International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 10 (2018) pp. 323-330

Received: 19 February, 2018; Accepted: 30 Oct, 2018; Published: 8 Nov, 2018

A Black-box Approach Evaluation on Conversational Agent using Loebner Prize Competition Datasets

Leong Pui Huang[1], Goh Ong Sing[2], Yogan Jaya Kumar[2] and Aditya Pradana[4]

[1] Faculty of Computing and Information Technology, Tunku Abdul Rahman University College (TAR UC),
Jalan Segamat/Labis 85000 Segamat, Johor Malaysia
leongph@tarc.edu.my

[2, 3, 4] Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM)
Jalan Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
[2]*goh@utem.edu.my*, [3]*yogan@utem.edu.my*, [4]*zastiex@yahoo.com*

Abstract: **Conversational Agents, or more commonly known as chatbots have been well-accepted in most enterprises. Chatbots have been widely used in major enterprises to perform the main purpose of customer service in delivering routine and frequently asked questions on behalf of humans. Bearing the goal to create believable conversational agents in providing accurate, reliable and up-to-date information, researchers have practiced natural language processing, machine learning, and deep learning in chatbot creation. In order to create a chatbot which is indistinguishable from human, many researchers have attempted to create human-like chatbots by neglecting the necessity to evaluate the general knowledge of the chatbot before proceeding into the creation of domain-specific knowledge. This paper presents the black-box approach to evaluate the response quality of a developed conversational agent named Tarie through datasets obtained from Loebner Prize Competition. The black-box approach is therefore deemed necessary to be used as the evaluation approach to evaluate the response quality of the conversational agent. The intelligence of the chatbot denotes the correct answers given to the user's queries. The results obtained via the black-box approach is further deduced to indicate whether the chatbot is ready for the domain-specific knowledge creation or to further improve its current general knowledge. This paper presents the black-box approach to evaluate the response quality of a proposed conversational agent named Tarie through datasets obtained from Loebner Prize Competition.**

Keywords: Conversational agent, chatbot, Loebner Prize Competition, black-box approach, datasets, Tarie

I. Introduction

A conversational agent, also known as chatbot refers to the technological interaction or artificial conversation between computer and human. Many different phrases have been used to describe a chatbot such as: dialogue system, conversation system, chatterbot and virtual agent [2]. This new popular trend of online chatting has appeared to provide users the capabilities to chat with the customer service agents anytime and anywhere [1]. In this globalisation era with technology advancements, conversational agent has grown in parallel with high-technology electronic devices such as computer, mobile phone, kiosk, and many more to play an important role to simulate a natural conversation between human and machine. One well-known example of chatbot would be Siri in the iPhone mobile device. This chatbot is capable to provide a new wave on how humans can interact with artificial intelligence implanted in a mobile device.

The dream of artificial conversational agent researchers to create a chatbot which is indistinguishable from human is still vague at present. By keeping the above goal in mind, many researchers have been developing chatbots but neglecting the necessity to evaluate the general knowledge of the chatbot before proceeding into the creation of domain-specific knowledge. General knowledge is regarded as the knowledge about many different things, but not detailed knowledge about one particular subject; for instance, basic arithmetic. It is important for the chatbot to have general knowledge in order to be evaluated as the user is expecting a logical response. For most of the cases, a logical response comes from factual information. If the response is not relevant, the user might feel reluctant to continue chatting with the chatbot. Thus, in this work, we have taken the measure to evaluate the response quality of the chatbot's general knowledge via black-box testing. Domain-specific knowledge on the other hand is regarded as the knowledge of a specific, specialised discipline or field. For instance, every corporate has their own domain-specific knowledge which brands them uniquely among each another. The black-box approach is therefore deemed necessary to be carried out as the evaluation approach to evaluate the response quality of the conversational agent's general knowledge. The intelligence of the chatbot denotes the correct answers given to the users' queries. The results obtained via the black-box approach is further deduced to indicate whether the chatbot is ready for the domain-specific knowledge creation or to further improve its general knowledge.

