World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:4, No:5, 2010

Response Quality Evaluation in Heterogeneous Question Answering System: A Black-box Approach

Goh Ong Sing, Cemal Ardil, Wilson Wong, and Shahrin Sahib

Open Science Index, Computer and Information Engineering Vol:4, No:5, 2010 publications.waset.org/13286/pdf

***Abstract*—**The evaluation of the question answering system is a major research area that needs much attention. Before the rise of domain-oriented question answering systems based on natural language understanding and reasoning, evaluation is never a problem as information retrieval-based metrics are readily available for use. However, when question answering systems began to be more domains 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 achieve higher quality responses The research in this paper discusses the inappropriateness of the existing measure for response quality evaluation and in a later part, the call for new standard measures and the related considerations are brought forward. As a short-term solution for evaluating response quality of heterogeneous systems, and to demonstrate the challenges in evaluating systems of different nature, this research presents a black-box approach using observation, classification scheme and a scoring mechanism to assess and rank three example systems (i.e. AnswerBus, START and NaLURI).

***Keywords*—**Evaluation, question answering, response quality.

I. Introduction

THE common idea in question answering system is to be able to provide responses to questions in natural language format by finding the correct answer from some sources (e.g. web pages, plain texts, knowledge bases), or by generating explanations in the case of failures. Unlike information retrieval applications, like web search engines, the goal is to find a specific answer [9], rather than flooding the users with documents or even best-matching passages as most information retrieval systems currently do. With the increase in the number of online information seekers, the demand for automated question answering systems has rise accordingly.

The problem of question answered can be approached from different dimension [7]. Generally, question answering systems can be categorized into two groups based on the approach in each dimension. 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 respects to the dimensions in question answering. Some of the well known systems from the first approach are Webclopedia [8], AnswerBus [22] and MULDER [14], while examples of question answering systems from the second approach are the work in biomedicine [24], system for weather forecast [3], WEBCOOP [1][2] in tourism, NaLURI [18][19][20] in Cyberlaw and multimedia information system, START [10][11].

TABLE I
Characteristics of the Two Approaches in Question Answering

Dimensions	Question answering based on simple natural language processing and information retrieval	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-oriented
Response	Extracted snippets	Synthesized responses
Question	Questions using wh-words	Questions beyond wh-words
Evaluation	Use existing information retrieval metrics	N/A

Referring back to Table I, unlike other dimensions of problem in question answering, evaluation is the most poorly defined. As this is as important as other dimensions, the lack of standards in evaluation has resulted in benchmarking the success of any proposed question answering based systems. 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 [16]. 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 NaLURI, as there is no baseline or comparable systems in certain domains. Besides, developing a set of test

Manuscript received August 2005.
Goh Ong Sing is with the Murdoch University, Perth, Western Australia (e-mail: osgoh88@gmail.com).
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).
Shahrin Sahib is with Technical College University of Malaysia, 75450, Melaka, Malaysia (e-mail: shahrinsahib@kutkm.edu.my).

Open Science Index, Computer and Information Engineering Vol:4, No:5, 2010 publications.waset.org/13286/pdf

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

II. Existing Metrics for Question Answering

Evaluation is one of the important dimensions in question answering which involve the process of assessing, comparing and ranking to measure the progress in the field. Surprisingly, the 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 [13]. 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. 1 Requirements for F-measure

The most notable evaluation for question answering has to be the question answering track in the TREC evaluation [21]. 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 as shown in Fig. 1.

There are several inherent requirements with F-measure that makes it inappropriate for evaluations 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 and 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 NaLURI due to the absence of large quantity of domain-related documents collection. Besides, like most other systems based on understanding and reasoning, NaLURI 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 system based on understanding and reasoning such as START and NaLURI are descriptive in nature and thus, cannot be merely classified into correct and 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 has domain portability as their main aim by starting out as a domain-restricted system and slowly grows or moves 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 [5] 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 1.0, and is the simple 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 move on with time. 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, general model is provided by Guida & Mauri [6] that acts as a basis of a quantitative measure for evaluating how well a system can understand natural language. But how well a system can understand natural language only provides for half of the actual ability required to generate high-quality responses. Hence, such general model is inadequate for more specific application of natural language understanding like question answering.

