERISTICS OF THE TWO APPROACHES IN QUESTION ANSWERING

Dimensions	Question answering based on simple natural language processing and information retrieval	Conversational Agents question answering based on natural language understanding and reasoning
Technique	Syntax processing, named-entity tagging and information retrieval	Semantic analysis or higher, and reasoning
Source	Free-text documents	Knowledge base
Domain	Open-domain	Domain-specific
Response	Extracted snippets	Synthesized responses
Question	Questions using wh-words	Questions beyond wh-words
Evaluation	Use existing information retrieval metrics	N/A

With reference to Table I, unlike other dimensions of problem in question answering, evaluation is the most poorly defined. As evaluation is an important dimension, the lack of standards has resulted in benchmarking the success of any proposed question answering based systems a difficult task. The evaluation of question answering systems for non-dynamic responses has been largely reliant on the use of (TREC) corpus. It is easy to evaluate systems in which there is a clearly defined answer, however, for most natural language questions there is no single correct answer [14]. For example, only the question answering systems based on simple natural language processing and information retrieval like AnswerBus that have the corpora and test questions readily available can use recall and precision as evaluation criteria.

Evaluation can turn into a very subjective matter especially when dealing with different types of natural language systems in different domains. It gets more difficult to evaluate systems based on natural language understanding and reasoning like START and AINI, as there is no baseline or comparable systems in certain domains. Besides, developing a set of test questions is a complicated task because unlike the open-domain evaluations, where test questions can be mined from question logs like Encarta, no question sets are at the disposal for domain-oriented evaluations. Furthermore, due to the dynamic nature of the responses, there is no right or wrong

Manuscript received April 2006.
Ong Sing Goh is with the Murdoch University, Perth, Western Australia (e-mail: os.goh@murdoch.edu.au).
Cemal Ardil is with the National Academy of Azerbaijan, Baku, Azerbaijan (e-mail: cemalardil@gmail.com)
Wilson Wong is with National Technical College University of Malaysia, 75450, Melaka, Malaysia (e-mail: wilson@kutkm.edu.my).
Chun Che Fung is with the Murdoch University, Perth, Western Australia (e-mail: l.fung@murdoch.edu.au).

answer as there are always responses to justify the absence of an answer. For other domain-oriented question answering systems, the task of evaluating the system is not that straightforward and is usually a controversial issue.

II. AINI'S CONVERSATION ENGINE

The ability of computers to converse with users in natural language would arguably increase their usefulness and flexibility even further. Research in practical dialogue systems, while still in its infancy, has matured tremendously in recent years [15] [16]. Today's dialogue systems typically focus on helping users complete a specific task, such as information search, planning, event management, or diagnosis. Recent advances in Natural Language Processing (NLP) and Artificial Intelligence (AI) in general have approached this dream world to the point where it mixes with reality. Several known futurists believe that computers will reach capabilities comparable to human reasoning and understanding of languages by 2020 [17].

knowledge model, multimodal human-computer communication interface and multilevel natural language query, communicates with one another via TCP/IP can be used. AINI is a conversation agent designed by the authors that is capable of having a meaningful conversation with users who interact with her. This is a combination of natural language processing and multimodal communication. A human user can communicate with the developed system using typed natural language conversation. The embodied conversation agent system will reply text-prompts or Text-to-Speech Synthesis together with appropriate facial-expressions. For the purposes of this research, the application area chosen for designing the conversation agent is primarily grounded in an ability to communicate based upon scripting and/or artificial intelligence programming in the field of legal domain.

As shown in Figure. 1, AINI adopts a hybrid architecture that combines the utility of knowledge bases model, multimodal interface and multilevel natural language query.