Countless conversational agents have been built every year, but only a few made it to the top of the chatbot ranking. In order to inspire more researchers to create human-like chatbots, Loebner Prize Competition is introduced. In this competition, the most human-like chatbot will be rewarded. More detailed information regarding Loebner Prize Competition will be discussed in Section IV. Several approaches have been used to evalute the performance of chatbots. However, this paper will be focusing on evaluation via the black-box approach using questions from Loebner Prize Competition as the datasets. In this research, a chatbot named Tarie is developed. The goal of this study is to conduct an evaluation on Tarie to indicate its quality response in conversing via a qualitative assessment known as Black-box testing using datasets obtained from Loebner Prize Competition. Through the evaluation, future enhancement will be applied to the chatbot to increase its intelligence in delivering human-like information.

The goal of this paper is to conduct an evaluation on the chatbot named Tarie to indicate its level of intelligence in conversing via a qualitative assessment known as Black-box testing using datasets obtained from Loebner Prize Competition. The flow of this paper is organised as follows; Section II describes the review of literature followed by Section III which describes the black-box approach. Section IV discusses the Loebner Prize Competition whereas Section V discusses the datasets. The flow of this paper then proceeds with Section VI to discuss the conversational agent named Tarie. Section VII portrays the results and findings and finally the conclusion of the paper is discussed in Section VIII.

II. Related Work

The application of chatbot and its technology has seen a drastic improvement since the very first chatbot was created in 1966; for instance, the attempt to create a novel approach for medical support via chatbot to predict the disease based on the symptoms given by the user [2]. The chatbot is also capable to provide user with a proper cure or treatment. Despite of this, the accuracy of the diseases identified and treatments provided are still an issue. It is therefore important to perform the evaluation of the chatbot. More recently, a counselling chatbot to provide conversation service for mental health care purposes has been introduced [3]. The chatbot is capable of recognising emotions through various training datasets from videos, audios, texts and even images. Their research is then further enhanced to include emotion recognition process to provide better mental health conversations [5]. This research involves high-level natural language process technique to provide an emphatic response to patients which indirectly improves the psychiatric conversation.

Likewise, the research on creating an educational chatbot for visually-impaired users has also been introduced [4]. This chatbot provides the capabilities to converse or interact with users using speech or voice recognition via the knowledge obtained from Wikipedia. Meanwhile, instead of using the knowledge obtained from Wikipedia, other approach includese obtaining knowledge from the database built based on the relational database management system, Pascal, and also Java programming language [6]. Furthermore, an e-business chatbot has been developed via Artificial Intelligent Markup Language (AIML) and Latent Semantic Analysis (LSA) [7]. The authors claimed that with the combination of these two techniques, the chatbot is more intelligent in providing precise answers to the users. An attempt has been made to design a chatbot with a 3D Avatar, facial expression and also voice interaction based on APIs such as Chatbot API, speech recognition API, and voice processing API to enhance the interaction style with users [8]. An ergonomics evaluation to chatbot has been performed by researchers in [9]. Their findings claimed that chatbots with higher amount of factual knowledge are directly proportional to the level of user satisfaction. In their proposed system, a strategy centre is embedded to solve living questions. FAQ and NLP module have also supported the chatbot and provided up to more than half of the knowledge-sense questions.

Moreover, the chatbot has also been used in a world epidermic crisis, which is the Severe Acute Respiratory Syndrome (SARS) [20]. Furthermore, a chatbot has been developed to be used in the web interface regarding pandemic crisis communication. From their research, it is shown that the proposed chatbot is able to provide useful application in other domains [26]. Besides that, numerous efforts have been made to create chatbots for educational purposes [10], [13]. A chatbot which acts as an undergraduate advisor has been created [10]. The authors evaluated the chatbot with three experimental cases: a dialogue system with natural language knowledge, a domain knowledge system trained with information content, and finally a combined system which fused the dialogue system with thedomain knowledge system. The evaluation results showed that the dialogue system mixed with domain knowledge system yielded the highest conversation satisfaction result. An e-tutor chatbot to guide students in their study or performance has also been developed [13].

In addition to this, the idea of using the ontology concept in chatbot creation has been presented [11], [12]. The authors provided an approach to enhance chatbot through ontology and sentence reducer module [11]. Via the ontology, the chatbot is capable to automatically populate chatbot knowledge base whereas via the reducer module, sentences can be reduced into simpler phrases, allowing for an easier pattern matching to gain the correct answer, thus improving the performance of the chatbot. On the other hand, the approach in creating a chatbot via ontology concept to assist in the recommendation of test and measurement instruments has been deveopled [12]. Likewise, several researchers have been making an effort to conduct a survey on award-winning chatbots through Loebner Prize Competition [14], [18]. The results of the survey demonstrated that no significant advancement was made on the chatbot technologies, but changes were made only from simple pattern matching system to complicated patterns with knowledge base and ontology.