Srivastava & Rajaraman [17] 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 is 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

Open Science Index, Computer and Information Engineering Vol:4, No:5, 2010 publications.waset.org/13286/pdf

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 [25] and Nyberg & Mitamura [26] have also suggested a type of black-box evaluation where we evaluate a system to see how good it is at producing the quality or desirable answers. Diekema et al. [4] 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.

III. Considerations for Alternative Measure

Question answering 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 [15], 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/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 uses 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?

All in all, only with the presence of new and non-refutable metrics can the formal evaluation for this new question answering approach be performed. 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.

IV. Black-Box Approach for Quality Evaluation

In this paper, we present a short-term solution to answer the call for standardise 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 [25], Nyberg & Mitamura [26], Diekema *et al.* [4] 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 level of response generation complexity namely AnswerBus, START and NaLURI.

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 prepare 45 questions of various natures on the Cyberlaw domain. These questions will be used to probe the systems and the actual responses are gathered for later use. Details of the questions and responses for the three systems are available in [18].

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 detail of any systems. This makes the scheme generally applicable to all evaluations of 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, and one 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 systems 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 systems may increase and in such cases, evaluators can define evaluation-specific categories. For this evaluation, we define $O1_\theta$ for quality of response in the event of no answer and $O2_\theta$ for response that suggest possible spelling mistake. In this evaluation, the initials for AnswerBus, START and NaLURI 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

Open Science Index, Computer and Information Engineering Vol:4, No:5, 2010 publications.waset.org/13286/pdf

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 $O1_\alpha$. 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 the spamming case against Howard Carmark?	I'm not sure what you mean.

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 NaLURI, 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 NALURI

	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 from which system are of 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;

Open Science Index, Computer and Information Engineering Vol:4, No:5, 2010 publications.waset.org/13286/pdf

- 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, NaLURI 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. NaLURI		START vs. NaLURI		START vs. AnswerBus	
	AnswerBus	NaLURI	START	NaLURI	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**

V. IMPLICATIONS AND VALIDITY OF THE RESULTS

From the total scores of the three systems, NaLURI ranked first with 4 points, followed by START with 2 points and lastly, AnswerBus with 0 points. This makes the quality of responses generated by NaLURI relatively ***better*** compare with START and AnswerBus. The condition is to assume 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. Please refer to Table VII for the summary of components implemented by each of the three systems evaluated.

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

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

For instance, one of the criteria that have contributed to the higher score of NaLURI 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 NaLURI 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 neatly shows how each of the categories of responses are achieved through the different approach towards question answering that implements diverse components in information retrieval, natural language understanding and reasoning.

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

Categories of responses	AnswerBus	START	NaLURI
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

After having concluded that NaLURI is comparatively better than the other two systems, skeptical thoughts may

Open Science Index, Computer and Information Engineering Vol:4, No:5, 2010 publications.waset.org/13286/pdf

arise. Firstly, thoughts may arise concerning to the domain of the question. People may question that the evaluation is inclined towards NaLURI because the question set is prepared in the same domain as NaLURI, which is Cyberlaw. But, what is the domain of START and AnswerBus? "*AnswerBus is an open-domain question answering…*" [23] while START is capable of handling many domains based on the 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*. [12].

Hence, the authors do not see any problem in START and AnswerBus handling Cyberlaw questions. Secondly, thoughts may arise concerning to the nature of the question. People may question that the evaluation is inequitable towards START and AnswerBus because the nature of the questions used to evaluate vary greatly and cover beyond wh-questions. But, we would like the readers to recall that the aim of this evaluation is to assess and rank systems of any approach based on the quality of responses generated. How can we rank these systems if we merely use wh-questions, knowing that given the present state of question answering technology, handling wh-questions is no more a challenge? Hence, benchmark for question answering systems has to progress with time by considering various state-of-the-art factors instead of dwelling in the past.

VI. Conclusion

In this paper, we have highlighted the increasing need for standard metrics to assess and measure the quality of responses produced by systems of different approaches and domain. By considering the fact that as more researchers 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 very different systems.

We see that this initial work has at least lay the 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. Hopefully, this work will bring to the attention of many researchers and to bring more interest in this area. There is a need for more focus research in the area of question answering evaluation for systems that are increasingly diverse in many aspects like domain, responses, techniques, etc.