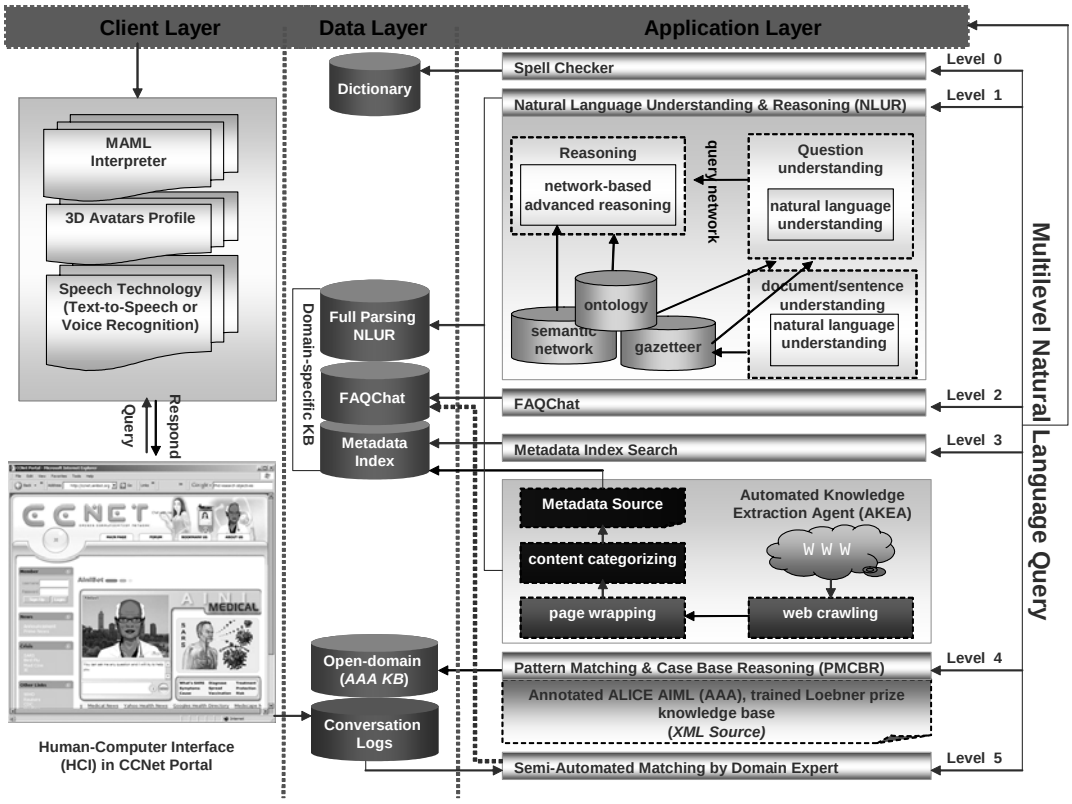

Fig. 1 AINI's Conversational Agent Architecture

This research project involves the establishment of an embodied conversational agent, called Artificial Intelligent Neural-network Identity (AINI) [11] chatterbot as the basic architecture. An AINI chatterbot is a computer program designed to simulate an intelligent conversation with one or more human users via auditory or textual methods. AINI is different from other systems in the sense that it moves away from task-oriented dialogue while many other systems only appear to be intelligent by interpreting the human input prior to providing a response. Our real-time prototype relies on distributed agent architecture designed specifically for the Web. A software agent, such as the conversation engine,

Given a question, AINI first performs question analysis by extracting pertinent information to be used in query formulation, such as the Noun Phrases and Verb Phrases. AINI employs an Internet three-tier, thin-client architecture that may be configured to work with any web application. It comprises of a data server, application and client layers. This Internet specific architecture offers a flexible solution to the unique implementation requirements of the AINI system. The data server layer serves as storage for permanent data required by the system, where the legal knowledge bases are stored. These databases are Dictionary, Domain-Specific and conversation logs. The dictionary is *ispell* which was first

deployed on TOPS-20 systems at the MIT-AI lab[1]. Domain-Specific database is extracted by the Automated Knowledge Extraction Agent (AKEA)[18]. These web-enabled databases are accessible via the SQL query standard for database connectivity using MySQL database.

The application server layer handles the processing of logic and information requests. Here, one or more application servers are configured to compute the dialogue logic through the hybrid approach multilevel natural language query algorithm as shown in Fig 1. Recently in the field of AI, researchers are debating whether bottom-up or top-down approach can be best used to model human brain. Mentalese or `language of thought' and conceptual representation support the idea of a top-down approach [19]. However, the MIT Cog Robot Team fervently supports the bottom-up approach when modeling the human brain [20]. The top-down approach seems to be a good model to explain how humans use their knowledge in a conversation. After much literature search, we concluded that in the field of NLP, it seems that the top-down approach is by far the best approach. Therefore, we use the top-down approach as our natural language query. As shown in Fig. 1, our top-down natural language query approach consists of 6 levels of queries, namely Spell Checker, Full-discourse Natural Language Understanding and Reasoning (NLUR), FAQChat, Metadata Index Search, Pattern Matching and Case Base Reasoning (PMCBR) and Semi-Automated Machine Learning Approach [21].

The user interface resides in the thin-client layer and is completely browser based, employing Multimodal Agent Markup Language (MAML) interpreter or Microsoft SAPI to handle the users interface. MAML is a prototype multimodal markup language based on XML that enables animated presentation agents or avatars. It involves a talking virtual lifelike 3D agent character that is capable of involving in a fairly meaningful conversation. The conversation engine is Web-based and is implemented with an architectural open-source practice by employing PHP, Perl scripting language, Apache Server and knowledge base stored in a MySQL server.

III. DOMAIN KNOWLEDGE MODEL IN AINI'S CONVERSATIONAL SYSTEM

Another significant difference between this research and other conversational agents is the domain knowledge model. Dahlbäck and Jönsson [16] stressed that the Domain Model represents the structure of the knowledge which comprises a subset of general world knowledge. In our research, the domain model is the taxonomy of knowledge related to the topic of the presentation, or XML-like metadata model. This will reduce the workload of the author to predict every input typed by the user. Instead, this allows the author to put more effort on scripting conversation within a specified domain or conversation Domain-Specific.

