Framework for the Formulation of Metrics for Conversational Agent Evaluation

Mohammed Kaleem, Omar Alobadi, James O'Shea, Keeley Crockett
Intelligent Systems Research Group
Department of Computing, Mathematics and Digital Technology
Manchester Metropolitan University
Manchester, United Kingdom
E-mail: (m.kaleem, j.d.oshea, k.crockett)@mmu.ac.uk, oalobadi@yahoo.com

Abstract

The evaluation of conversational agents is an area that has not seen much progress since the initial developments during the early 90's. The initial challenge faced when evaluating conversational agents is trying to formulate the metrics that need to be captured and measured in order to gauge the success of a particular conversational agent. Although frameworks exist they overlook the individual objectives of modern conversational agents which are much more than just question answering systems. This paper presents a new framework that has been utilised to formulate metrics to evaluate two conversational agents deployed in two significantly different contexts.

Keywords: Conversational agents, dialog agents, evaluation, evaluation metrics

1 Introduction

This paper illustrates the application of a software quality model in Conversational Agent (CA) evaluation. According to the IEEE Glossary of Software System Engineering Terminology (IEEE, 2000), quality is defined as the degree to which a system, a component, or a process meets customer or user needs or expectations. Roy and Graham (2008), posit that the quality of software is measured primarily against the degree to which requirements, such as correctness, reliability and usability are met. The factors that affect quality are termed as quality attributes or metrics. There are different categorizations of quality metrics. Roy and Graham (2008), further state that quality metrics can be categorized into two broad groups: metrics that can be directly measured (e.g. performance) and metrics that can be indirectly measured (e.g. usability). These metrics can be translated into objective and subjective metrics respectively. In order to build a successful dialogue system, data is needed on how users behave and their perceptions when interacting with the system (Skantze and Hjalmarsson, 2013). Recent work in the field has produced CAs in very diverse applications (Keeling et al., 2004 Alobaidi et al., 2013; Latham et al., 2014), therefore the evaluation of such systems has to suit the individual goals of the application domain. As such the evaluation metrics cannot be generalized in to a "one size fits all" evaluation framework. The weakness with existing frameworks is that they fail to identify the individual evaluation metrics that need to be gauged, metrics that are unique to the goal of the CA developed.

This paper is structured as follows: Section 2 outlines the concept of CAs and presents the existing work conducted in the field of CA evaluation. Section 3 details the proposed framework for formulating CA evaluation metrics and details two case studies where the framework was used to evaluate two different CAs. Section 4 discusses the results of the case studies. Section 5 details the conclusion drawn from the case studies.

2 Background

2.1 Conversation Agents

The term "Conversational Agent" (CA) is interpreted in different ways by different researchers; however the essence of CAs is natural language dialogue between the human and an application running on a computer (O'Shea et al., 2011). Recent developments in the field of CAs have utilized complex artificial intelligence techniques in order to facilitate a rich goal driven conversation with the user. These types of CAs have been applied in a wide array of contexts such as a CA Help Desk: responding to employee or customer questions related to complex processes or procedures (Lester et al., 2004; Kaleem et al., 2014), Website Navigation/Concierge: guiding customers to relevant portions of complex websites (Shimazu, 2002), Guided Selling: providing answers and guidance in the sales process, particularly for complex products being sold to novice customers (Keeling et al., 2004), Education: known as Conversational Intelligent Tutoring Systems (CITS) (Alobaidi et al., 2013; Latham et al., 2014) and HR Bully and Harassment Help System (Latham et al., 2010).

2.2 Conversation Agent Evaluation

According to Martinez et al. (2008), it is quite difficult to evaluate dialogue systems. In addition to the lack of evaluation standards within the dialogue community, it is difficult to find performance figures from real world applications that can be extrapolated to other systems or be accepted worldwide, as all of them are directly related to one specific dialogue system. Although CA/Dialogue system evaluation frameworks exist, these frameworks are dated and moreover they generalize the metrics tested between individual systems, therefore overlooking the increasingly complex developments in the field of CAs. CAs are now more than just question answer systems.