The works demonstrated by the researchers above show that only a few chatbots have gone through the evaluation process to indicate their level of intelligence in conversing with humans. Most of the researchers have been keeping in pace to create human-like chatbots by neglecting the necessity to evaluate the general knowledge of the chatbot before proceeding into the creation of domain-specific. It can be concuded that chatbot evaluations have not been put into much considerations and thus it motivates the present study.

III. Black-box Aproach

Black-box testing is originally used to perform software testing to ensure all the functionalities work in compliance to the software requirements and specifications without looking into the implementation details, knowledge creation, and the internal structures of a programme. In simpler words, this testing is only focusing on the inputs and outputs of the software without examining the internal structures, paths and knowledge of the software to indicate whether the functionalities meet the users' expectations.

Black-box testing is regarded as a qualitative assessment and is widely used by other researchers. For instance, the Turing test, which has a renowned reputation in the field of Artificial Intelligence is also using the black-box approach to conduct the evaluation [15]. Moreover, several researchers have applied the black-box approach for evaluation purposes in determining the response quality of conversational agents [16], [17]. In addition to this, one of the researchers was performing black-box testing on the conversational agent using Loebner Competition as the datasets [24]. As the main function of the conversational agent is to deliver appropriate responses to the users without looking into the internal processing, implementation details, knowledge creation, and the internal structures, the black-box approach is therefore deemed necessary to be used as the evaluation approach to evaluate the response quality of the general knowledge of the developed conversational agent, Tarie. In order to start up with the black-box approach, a set of appropriate questions are required to examine the response quality of the conversational agent. For the evaluation purpose, a total of 100 questions retrieved from the recent five years of Loebner Prize Competition are used as the datasets. The Loebner Prize Competition and datasets are further elaborated in Section IV and Section V.

IV. Loebner Prize Competition

The Loebner Prize Competition is a yearly event in the field of Artificial Intelligence in order to discover the most human-like chatbot as reviewed by the judges [19]. It is regarded as the oldest Turing Test competition. In 1990, Hugh Loebner and the Cambridge Centre for Behavioural Studies agreed to start this Turing Test contest. Dr Hugh Loebner would reward $100,000 and a Gold Medal for the first computer or chatbot which responses were not distinguishable from humans. From that year onwards, numerous institutions across the globe have been hosting the competition. The prize committee spent almost two years in planning the structure of the tournament [23]. The questions from the competition are generated after careful considerations in various aspects to test the intelligence of the chatbot's general knowledge [19], [25]. The tested aspects are:

- general knowledge
- reasoning knowledge
- memory knowledge
- personality knowledge

General knowledge is the knowledge of various subjects including people, places and also factual information. For example, "What colour is the sea?". Reasoning knowledge on the other hand is the knowledge of drawing the inferences or conclusions through the use of reason. For example, "If a bed doesn't fit in a room because it's too big, what is too big?". Furthermore, memory knowledge is the knowledge to recall previous information. For example, the user told the chatbot that his name was Andrew. Then, in the middle of the conversation, the user asked the chatbot again "What is my name?". In this case, the chatbot should be able to answer "Your name is Andrew". Personality knowledge is the knowledge related to the profile of the respective chatbot. For example, "Are you married?"

The competition format is based on the Turing Test, named after Alan Turing, who was a famous computer pioneer and British mathematician [21]. He suggested the test in a paper (year 1950) with the title of Computing Machinery and Intelligence. A human judge will test the chatbot using a computer programme, computer and keyboard, then key-in the appropriate questions to communicate with the chatbot. Based on the answers or responses given by chatbot, the judge will present the marks and rank the most human-like to least human-like conversation partners. The chatbot with the highest ranking is considered the winner of the competition and awarded with a cash prize and medal [21]. The compilation of the Loebner Prize Competition winners and its technique used since 1991 until the most recent year, 2017 can be accessed can be accessed online.[1]

[1] https://docs.google.com/document/d/1pfyjYVRs7GmlnBkb6r1SD4woiNUov1fOj4id5AW4HoQ/edit