We believe that the ultimate conversational human-computer interface uses and requires different kinds of approaches. Therefore, we have been working to develop a domain knowledge model for building conversation and interactive systems. For example, according to S. Kshirsagar and N. Magnenat-Thalmann [17], having a small conversation about the weather requires a lot less resources than a philosophical discussion about the meaning of life. In our research, we defined our conversation system as a collective specific conversation units; every unit handles a specific conversation between user and computer. In our case, Domain-Specific knowledge base is extracted by AKEA from the online news articles from ZDNet[2].

Domain is one of the dimensions that determines the focus or direction of a conversational system. An Open-Domain will practice techniques based on probabilistic measures and has a wider range of information source. For a system that focuses on certain domains, it is more likely that the techniques are more logic-based and well-founded, with relatively limited sources as compared to an Open-Domain. A domain-oriented conversational system deals with questions under a Domain-Specific environment, and can be seen as a richer approach. This is because natural language processing systems can exploit domain knowledge and ontologies. Advanced reasoning such as providing explanations for answers, generalizing questions, etc is not possible in Open-Domain systems. Open-Domain question answering systems need to deal with questions about nearly everything and it is very difficult to rely on ontological information due to the absence of wide and yet detailed world knowledge. On the other hand, these systems have much more data to exploit in the process of extracting the answers. Therefore this leads to research on the responses quality evaluation of conversational agents system as presented in this paper.

IV. EXISTING METRICS FOR QUESTION ANSWERING

Evaluation is one of the important dimensions in question answering systems which involve the process of assessing, comparing and ranking to measure the progress in the field of interest. Surprisingly, literatures on evaluation are relatively sparse given its state of importance and are mostly available in the form of evaluating general natural language systems. One of the factors may be due to the bad reputation earned during the early days of evaluating natural language systems [22]. Nonetheless, we will attempt to highlight several works that strive for a standard metric or formal framework in evaluating general natural language understanding systems.

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

where
Precision, P = correct answers produced/answers produced = PC/(PC+PI)
Recall, R = correct answers produced/total possible correct answers = PC/(PC+NC)
where PC and PI are

	Correct	*Incorrect*
Produced	*PC*	*PI*
Not produced	*NC*	*NI*

β = parameter indicating the importance of recall to precision. (e.g. if β was set to 5, then evaluator is trying to indicate that recall was five times as important as precision)
α = inverse of β

Fig. 2 Requirements for F-measure

[1] http://www.mit.edu/afs/sipb/project/sipb-athena/src/ispell/

[2] http://www.zdnet.com

The most notable evaluation for question answering systems has to be the question answering track in the TREC evaluation [23]. Evaluation in TREC is essentially based on the F-measure to assess the quality of response in terms of precision and recall. Such mode of evaluation is tailored for all question answering systems based on shallow natural language processing and information retrieval like AnswerBus where information retrieval is the backbone of such systems. To enable F-measure, a large query and document ensemble is required where the document collection is manually read and tagged as correct or incorrect for one question out of a list of predefined answers as shown in Fig. 2.

There are several inherent requirements with F-measure that makes it inappropriate for evaluation of domain-oriented question answering systems based on understanding and reasoning:

- assessments should average over large corpus or query collection;
- assessments have to be binary where answers can only be classified as correct or incorrect; and
- assessments would be heavily skewed by corpus, making the results not translatable from one domain to another.

The first requirement actually makes it extremely difficult to evaluate domain-oriented systems like START and AINI due to the absence of large quantity of domain-related documents collection. Besides, like most other systems based on understanding and reasoning, AINI uses knowledge base as information source instead of a large document collection, making F-measure impossible. For modern-day question answering systems, the large corpus requirement has been handled by TREC.

Secondly, responses produced in question answering systems based on understanding and reasoning such as START and AINI are descriptive in nature and thus, cannot be merely classified into correct or incorrect. Moreover, the classification is manually done by human experts, making the results extremely subjective and non-definite.

Lastly, most systems based on understanding and reasoning actually have domain portability as their main aim. A system normally starts out as domain-restricted and then slowly grows or progresses to other domains. The characteristic of F-measure that skews according to domains makes it inappropriate for evaluation of such systems.

There are also other measures but are mostly designed for general tasks related to natural language processing like translation, database query, etc. Facemire [24] proposes that a simple number scale be established for the evaluation of natural language text processing systems. This metric is to be based on human linguistic performance, taken as 0 to 1, and is an average of four subcomponents which are the size of the lexicon, the speed and accuracy of the parse and the overall experience of the system. The author has also oversimplified matters by equating the ability of understanding to mere sentence parsing. Also, the use of the criteria of speed and accuracy in parsing has limited the metric's ability to keep the pace of technological advances. As the computing strength increases in terms of hardware and software, the factor of speed and accuracy can no longer be discriminative enough to separate one system from another.

Unlike the previous approach, a general model is provided by Guida & Mauri [25] that acts as a basis of a quantitative measure for evaluating how well a system can understand natural language. However, such model only provides for half of the actual ability required to generate high-quality responses. Such general model is inadequate for more specific application of natural language understanding like question answering.