An early example of evaluating the success of dialog based software is the Turing test. The Turing test (Turing, 1950) was primarily aimed at making a human believe that they were speaking to another human, when in fact they were speaking to a computer program. This approach however is not suitable to gauge the effectiveness or usability of a modern goal orientated conversational agent as the intrinsic nature behind the two applications are completely different.

Existing CA evaluation frameworks such as PARADISE were devised almost 20 years ago and while they were suited to evaluate the CAs of their time they are not entirely suitable to evaluate modern day conversational agents which are much more technologically advanced,

taking advantage of artificial intelligence to achieve much more diverse goals like tutoring (Latham et al., 2014) and offering specialist advice (Latham et al., 2010).

A substantial amount of work has been done on evaluating CAs as a whole. The seminal work in this area was done by Walker et al. (1997) who created the PARADISE framework which is a general framework for evaluating spoken dialogue systems. For determining the quality of Spoken Dialogue Systems, several aspects are of interest. Moller et al. (2009), presented a taxonomy of quality criteria. They describe quality as two separate issues consisting of Quality of Service and Quality of Experience. Quality of Service describes objective criteria like dialogue duration or number of turns or utterances it takes to achieve the desired outcome. While these are well-defined items that can be determined easily, Quality of Experience, which describes the user experience with subjective criteria, is a more vague area and without a sound definition, e.g. User Satisfaction.

There is a general agreement on "usability" as the most important performance figure in CAs (Turunen et al., 2006) even more than others widely used like "naturalness" or "flexibility". However functionality may be more important, but without usability the system will not get the chance to demonstrate functionality. Therefore, besides quality and efficiency metrics, automatically logged or computed, subjective tests must also be performed in order to assess the impact of the capabilities of the system on user satisfaction and to get a valuable insight on the shortcomings and advantages of the system (Martinez et al., 2008). According to Silvervarg and Jönsson (2011), the evaluation of CA/dialogue systems is mainly done either by distributing a questionnaire to the users trying to the reveal their subjective assessment of using the dialogue system or by studying the resulting dialogue. Artstein et al. (2009), refer to this as "soft" numbers versus "hard" numbers and propose a "semi-formal" evaluation method combining the two evaluation methodologies. This notion is supported by more recent research conducted by Rauschenberger et al. (2013) who propose a framework to measure user experience and software quality in interactive software applications through User Evaluation Questionnaires (UEQ). They state that the evaluation of interactive software quality falls into two distinct categories, these being "pragmatic quality" and "hedonic quality". Pragmatic quality relates to task orientated quality like task completion effectiveness and efficiency. Hedonic quality is related to non-task orientated aspects like aesthetic impressions and user stimulation. These two categories can be translated into objective measures and subjective measures respectively.

Subjective aspects like user satisfaction are usually determined by using questionnaires with Likert Scale questions (Hone and Graham, 2000; Silvervarg and Jönsson, 2011; Rauschenberger et al., 2013). Objective metrics can be measured through records and logs of the user's dialogue with the CA. These metrics are captured whilst a user is undergoing an evaluation session to achieve a pre-set task. The records/logs are used to capture and store several variables related to the dialogue such as rule fired, similarity strength, user utterance, CA response etc. Based on these captured variables which are stored in the log file, the CA can be evaluated for effectiveness accuracy and robustness, through statistical analysis. The general consensus among researchers in the field from the early days to the present day is that the effectiveness of a CA/Dialogue system should be evaluated through a combination of subjective and objective measures (Alobaidi et al., 2013; O'Shea et al., 2011; Rauschenberger et al., 2013; Walker et al., 1997). This ensures that not only is the effectiveness of the CA's functionality tested but the usability from the user perspective is also tested. It has been established that there is standard set of metrics related to usability and task completion that are to some extent universal for the evaluation for all dialogue systems.

However there is no framework which can be followed to derive the individual metrics that need to be tested in order to evaluate the success of modern day conversational agents that are much more than just general chat applications - they are tutoring users on diverse topics and explaining complex procedures.

As there has been no formal development of the CA evaluation frameworks over the years, alternative existing approaches/evaluation frameworks that can be adopted are software evaluation frameworks that that are utilized to test new software applications in terms of functionality and usability (i.e. objective and subjective metrics).