V. Datasets

Datasets are one of the important aspects in this research. Datasets are needed to evaluate the response quality of the general knowledge generated by the chatbot. Hence, for the purpose of the evaluation, a total of 100 questions posed in the Loebner Prize Competition are selected as the datasets [19], [25]. As mentioned in Section IV, the competition has since started in 1991 and continued to be held until the current year, 2017. The entire datasets since the beginning year would be too long for the evaluation. Hence, for brevity, only the datasets from the recent five years, ranging from 2013 to 2017 would be chosen for the evaluation purpose. For more information, the compilation of the recent five years' datasets from Leobner Prize Competition can be accessed online.[2]

VI. Conversational Agent – Tarie

The proposed chatbot, Tarie is an artificial conversational agent developed in this research to be used as the subject in the black-box evaluation. Tarie is capable of responding to general knowledge related questions probed by users. Tarie's general knowledge is obtained from the AAA (Annotated A.L.I.C.E. AIML) that comes from the brain of the award-winning artificial conversation chatbot, ALICE [22]. The general knowledge is then improved by adding new general knowledge from time to time. Before Tarie could be used to serve its purpose to deliver appropriate general information to the users, it is indisputable to say that the evaluation of Tarie's general knowledge plays a crucial role. The evaluation results are able to indicate the quality of responses from Tarie and to determine whether the chatbot is ready for the domain-specific creation or to further improve its general knowledge.

VII. Results and Findings

This section shows the results of the Tarie black-box approach evaluation. It is further divided into a few sub-sections to discuss the scoring method, the responses of Tarie and the results of the analysis.

A. Scoring Method

The scoring method is conducted in accordance to the recent Loebner Prize Competition. The compilation of the questions posed for the evaluation purpose is illustrated in Section V. As discussed in Section IV, these questions are generated after careful considerations in various aspects to test the intelligence of the chatbot. The scoring method is similar to the latest Loebner Prize Competition in 2017, in which the response to each question is weighed based on three criteria specifically relevance, correctness, and clarity of expression or grammar as per the guideline given in [19], [25]. By using the guideline as reference, an expert is selected to allocate the score. The score 0, 1 or 2 is then allocated to the response of the chatbot in conjunction to the degree where it meets the criteria. The score 0, 1 or 2 is depicted as below:

0 = The response does not meet the criteria
1 = The response partially meets the criteria
2 = The response fully meets the criteria

An example of the scoring guidelines for the question "What is the answer of 1 + 1?" is depicted as in Table 1.

```
ID: 173
Date Time: 2017-10-14 08:36:06
Judge: Hello, my name is Andrew. What's your name?
Tarie: Hello, Andrew. My name is Tarie, nice to meet you.

ID: 174
Date Time: 2017-10-14 08:39:07
Judge: Why don't you tell me a little more about yourself?
Tarie: I am friendly and fairly good looking.

ID: 175
Date Time: 2017-10-14 08:41:07
Judge: I am a researcher in Artificial Intelligence at Goldsmiths University,
do you know what that is?
Tarie: I am not sure about that.

ID: 176
Date Time: 2017-10-14 08:44:30
Judge: Will you tell me about your dreams?
Tarie: I dream of becoming a scientist.

ID: 177
Date Time: 2017-10-14 08:46:03
Judge: Can you rephrase that?
Tarie: Rephrase what?
```

Figure 1. Sample of Datasets on Judges' Questions in 2017 with Tarie

[2] https://docs.google.com/document/d/16-g7zej18_rGVQc2kZqu9xiT6Yqg5iXdZs0Nh1P6HM4/edit?usp=sharing

Score	Response	Explanation
0	You mean hobby?	0 point as the bot did not know how to answer.
1	I am a bot, not a math teacher.	1 point as the bot did not provide a correct answer but was able to answer within the scope.
2	The answer is 2.	2 points for answering correctly.

Table 1. Scoring Guidelines

B. *Evaluation of Results and Analysis*

As the entire responses given by Tarie alongside with the scores given would be too long, the responses can be assessed online[3]. The sample of datasets on the judges' questions in year 2017 with Tarie is as depicted in Figure 1. After that, the results obtained via the black-box testing are further deduced. The scores are further grouped to observe the level of occurrence as shown in Figure 2. Based on Figure 2, it could be perceived that majority of the responses fell under score-2. Score-2 indicates that the responses have fully met the criteria. This demonstrates that Tarie was able to deliver accurate and reliable answers for most of the questions.