Srivastava & Rajaraman [26] have also attempted to devise an experimental validation for intelligence parameters of a system. The authors concluded that intelligence of a question answering system is not a scalar value but rather, a vector quantity. The set of parameters that define intelligence are knowledge content of a system, efficiency of a system and correctness of a system. In this approach, the answerer is an entity that has the answer in mind and the questioner must attempt to guess what is in the mind of the answerer with the help of the least number of questions. The questioner that manages to figure out the answer using the minimal number of questions is considered as intelligent. Hence, to apply this approach for evaluating the quality of responses in a standard setting of question and answering is not possible.

Allen [27] and Nyberg & Mitamura [28] have also suggested a type of black-box evaluation where we evaluate a system to see how good it is at producing quality or desirable answers. Diekema et al. [29] further characterize the black-box evaluation and suggested that systems can be evaluated on their answer providing ability that includes measures for answer completeness, accuracy and relevancy. The authors also state that evaluation measures should include more fine grained scoring procedures to cater answers to different types of question. The authors give examples of answers that are explanations or summaries or biographies or comparative evaluations that cannot be meaningfully rated as simply right or wrong. We consider this black-box approach as comprehensive in assessing how well question answering systems produce responses required by users and how capable are these systems in handling various types of situations and questions. Despite the merits of the evaluation approach, none of the authors provide further details on the formal measures used for scoring and ranking the systems under evaluation.

V. CONSIDERATIONS FOR ALTERNATIVE MEASURE

Question answering in conversational agents is a multi-dimensional research area and with the rise of using natural language understanding and reasoning in question answering system as suggested by Maybury [30], there is a growing need to look for a common evaluation metric. Thus, to evaluate systems based on natural language understanding and reasoning for response quality, an alternative measure that is agreed upon by members of the community in the field is required. The new method should be capable of handling information in the knowledge domain, and classification of response extending beyond logical correct or incorrect.

The new measure must take into consideration the three crucial factors related to the inherent nature of question answering systems based on natural language understanding and reasoning:

- systems based on understanding and reasoning use knowledge base as information source and there are no numerical measurements for such unit of information. In systems where information retrieval is their backbone, the unit of information has always been a document. It is commonly known that *"out of the three documents retrieved, two answers the question"*. However, we cannot state that "*two out of the three meaning or knowledge produced answers the question*"; and
- responses generated by such systems are subjective; there is a need for a scale whereby everyone in the research community of understanding and reasoning agrees on for measuring the quality of responses. For example, a scale where everyone can actually refer to and say that a response to a question is 45% correct is needed.
- preparation of the questions set must put into consideration that the peer systems under evaluation are from the same domain. For example, there are two systems to be evaluated where one supports the biological disease domain while the other handles agricultural domain. How are we going to craft or prepare the questions in a way to prevent any controversy concerning the fairness of the evaluation?

Therefore, there is an urgent need of new and non-refutable metrics that can be used for the formal evaluation of the new question answering systems. Until then, the validity of comparing and evaluating question answering systems based on understanding and reasoning will always be a topic of research. A formal evaluation is crucial to promote further research interest and growth in this area, as well as providing a framework for benchmarking research in this area.

VI. BLACK-BOX APPROACH FOR QUALITY EVALUATION

In this paper, we present a short-term solution to answer the call for standardized metrics for evaluating response quality: a black-box approach through observation, and classification with a scoring mechanism. This black-box approach is based on the work of Allen [27], Nyberg & Mitamura [28], Diekema *et al.* [29] as discussed in previous sections for evaluating response quality. We further refine this approach by proposing a response classification scheme and a scoring mechanism. To demonstrate this approach, we have selected three question answering systems that represent different levels of response generation complexity namely, AnswerBus, START and AINI.

To begin with, this black-box approach requires a set of questions that can sufficiently examines the response generation strength of all systems under evaluation. For this purpose, we prepared 45 questions of various natures on the legal domain. These questions will be used to probe the systems and the actual responses are gathered for later analysis.

For this approach, we propose a classification scheme that consists of categories to encompass all possible types of response from all systems under evaluation. This scheme consists of three category codes and was designed based on the quality of responses as perceived by general users and is not tied down to any implementation details of any systems. This makes the scheme generally applicable to evaluation of all question answering systems with different approaches. Under this scheme, we define two general categories *BQ_θ* and *LQ_θ*, where *θ* is systems initial, which represent the best and lowest quality response for each system respectively. There is also a dynamic category *Oj_θ*, where *j* is an integer, which represents other evaluation-specific criteria.

Evaluators can create as many new categories as required by the types of system under evaluation using *Oj_θ*. The *Oj_θ* category not only makes this scheme expandable but also dynamic because as technology progresses, the response generation capability of the systems may increase and in such cases, evaluators can define evaluation-specific categories. For this evaluation, we define *O1_θ* for quality of response in the event of no answer and *O2_θ* for responses that suggest possible spelling mistakes. In this evaluation, the initials for AnswerBus, START and AINI are *A*, *S* and *N* respectively.

Next, using these codes, the evaluators will try to observe and classify each response into one of the categories. The classification is done based on the manual observation by evaluators who are guided by the criteria of each category. For example, if the evaluator comes across a response that is generated by system *α* and the response appears to be an uninformative attempt to notify the user that no valid answer can be found, then we can classify that response as *O1_α*. This is to say that system *α* generates uninformative response in the event of no answer.