References

[1] Benamara, F., Cooperative Question Answering in Restricted Domains: the WEBCOOP Experiment. In Proceedings of the ACL Workshop on Question Answering in Restricted Domains, 2004.

[2] Benamara, F. & Saint-Dizier, P., *Advanced Relaxation for Cooperative Question Answering*. In New Directions in Question Answering. MIT Press, 2004.

[3] Chung, H., Han, K., Rim, H., Kim, S., Lee, J., Song, Y. & Yoon, D., A Practical QA System in Restricted Domains. In *Proceedings of the ACL Workshop on Question Answering in Restricted Domains*, 2004.

[4] Diekema, A., Yilmazel, O. & Liddy, E., Evaluation of Restricted Domain Question-Answering Systems. In *Proceedings of the ACL Workshop on Question Answering in Restricted Domains*, 2004.

[5] Facemire, J., A Proposed Metric for the Evaluation of Natural Language Systems. In *Proceedings of the IEEE Energy and Information Technologies in the Southeast*, 1989.

[6] Guida, G. & Mauri, G., A Formal Basis for Performance Evaluation of Natural Language Understanding Systems. *Computational Linguistics, 10*(1):15-30, 1984.

[7] Hirschman, L. & Gaizauskas, R., Natural Language Question Answering: The View from Here. *Natural Language Engineering, 7*(4):275-300, 2001.

[8] Hermjakob, U., Parsing and Question Classification for Question Answering. In *Proceedings of the ACL Workshop on Open-Domain Question Answering*, 2001.

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

[10] Katz, B. & Lin, J., START and Beyond. In Proceedings of the 6th World Multiconference Systemics, Cybernetics and Informatics, 2002.

[11] Katz, B., Annotating the World Wide Web using Natural Language. In Proceedings of the 5th Conference on Computer Assisted Information Searching on the Internet, 1997.

[12] Katz, B., Felshin, S. & Lin, J., The START Multimedia Information System: Current Technology and Future Directions. In *Proceedings of the International Workshop on Multimedia Information Systems*, 2002.

[13] King, M., Evaluating Natural Language Processing Systems. *Communications of the ACM, 39*(1):73-79, 1996.

[14] Kwok, C., Weld, D. & Etzioni, O., Scaling Question Answering to the Web. *ACM Transactions on Information Systems, 19*(3):242-262, 2001.

[15] Maybury, M., Toward a Question Answering Roadmap. In Proceedings of the AAAI Spring Symposium on New Directions in Question Answering, pp. vii-xi, 2003.

[16] Moldovan, D., Pasca, M., Surdeanu, M. & Harabagiu, S., Performance Issues and Error Analysis in an Open-Domain Question Answering System. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[17] Srivastava, A. & Rajaraman, V., A Vector Measure for the Intelligence of a Question-Answering (Q-A) System. *IEEE Transactions on Systems_Man and Cybernetics, 25*(5):814-823, 1995.

[18] Wong, W., *Practical Approach to Knowledge-based Question Answering with Natural Language Understanding and Advanced Reasoning*. Thesis (MSc), Kolej Universiti Teknikal Kebangsaan Malaysia, 2004.

[19] Wong, W., Sing, G. O., Mohammad-Ishak, D. & Shahrin, S., Online Cyberlaw Knowledge Base Construction using Semantic Network. In *Proceedings of the IASTED International Conference on Applied Simulation and Modeling*, 2004a.

[20] Wong, W., Sing, G. O. & Mokhtar, M., Syntax Preprocessing in Cyberlaw Web Knowledge Base Construction. In *Proceedings of the International Conference on Intelligent Agents, Web Technologies and Internet Commerce*, 2004b.

[21] Voorhees, E., Overview of TREC 2003. In Proceedings of the 12th Text Retrieval Conference, 2003.

[22] Zheng, Z., Developing a Web-based Question Answering System. In *Proceedings of the 11th International Conference on World Wide Web*, 2002a.

[23] Zheng, Z., AnswerBus Question Answering System. In *Proceedings of the Conference on Human Language Technology*, 2002b.

[24] Zweigenbaum, P., Question Answering in Biomedicine. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, 2003.

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

[26] Nyberg, E. & Mitamura, T., Evaluating QA Systems on Multiple Dimensions. In *Proceedings of the Workshop on QA Strategy and Resources*, 2002.