As discussed in Section IV, the responses were tested via four aspects. The analysis shows that Tarie was capable to provide relatively accurate responses in terms of general knowledge-type questions and personality-type questions. For instance, when Tarie was asked a personality-type question "Why don't you tell me a little more about yourself?", Tarie was able to deliver a logical response by saying "I am friendly and fairly good-looking". As for general knowledge-type questions, when Tarie was asked "What do you think of Trump?", Tarie was able to answer correctly by saying "Trump was a successful businessman before becoming the president of America". Apart from this, the second highest occurrence fell under score-0 which indicates that the response did not meet the criteria.

This shows that Tarie was unable to response to certain questions asked. The answers provided by Tarie were rather inaccurate and irrelevant, especially when it came to reasoning-type questions. For instance, when Tarie was asked "The trophy doesn't fit into the brown suitcase because it's too small. What is too small?" and Tarie responded the reasoning-type question with "Atomic particles". Since the response given by Tarie was inaccurate, Tarie was given the score-0. The least occurrence fell under score-1 indicating that the responses partially met the criteria. For instance, "How many presidents of the US were called Bush?". Tarie responded with "Bush is one of the presidents of US.". The response given by Tarie indicated that Tarie was not able to provide the correct answer but able to answer within the scope. Hence, score-1 was given in this question. Even though Tarie was unable to deliver the correct answers, it somehow demonstrated the capability to provide relevant answers to prolong the conversations. In addition to this, Tarie was seen as uncertain in providing accurate responses in memory-type questions.