We first look at the responses generated by AnswerBus, and from the nature of their responses, we can group them into three relevant categories:

- BQ_A: responses with some elements that meet the questions' requirements (e.g. question 10 in Table II);
- LQ_A: responses with completely irrelevant information (e.g. question 37 in Table II);
- O1_A: uninformative responses in the event of no answers (e.g. question 43 in Table II); and
- O2_A: not relevant.

TABLE II
PART OF THE RESPONSES BY ANSWERBUS

	Question	Response
question 10	List the companies who sued Microsoft?	• The United States Department of Justice (DOJ) and the 19 US states who successfully sued Microsoft over Anti-Trust violations have requested that the US court spilt Microsoft in to two separate companies. • None of these companies sued Microsoft when it began giving away a free codec in Windows Media Player, because they could not be sure their technology was in there.
question 37	Which judge presided the ruling of the case by RealNetworks against Microsoft?	• Grimaldi, Microsoft Judge Says Ruling at Risk, Wash. • A month ago, lawyers for the accused ISV ElcomSoft asked a US District Judge Ronald Whyte, who presided over the Sun-Microsoft Java suit, to drop the case because the DMCA is unconstitutional.
question 43	Who presided the closing of	I'm not sure what you mean.

	the spamming case against Howard Carmark?	

Following that, we move on to study the responses generated by START. Similarly, based on the nature of the responses produced, we can group the responses into three relevant categories:

- BQ_S: responses that provide accurate and direct answers to questions;
- LQ_S: not relevant;
- O1_S: uninformative responses in the event of no answers (e.g. question 38 in Table III); and
- O2_S: responses that suggest possible spelling mistake (e.g. question 29 in Table III).

TABLE III
PART OF THE RESPONSES BY START

	Question	Response
question 29	When was the closing of the caset against Microsoft?	The word CASET may be misspelled
question 38	Who chaired the closing of the case by RealNetworks against Microsoft?	Unfortunately, I don't know who chaired the closing of the case against Microsoft.

Lastly, based on the nature of the responses produced by AINI, we group the responses into three relevant categories:

- BQ_N: responses that provide accurate and direct answers to questions (e.g. question 43 in Table IV);
- LQ_N: not relevant;
- O1_N: informative responses in the event of no answers (e.g. question 32 in Table IV); and
- O2_N: responses that suggest possible spelling mistake (e.g. question 29 in Table IV).

TABLE IV
PART OF THE RESPONSES BY AINI

	Question	Response
question 29	When was the closing of the caset against Microsoft?	There are some spelling errors in the question. When was the closing of the **caset** against Microsoft
question 32	When was the filing of the case against Microsoft by RealNetworks?	There is no such filing event involving RealNetworks as plaintiff.
question 43	Who presided the closing of the spamming case against Howard Carmark?	Attorney General Eliot Spitzer New York chaired the resolution of the case

After classification of the responses is done, a scoring mechanism is used to determine responses which system provides the best overall quality. A pair-wise relative comparison is performed and points are assigned based on superiority of responses of the same category. If there are n systems under evaluation, then there should be ${}_nC_2 = k$ pairs. Let λ_i represents the pair of system θ_{xi} and θ_{yi}. To perform the scoring, a table is constructed as shown in Table V where the column header represents all the $\lambda_1, \lambda_2, \ldots, \lambda_k$ pairs. The row header will consists of the two general categories BQ_θ and LQ_θ and other evaluation-specific categories Oj_θ.

TABLE V
TEMPLATE FOR SCORING MECHANISM

Category	λ_1		λ_2		…	λ_k	
	θ_{x1}	θ_{y1}	θ_{x2}	θ_{y2}		θ_{xk}	θ_{y1k}
BQ_θ							
LQ_θ							
Oj_θ							
Total							

Then for every λ_i, we compare BQ_θ_{xi} with BQ_θ_{yi}, LQ_θ_{xi} with LQ_θ_{yi} and other Oj_θ_{xi} with Oj_θ_{yi}. The rules for superiority comparison and assigning of score are as follows:

- if the description of the responses for θ_{xi} is better than θ_{yi} under a particular category, then θ_{xi} is assigned with 1 and θ_{yi} is assigned with 0 under the same category;
- if the description of the responses for θ_{xi} is inferior compared to θ_{yi} under a particular category, then θ_{xi} is assigned with 0 and θ_{yi} is assigned with 1 under the same category; and
- if the description of the responses for θ_{xi} is the same as θ_{yi} under a particular category, then both θ_{xi} and θ_{yi} are assigned with 0 under the same category.

After filling up all the cells in the score table, summation of scores for every θ_{xi} and θ_{yi} under all categories is performed.