3 Novel Framework for the Formulation of Evaluation Metrics Suited to Individual CA Goals

As with any engineering discipline, software development requires a measurement mechanism for feedback and evaluation. Measurement is an aid in answering a variety of questions associated with the enactment of any software. It allows the determination of the strengths and weaknesses of the current processes and allows us to evaluate the quality of specific processes and products (Van Solingen et al., 2002). A particular measurement/evaluation is useful only if it helps you to understand the underlying process or one of its resultant products. In turn, recognizing improvement of the process and products can occur only when the project has clearly defined goals for process and products. In other words, you cannot tell if you are going in the right direction until you determine your destination. (Fenton and Pfleeger, 1998).

According to (Fenton and Pfleeger, 1998) an evaluation strategy can be more successful if it is designed with the goals of the project in mind. One such strategy is the Goal Question Metric (GQM) approach, which is based upon the assumption that for an system to be measured in a focused way the goals of the system must be identified first, then those goals can be traced to the questions that are intended to answer those goals operationally. Finally provide a framework for interpreting the questions with respect to the stated goals in to measurable metrics. Thus it is important to make clear, at least in general terms, what the goals of each CA are so that these goals can be quantified whenever possible, and the quantified information can be analyzed as to whether or not the goals are achieved. The GQM approach proposed by Fenton and Pfleeger (1998) provides a framework involving three steps:

GOAL - List the major goals of the system.

QUESTION - Derive from each goal the questions that must be answered to determine if the goals are being met. Questions try to characterize the object of measurement (product, process, resource) with respect to a selected quality issue and to determine its quality from the selected

viewpoint. Once the questions have been developed, the next step involves associating the question with appropriate metrics that will help in answering the question.

METRIC - Decide what must be measured in order to be able to answer the questions adequately. A set of metrics is associated with every question in order to answer it in a quantitative way. The metrics can be classified as either:

Objective: If they depend only on the object that is being measured and not on the viewpoint from which they are taken; e.g., number of versions of a document, staff hours spent on a task, size of a program.

Subjective: If they depend on both the object that is being measured and the viewpoint from which they are taken; e.g., readability of a text, level of user satisfaction. (Fenton and Pfleeger, 1998; Van Solingen et al., 2002).

The GQM model is a top down hierarchical model as illustrated in Figure 1, the top level starts with a goal (specifying purpose of measurement, object to be measured, issue to be measured, and viewpoint from which the measure is taken).

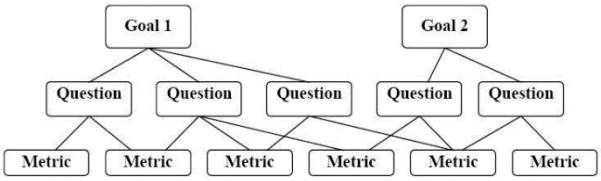

Figure 1: GQM Model

The goal is refined into several questions that usually break down the issue into its major components. Each question is a metric, some of them objective, some of them subjective. The same metric can be used in order to answer different questions under the same goal (Van Solingen et al., 2002).

This new CA evaluation framework was adopted for the evaluation of two novel conversational agent systems, UMAIR (Kaleem et al., 2014) and Abdullah CITS (Alobaidi et al., 2013).

3.1 Application of new Framework for the evaluation of the UMAIR CA

UMAIR was developed to serve as a customer service agent for the National Database and Registration Authority (NADRA) of Pakistan. The objective of UMAIR's development was to guide users through the complex process of ID card and Passport application. UMAIR conversed with the user in Urdu and was the first Urdu CA developed requiring a new CA architecture, using novel language processing techniques and algorithms. UMAIR was evaluated by utilizing proposed framework to formulate which metrics need to be evaluated to determine the success and robustness of the system and its newly developed algorithms. The results of applying the new evaluation framework to establish evaluation metrics for UMAIR are illustrated in Figure 2.

3.2 Application of new Framework for the evaluation of Abdullah CITS

The Abdullah CITS was developed to serve as on online tutor that teaches young children topics related to Islam. The aim of Abdullah was to mimic a human tutor by utilizing several teaching methodologies to deliver the tutorial through conversation with the users. Abdullah employed novel methods of detecting the users' level of knowledge and learning styles in order to adapt the tutorial conversation to suit that individual users' ability and learning style (Alobaidi et al., 2013). One of the aims behind the evaluation of the Abdullah CITS was to verify if Abdullah was an effective tutor. The results of applying the new evaluation framework to establish evaluation metrics for gauging the effectiveness of Abdullah CITS as a tutor are illustrated in Figure 3.