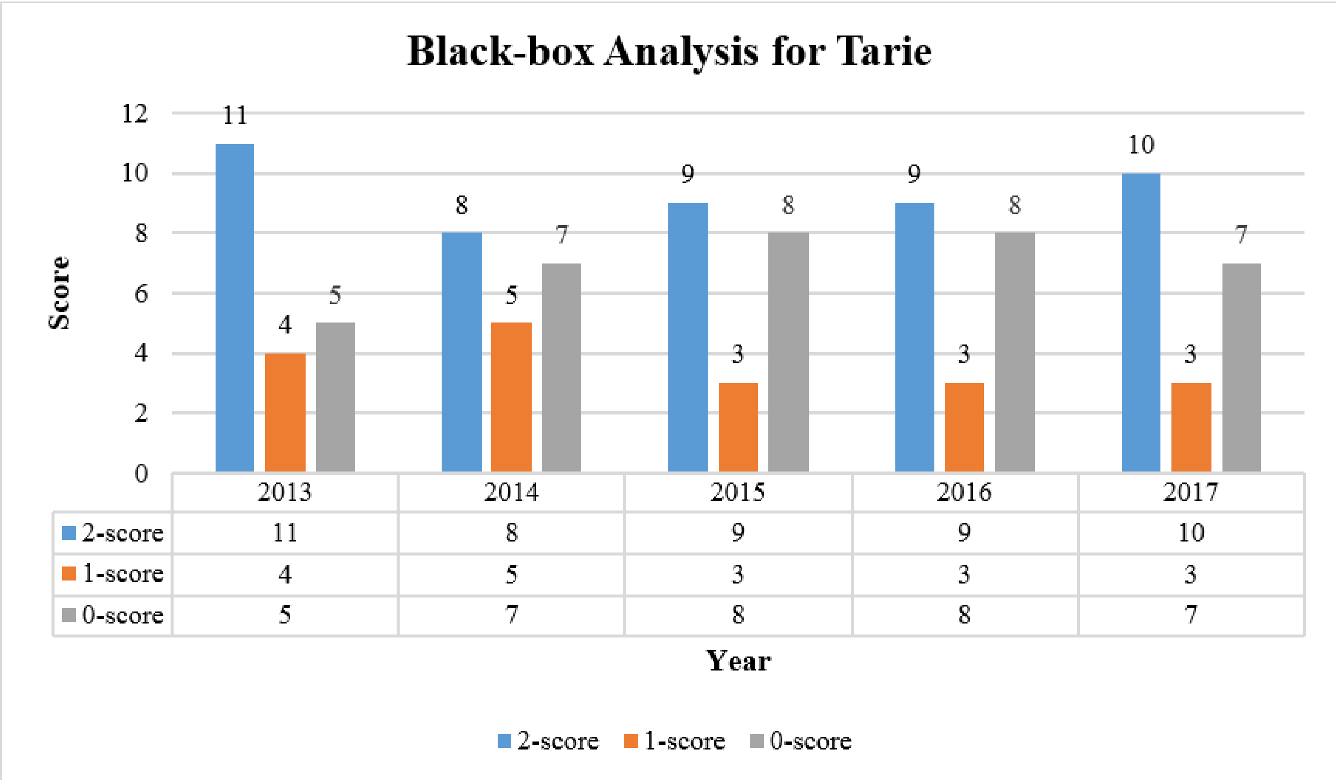

Figure 2. Black-box Analysis for Tarie

[3] https://docs.google.com/document/d/1bcGmJu1z-1nPU4ET2A3t83agpTC94B3ggv5kbC03rkg/edit?usp=sharing

The occurrences of score-0 and score-1 somehow verified that Tarie is currently still lacking in providing precise responses to memory-type questions. For instance, in the dataset year 2017, the conversation in question 3 indirectly informed Tarie that "I am a researcher in Artificial Intelligence at Goldsmiths University, do you know what that is?" Tarie responded that "I am not sure about that". Score-0 was given for this question. After that, Tarie was asked again in question 6 "What is my occupation?". Tarie responded with "I believe I have yet to hear from you.". Since Tarie was unable to recall the past conversation, the score given for this memory-type question was score-0. As this is a memory-type question, Tarie should be able to relate with the previous conversation and responded with "Your occupation is a researcher in Artificial Intelligence at Goldsmiths University.". Hence, there is still a room for improvement for the development of Tarie's knowledge base, especially with reasoning-type questions and memory-type questions.

In addition, Tarie is also compared with different chatbots, ranging from year 2013 to year 2017 as depicted in Figure 3 below. The scores from the chatbots in Loebner Prize Competition for the past five years are then compared with the score from Tarie. The scores are captured in terms of three aspects which are the yearly winner which has the highest score, the average score of the competing chatbots and the score from the lowest chatbot. Furthermore, the percentages of the highest score, average score and lowest score are gathered from [19], [24], [25]. The names of the chatbots are left to be anonymous to respect the works of other researchers. Figure 3 illustrates the comparison of scores between Tarie and Loebner Prize Competition Chatbots (in %).

Based on Figure 3, in year 2013 and year 2017, Tarie is able to score above average compared to other chatbots. For instance, in year 2017, the highest score captured is 67.5%, the average score captured is 39.26% and the lowest score captured is 5% whereas the score captured for Tarie is 57.5%, which is above average. Likewise, in year 2014 to 2016, Tarie's score is within the average compared to other chatbots. The average score is partly due to Tarie being unable to answer some of the questions posed.