Here are a few examples to demonstrate the working behind the scoring mechanism. The best quality responses of AnswerBus, BQ_A have the possibility of containing irrelevant elements, whereas responses generated by START are always correct and directly answer the questions. Due to this, the best quality responses from START, which belongs to BQ_S, are a level higher than the best quality responses of AnswerBus, BQ_A. Hence, for the pair "*START vs. AnswerBus*", START will be assigned with one point. In the case of ties, like other categories O_1S and O_1A which demonstrate the same quality of responses in the event of no answers, no points will be given for either side of the pair "*START vs. AnswerBus*". Consider another example where the responses from O_2S, which attempt to alert the users of possible spelling mistake, make START an additional level higher than AnswerBus. This provides START with another additional point in the pair "*START vs. AnswerBus*". The comparison will be done on all the three systems, giving us three possible pairs.

From Table VI, we can observe that AnswerBus has the total score of 0 + 0 = 0, AINI with the total score of 3 + 1 = 4 and START with the total score of 0 + 2 = 2.

TABLE VI
SCORING TABLE FOR QUALITY EVALUATION USING PAIR-WISE RELATIVE COMPARISON

Category	AnswerBus vs. AINI		START vs. AINI		START vs. AnswerBus	
	AnswerBus	AINI	START	AINI	START	AnswerBus
BQ_ LQ_	0	1	0	0	1	0
O_1	0	1	0	1	0	0
O_2	0	1	0	0	1	0
Total	**0**	**3**	**0**	**1**	**2**	**0**

VII. IMPLICATIONS AND VALIDITY OF THE RESULTS

From the total scores of the three systems based on the sample questions in the experiment, AINI ranked first with 4 points, followed by START with 2 points and lastly, AnswerBus with 0 point. This makes the quality of responses generated by AINI relatively better as compared with START and AnswerBus. The condition is assuming that the evaluators' observations and classifications are consistent throughout, and the set of questions used for evaluation is exhaustive enough to trigger all possible responses. In the case of new systems being added to the evaluation, the observation, classification and scoring process needs to be redone. The approach of evaluating the response quality through observation, classification and a scoring mechanism has revealed to us that the lack or addition of components has great impact on the response quality. Table VII gives a summary of components implemented by each of the three systems evaluated for this specific pilot study. It should be noted that the results will be dependent on the components and features to be evaluated and the set of questions being used.

TABLE VII
UNDERSTANDING AND REASONING COMPONENTS IN ANSWERBUS, START AND AINI

components and other features	AnswerBus	START	AINI
sentence parsing	√	√	√
named-entity recognition	√		√
relation extraction		√	√
anaphora resolution			√
semantic unification			√
semantic representation		√	√
traceable answer discovery		√	√
explanation on failure			√
dynamic answer generation		√	√

For instance, one of the criteria that have contributed to the higher score of AINI is the capability of the system in generating dynamic responses to suit the various anomalous situations. For example, useful responses can be dynamically generated by AINI to cater the condition when no answers are available. This ability can be attributed to the inclusion of the two advanced reasoning components namely explanation on failure and dynamic answer generation. Such useful responses can help the users to clear any doubts related to the actual state of the knowledge base. This is obviously a desirable trait for a question answering system. Table VIII shows how each category of responses are achieved through the different approaches being used by the question answering systems that encompass diverse components of information retrieval, natural language understanding and reasoning.

TABLE VIII
RELATION BETWEEN QUALITY OF RESPONSES AND COMPONENTS IN QUESTION ANSWERING

Categories of responses	AnswerBus	START	AINI
responses with some elements that meet the questions' requirements, while the rest are irrelevant materials.	achieved through mere sentence parsing and information retrieval	n/a	n/a
responses that provide accurate and direct answers to questions	n/a	achieved through higher-level of natural language understanding and reasoning	achieved through higher-level of natural language understanding and reasoning
quality of responses in the event of no answers	uninformative due to the lack of advanced reasoning	uninformative due to the lack of advanced reasoning	informative due to the use of advanced reasoning
responses that suggest possible spelling mistake	n/a	achieved through additional linguistic feature	achieved through additional linguistic feature

Initial results have indicated that AINI is comparatively better than the other two systems. However, it is expected that concerns may arise on the nature and domain of the questions. One would speculate that the evaluation is inclined towards AINI because the question set is prepared in the same domain as AINI, which is legal document. In the case of AnswerBus, it was quoted "*AnswerBus is an open-domain question answering*…"[31]. START also claimed that their system is capable of handling many domains based on their statement "*our system answers millions of natural language questions about places (e.g., cities, countries, lakes, coordinates, weather, maps, demographics, political and economic systems), movies (e.g., titles, actors, directors), people (e.g., birth dates, biographies), dictionary definitions, and much, much more…*" by Katz *et al*. [32] Hence, it is reasonable to expect that the two systems should be able to handle the questions in the legal domain.

Secondly, queries may arise concerning to the nature of the questions. It may be asked that whether the questions and the evaluation are inequitable towards START and AnswerBus because the questions used to evaluate vary greatly and cover beyond wh-questions. In this aspect, we consider the focus and the aim of this evaluation is to assess and rank the systems based on the quality of responses generated. It may not be valid if the systems were ranked merely based on wh-questions. It is believed that benchmarking for question answering systems has to progress with time by considering various state-of-the-art factors.