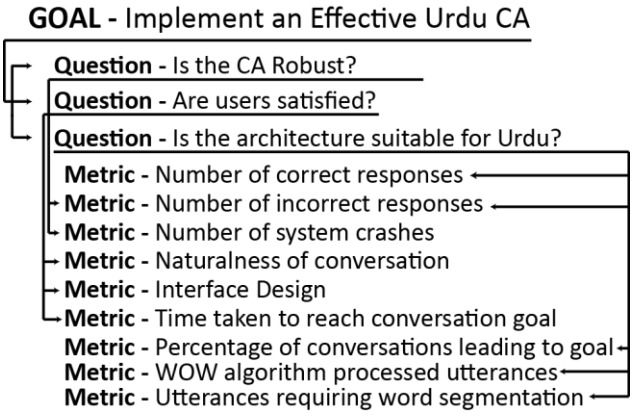

Figure 2 - UMAIR GQM Diagram

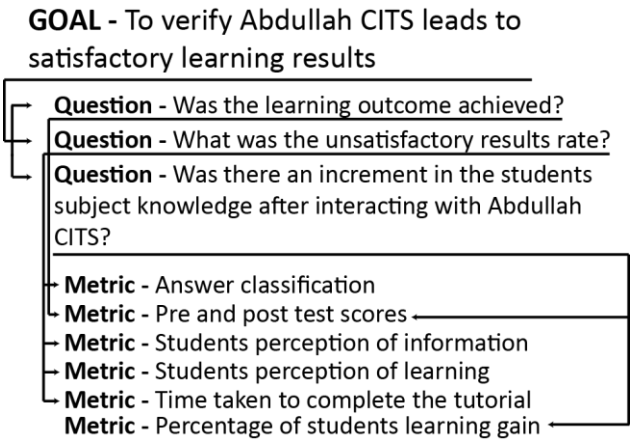

Figure 3 - Abdullah CITS GQM Diagram

4 Discussion

The results show that when the GQM method is applied to formulate the evaluation metrics for two different CAs the metrics derived suit the aims and objectives of that individual CA. It can be seen from the results that when GQM was applied to derive evaluation metrics for the UMAIR CA which is the first Urdu CA developed, the metrics derived are related to gauging the success of the components that make up the architecture of the CA. Whereas when GQM was applied to the Abdullah CITS the evaluation metrics derived were more related to the educational capabilities of the CA rather than measuring the success of the architecture components (i.e. was Abdullah an effective tutor). These metrics were utilized in the evaluation strategies of both the CAs by their respective researchers. Subsequent to the metrics being formulated they can be categorized in to their respective groups (i.e. subjective/objective) based on how the practitioner plans to capture the metrics for evaluation. This is traditionally done either through questionnaires for subjective metrics and some sort of log that captures the conversation and statistics related to the CA architecture and its components.

One of the advantages of the GQM approach is that multiple goals can be defined prior to evaluation and the

metrics can be formulated and categorized, which allows the pre-planning of how the metrics will be captured. One metric may be used to answer more than one question, therefore making the evaluation a more systematic process. Although there may be some overlap in common metrics such as conversation duration/time and conversation length etc. the metrics derived through GQM method are largely related to the development goals of the individual CA.

5 Conclusion

We have reported on the deployment of an adaptable framework for assessing CA quality in two distinctly different contexts. A new CA evaluation framework which is based on existing methods applied in a new context has been devised and tested which addresses the gap in current research related to the development and subsequent evaluation of natural language systems in general. The framework comprises of CA evaluation from an objective as well as subjective perspective in order to give an overall performance related CA evaluation. The proposed framework focuses on evaluating metrics related to the CAs ability to achieve the objective of its development by employing software evaluation methodologies (GQM). This approach allows CAs to be tested on an individual basis, meaning the metrics that are tested from system to system are derived based on the context of the systems implementation, thus allowing the evaluation metrics to be different depending on the development goals of the system being tested. Moreover it becomes easier to pre-determine better evaluation metrics when the proposed framework is used. The framework can be utilized by future research and practitioners to evaluate developed CAs, as the methodology is adaptable to suit individual CA development goals.

