Evaluation of Response Quality for Heterogeneous Question Answering Systems

Wilson Wong & Shahrin Sahib
Faculty of Information and Communication Technology, Kolej Universiti Teknikal Kebangsaan Malaysia, Melaka, Malaysia
{wilson, shahrinsahib}@kutkm.edu.my

Ong-Sing Goh
School of Information Technology, Murdoch University, Perth, Western Australia, 6150
osgoh88@gmail.com

Abstract

The research in this paper makes explicit why existing measures for response quality evaluation is not suitable for the ever-evolving field of question answering and following that, a short-term solution for evaluating response quality of heterogeneous systems is put forward. To demonstrate the challenges in evaluating systems of different nature, this research presents a black-box approach using a classification scheme and scoring mechanism to assess and rank three example systems.

1. Introduction

Generally, question answering systems can be categorized into two groups based on the approach in each dimension. The first is question answering based on shallow natural language processing and information retrieval and 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 like Webclopedia [1] and AnswerBus [2], while examples of question answering systems from the second are like WEBCOOP [3] in tourism, NaLURI [4] in Cyberlaw and START [5].

Table I. Characteristics of the two approaches in question answering

Dimension	Shallow natural language processing and information retrieval	Natural language understanding and reasoning
Technique	Syntax processing 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	Beyond wh-words
Evaluation	Information retrieval metrics	N/A

The evaluation of question answering systems for non-dynamic responses has been largely reliant on the 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 [6]. Evaluation can turn into a very subjective matter especially when dealing with different types of natural language systems in different domains due several reasons: no baseline or comparable systems in certain domains, developing test questions is not easy, and dynamic nature of the responses, there is no right or wrong answer as there are always responses to justify the absence of an answer.

2. Existing Metrics for Question Answering

The most notable evaluation for question answering has to be the question answering track in the TREC evaluation [7]. Evaluation in TREC assesses the quality of response in terms of precision and recall, and is well-suited for question answering systems based on shallow natural language processing and information retrieval like AnswerBus. There are several inherent requirements that make such evaluation inappropriate for 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.

There are also other measures but are mostly designed for general tasks related to natural language processing like translation, database query, etc. [8] proposes that a simple number scale be established for the evaluation of natural language text processing systems. This metric is to be based on the simple average of four things: size of the lexicon, the speed and accuracy of the parse and the overall experience of the system. The author oversimplified matters by equating the ability of understanding to mere sentence parsing and 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.

Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)
0-7695-2415-X/05 $20.00 © 2005 **IEEE**

Unlike the previous, general model is provided by [9] 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.

[10] and [12] 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. [13] 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 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 this evaluation approach, none of the authors provide further details on the formal measures used for scoring and ranking the systems under evaluation.

3. Black-box Approach for Evaluation

In this paper, we present an innovative measure 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 [10], [12] and [13] 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 [4].

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 as shown in Table II 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.

Table II: Categories in black-box approach

category	notation	desc.	purpose
general	*BQ_t*	*t* is systems initial	represents best and lowest quality response for each system
	LQ_t		
dynamic	*Oj_t*	*j* is an integer	represents other evaluation-specific criteria

Dynamic category allows evaluators to create as many new categories as required by the types of systems under evaluation. The *Oj_t* 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_t* for quality of response in the event of no answer and *O2_t* for response that suggest possible spelling mistake. In this evaluation, the initials for AnswerBus, START and NaLURI are *A*, *S* and *N* respectively.

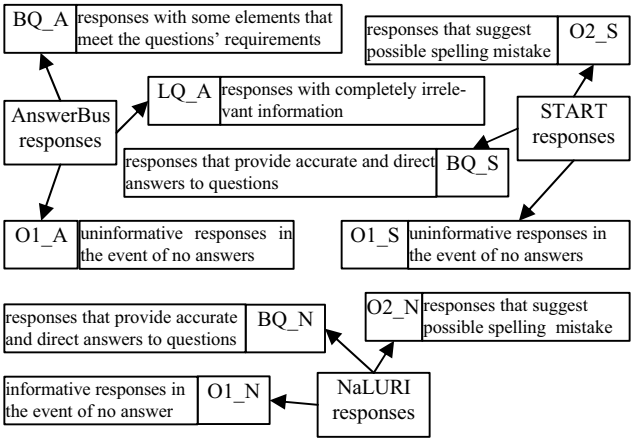

Figure 1: Grouping of responses into categories

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 *a* 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_a*. This is to say that system *a* generates uninformative response in the event of no answer. From the nature of the responses generated by the three systems, we can group them into relevant categories

Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)
0-7695-2415-X/05 $20.00 © 2005 **IEEE**