At the time of writing, the authors are inclining to consider another alternative by involving researchers or domain experts in the evaluation process. The idea is to extend the proposal in this paper by inviting questions from the participants or users. The decisions on the correctness and accuracy of the answers will be determined by independent judges based on the process as described in our proposed system. Over time, the set of question will grow but will be limited to a certain number so as to reduce the workload for the judges. This set of questions will be applied to all the systems under consideration. During the evaluation period, decisions from the independent observers or judges will be accumulated. This approach will give a fairer and unbiased system over a predefined period of time. This might also be considered as a comparison to the current Turing Test judging method. Further

development on this suggestion will be proposed in future papers.

VIII. Conclusion

In this paper, we have highlighted the increasing need for standard metrics to assess and measure the quality of responses produced by conversational agent systems based on different approaches and domains. Based on the fact that more researchers in conversational agents in question answering are adopting natural language understanding and reasoning, question answering systems will be more diverse in nature than before. Domains supported by the system will vary, and the responses produced can never be simply graded as just correct or wrong anymore. Following this, we have presented a short-term solution for the evaluation of the quality of responses in the form of a black-box approach through classification and a scoring mechanism using pair-wise relative comparison. To demonstrate the approach, we have also presented the data and results obtained through an evaluation performed on three different systems.

We see this initial work as a foundation for evaluating the quality of responses from question answering systems of different techniques and domains. This could also act as a first step to look for a unify method in this area as suggested in the latter part of this paper. It is hoped that this work will bring to the attention of many researchers and to arouse more interest in this area. There is a need for more focused research in the area of question answering evaluation for systems that are increasingly diverse in many aspects like domain, responses, techniques, etc. It is expected that with the establishment of a fair and unbiased evaluation system, conversational agent systems will advance further to become more versatile and robust in future applications.

References

[1] J. Lin, V. Sinha, B. Katz, K. Bakshi, D. Quan, D. Huynh, and D. Karger, "What Makes a Good Answer? The Role of Context in Question Answering," presented at.the 9th International Conference on Human-Computer Interaction, 2003.

[2] L. Hirschman and R. Gaizauskas., "Natural Language Question Answering: The View from Here," *Natural Language Engineering*, vol. 7, pp. 275-300, 2001.

[3] U. Hermjakob, "Parsing and Question Classification for Question Answering," presented at the ACL Workshop on Open-Domain Question Answering, 2001.

[4] Z. Zheng, "Developing a Web-based Question Answering System," presented at.the 11th International Conference on World Wide Web, 2002a.

[5] C. Kwok, D. Weld, and O. Etzioni, "Scaling Question Answering to the Web," *ACM Transactions on Information Systems*, vol. 19, pp. 242-262, 2001.

[6] P. Zweigenbaum, "Question Answering in Biomedicine," presented at the 10th Conference of the European Chapter of the Association for Computational Linguistics, 2003.

[7] H. Chung, K. Han, H. Rim, S. Kim, J. Lee, Y. Song, and D.Yoon, "A Practical QA System in Restricted Domains," presented at the ACL Workshop on Question Answering in Restricted Domains, 2004.

[8] F. Benamara, "Cooperative Question Answering in Restricted Domains: the WEBCOOP Experiment," presented at the ACL Workshop on Question Answering in Restricted Domains, 2004.

[9] F. Benamara and P. Saint-Dizier, "Advanced Relaxation for Cooperative Question Answering," in.*New Directions in Question Answering*: MIT Press, 2004.

[10] W. Wong, O. S. Goh, M. I. Desa, and S. Sahib, "Online Cyberlaw Knowledge Base Construction Using Semantic Network," presented at International Conference on Computational Intelligence for Modelling, Control and Automation, Rhodes, Greece, 2004.

[11] O. S. Goh, C. C. Fung, and M. P. Lee, "Intelligent Agents for an Internet-based Global Crisis Communication System," *Journal of Technology Management and Entrepreneurship*, vol. 2, pp. 65-78, 2005.

[12] B. Katz and J. Lin, "START and Beyond.," presented at the 6th World Multiconference Systemics, Cybernetics and Informatics, 2002.

[13] B. Katz, "Annotating the World Wide Web using Natural Language," presented at the 5th Conference on Computer Assisted Information Searching on the Internet., 1997.

[14] D. Moldovan, M. Pasca, M. Surdeanu, and S. Harabagiu., "Performance Issues and Error Analysis in an Open-Domain Question Answering System," presented at the 40th Annual Meeting of the Association for Computational Linguistics, 2002.

[15] J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, "Towards conversational human-computer interaction," *AI Magazine*, vol. 22, 2001.

[16] J. Cassell, "Embodied Conversation: Integrating Face and Gesture into Automatic Spoken Dialogue Systems," in *Spoken Dialogue Systems*, Luperfoy, Ed.: MIT Press, to appear.

[17] R. J. Lempert, S. W. Popper, and S. C. Bankes, *Shaping the next one hundred years: new methods for quantitative, long-term policy analysis*. Santa Monica, CA.: RAND, 2003.