6 References

Alobaidi, O. G., Crockett, K. A., O'shea, J. D. & Jarad, T. M. (2013) Abdullah: An Intelligent Arabic Conversational Tutoring System for Modern Islamic Education. Proceedings of the World Congress on Engineering.

Artstein, R., Gandhe, S., Gerten, J., Leuski, A. & Traum, D. (2009) Semi-formal evaluation of conversational characters. *Languages: From Formal to Natural.* Springer.

Fenton, N. E. & Pfleeger, S. L. (1998) *Software metrics: a rigorous and practical approach*, PWS Publishing Co.

Hone, K. S. & Graham, R. (2000) Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering,* 6**,** 287-303.

Ieee (2000) IEEE Recommended Practice for Architectural Description of Software-Intensive Systems. *IEEE Std 1471-2000***,** i-23.

Kaleem, M., O'shea, J. D. & Crockett, K. A. (2014) Word order variation and string similarity algorithm to reduce pattern scripting in pattern matching conversational agents. Computational Intelligence (UKCI), 2014 14th UK Workshop on. IEEE, 1-8.

Keeling, K., Beatty, S., Mcgoldrick, P. & Macaulay, L. (2004) Face Value? Customer views of appropriate formats for embodied conversational agents (ECAs) in online retailing. System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on. IEEE, 10 pp.

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

Latham, A., Crockett, K. A. & Bandar, Z. (2010) A Conversational Expert System Supporting Bullying and Harassment Policies. ICAART (1). 163-168.

Lester, J., Branting, K. & Mott, B. (2004) Conversational agents. *The Practical Handbook of Internet Computing*.

Martinez, F. F., Blázquez, J., Ferreiros, J., Barra, R., Macias-Guarasa, J. & Lucas-Cuesta, J. M. (2008) Evaluation of a spoken dialogue system for controlling a hifi audio system. Spoken Language Technology Workshop, 2008. SLT 2008. IEEE. IEEE, 137-140.

Moller, S., Engelbrecht, K.-P., Kuhnel, C., Wechsung, I. & Weiss, B. (2009) A taxonomy of quality of service and quality of experience of multimodal human-machine interaction. Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on. IEEE, 7-12.

O'shea, J., Bandar, Z. & Crockett, K. (2011) Systems Engineering and Conversational Agents. *In:* TOLK, A. & JAIN, L. (eds.) *Intelligence-Based Systems Engineering.* Springer Berlin Heidelberg.

Rauschenberger, M., Schrepp, M., Cota, M. P., Olschner, S. & Thomaschewski, J. (2013) Efficient measurement of the user experience of interactive products. How to use the user experience questionnaire (ueq). example: spanish language version. *IJIMAI,* 2**,** 39-45.

Roy, B. & Graham, T. N. (2008) Methods for evaluating software architecture: A survey. *School of Computing TR,* 545**,** 82.

Shimazu, H. (2002) ExpertClerk: A Conversational Case-Based Reasoning Tool forDeveloping Salesclerk Agents in E-Commerce Webshops. *Artificial Intelligence Review,* 18**,** 223-244.

Silvervarg, A. & Jönsson, A. (2011) Subjective and objective evaluation of conversational agents in learning environments for young teenagers. 7th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, Barcelona, Spain.

Skantze, G. & Hjalmarsson, A. (2013) Towards incremental speech generation in conversational systems. *Computer Speech & Language,* 27**,** 243-262.

Turing, A. M. (1950) Computing machinery and intelligence. *Mind***,** 433-460.

Turunen, M., Hakulinen, J. & Kainulainen, A. (2006) Evaluation of a spoken dialogue system with usability tests and long-term pilot studies: similarities and differences. INTERSPEECH.

Van Solingen, R., Basili, V., Caldiera, G. & Rombach, H. D. (2002) Goal Question Metric (GQM) Approach. *Encyclopedia of Software Engineering*.

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