[18] O. S. Goh and C. C. Fung, "Automated Knowledge Extraction from Internet for a Crisis Communication Portal," in *First International Conference on Natural Computation*. Changsha, China: Lecture Notes in Computer Science (LNCS), 2005, pp. 1226-1235.

[19] J. A. Fodor, *Elm and the Expert: An Introduction to Mentalese and Its Semantics*: Cambridge University Press, 1994.

[20] R. A. Brooks, "The Cog Project: Building a Humanoid Robot," presented at The 1st International Conference on Humanoid Robots and Human friendly Robots, Tsukuba, Japan, 1998.

[21] O. S. Goh, A. Depickere, C. C. Fung, and K. W. Wong, "Top-down Natural Language Query Approach for Embodied Conversational Agent," presented at the International MultiConference of Engineers and Computer Scientists 2006, Hong Kong, 2006.

[22] M. King, "Evaluating Natural Language Processing Systems," *Communications of the ACM.*, vol. 39, pp. 73-79, 1996.

[23] E. Voorhees, "Overview of TREC 2003," presented at the 12th Text Retrieval Conference, 2003.

[24] J. Facemire, "A Proposed Metric for the Evaluation of Natural Language Systems," presented at the IEEE Energy and Information Technologies in the Southeast,, 1989.

[25] G. Guida and G. Mauri, "A Formal Basis for Performance Evaluation of Natural Language Understanding Systems.," *Computational Linguistics.*, vol. 10, pp. 15-30, 1984.

[26] A. Srivastava and V. Rajaraman, "A Vector Measure for the Intelligence of a Question-Answering (Q-A) System," *IEEE Transactions on Systems: Man and Cybernetics.*, vol. 25, pp. 814-823, 1995.

[27] J. Allen, *Natural Language Understanding*: Benjamin/Cummins Publishing, 1995.

[28] E. Nyberg and T. Mitamura, "Evaluating QA Systems on Multiple Dimensions," presented at the Workshop on QA Strategy and Resources, 2002.

[29] A. Diekema, O. Yilmazel, and E. Liddy., "Evaluation of Restricted Domain Question-Answering Systems," presented at the ACL Workshop on Question Answering in Restricted Domains, 2004.

[30] M. Maybury, "Toward a Question Answering Roadmap," presented at the AAAI Spring Symposium on New Directions in Question Answering, 2003.

[31] Z. Zheng, "AnswerBus Question Answering System," presented at the Conference on Human Language Technology, 2002b.

[32] B. Katz, S. Felshin, and J. Lin, "The START Multimedia Information System: Current Technology and Future Directions," presented at the International Workshop on Multimedia Information Systems, 2002.

Associate Professor Ong Sing Goh is a researcher and a member of the Game & Simulation Research Group based at Murdoch University. He received his B. A. Ed. (Hons) from University Science Malaysia, Master of Science in Machine Translation from University of Manchester Institute of Science and Technology and PhD in IT from Murdoch University. He was a founding Chairman of the Center of Excellence for Computer Languages (CCL) at Multimedia University, Malaysia. Prior to his current appointment, he was a Associate Professor, Director of University Press and the Chairman of the Center for Artificial Intelligence and Modeling at the National Technical College University of Malaysia from August 2002. His main research interest is in the development of intelligent agent, natural language processing and speech technology to facilitate graceful human-computer interactions, conversational robot and mobile services. He is the author of more than 50 peer-reviewed scientific journal, books and conference papers. He has led and worked on research grants funded by Malaysian Government's Intensified Research in Priority Areas and Malaysia Technology Development Corporation. He is a member of IEEE Computer Society, Member of Malaysia Invention and Design Society (MINDS, Scientific Member of World Enformatika Society and member of International Association of Engineers.

Wilson Wong is currently a full-time PhD student at the University of Western Australia. He received his B.IT(HONS) from Multimedia University, Malaysia and M.Sc(ICT) from the National Technical University College of Malaysia in the field of Natural Language Understanding. His research interests include natural language processing, and knowledge representation and reasoning.

Associate Professor Chun Che Fung (Lance) is a member of IEEE since 1992. He received his PhD degree from the University of Western Australia in 1994, a Master of Engineering degree, and a Bachelor of Science degree in Maritime Technology with First class honors from the University of Wales, Institute of Science and Technology in 1982 and 1981 respective. He also received a Graduate Diploma in Business from Curtin University of Technology in 2000.

Since 2003, Lance is employed as Associate Professor at the School of Information Technology, Murdoch University, Western Australia. Prior to this appointment, he has served as a lecturer and senior lecturer at Curtin University of Technology, Perth from 1989 to 2003. He also lectured at the Singapore Polytechnic from 1982 to 1988.

Dr Fung has served in various executive positions in the IEEE WA Section and several technical chapters during the past 10 years. He received the IEEE Third Millennium Award in 2000 in recognition of for his services. He has also contributed to many national and international conferences as member of the organizing and program committees. He has published over 150 articles and papers in book chapters, journal transactions and conference proceedings. His research interests are in the practical applications of artificial intelligent systems. He is also a member of Institute of Engineers Australia (IEAust) and Australian Computer Society (ACS).