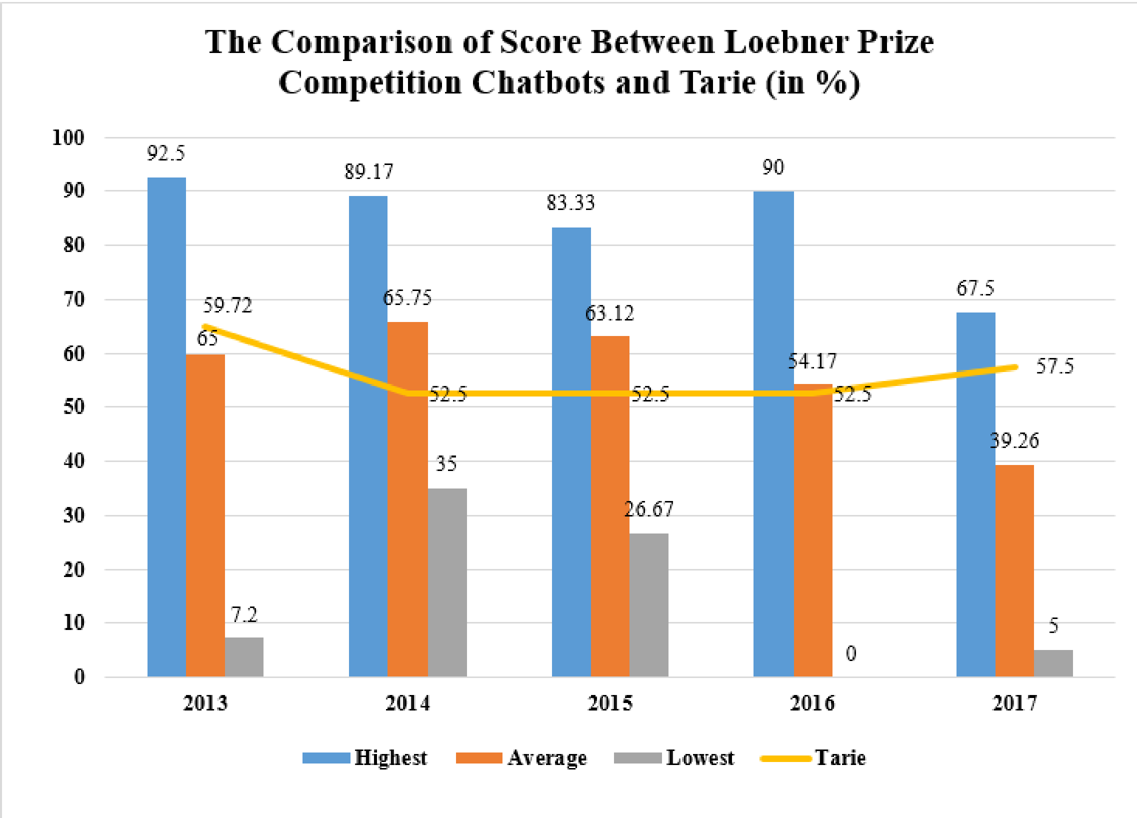

Figure 3. The Comparison of Score Between Tarie and Loebner Prize Competition Chatbots

Nevertheless, the analysis shows that Tarie has obtained average results in comparison to other competing chatbots in a five year period. This analysis shows that Tarie is capable to deliver fairly accurate responses.

VIII. Conclusions and Future Works

It is concluded that there is still limited research that emphasise on the evaluation of the chatbot's general knowledge before proceeding to the creation of the domain-specific knowledge. The goal of this paper is to conduct an evaluation on the chatbot named Tarie to indicate its quality response in conversing general knowledge, reasoning knowledge, memory knowledge as well as personality knowledge via a qualitative assessment known as black-box testing using datasets obtained from Loebner Prize Competition. The evaluation results indicate that the majority of responses fell under score-2, especially to general knowledge-type questions and personality-type questions. Nevertheless, Tarie was unable to answer certain questions when it comes to reasoning questions and memory-type questions. The black-box approach deduced that the current Tarie is capable in answering questions in general knowledge and personality knowledge but still lacking in terms of reasoning knowledge and memory knowledge. As the brain of the award-winning artificial conversation chatbot ALICE will be updated from time to time, so to will the general knowledge of Tarie be improved before proceeding to the creation of domain-specific knowledge. Our future work would be focusing on the domain-specific knowledge using machine learning via retrieval-based models. Likewise, for better interactivity, the chat interface would be integrated with predefined questions recommendation. The predefined responses would be generated dynamically to the chat interface based on the popularity of the domain-specific knowledge. Another future work would be using the crawler technology to automatically extract frequently asked questions (FAQ) from existing websites for the knowledge base creation.

Acknowledgement

Special thanks to the late Hugh Loebner for launching the Loebner Prize Competition to motivate more researchers to engage in the field of artificial conversational agent, the grant from UTeM and Big Data Technology Pty Ltd, Australia (GLuar/BIGTECH/2017/FTMK-CACT/A00009) and the scholarship from MyBrain15 awarded by the Ministry of Higher Education, Malaysia.

References

[1] G. M. D’silva, S. Thakare, S. More, and J. Kuriakose. “Real World Smart Chatbot for Customer Care using a Software as a Service (SaaS) Architecture”. In *International conference on I-SMAC*, pp. 658–664, 2017.

[2] D. Madhu, N. Jain, E. Sebastain, S. Shaji, and A. Ajayakumar. “A Novel Approach for Medical Assistance Using Trained Chatbot”. In *International Conference on Inventive Communication and Computational Technologies*, pp. 243–246, 2017.

[3] K. Oh, D. Lee, B. Ko, and H. Choi. “A Chatbot for Psychiatric Counseling in Mental Healthcare Service Based on Emotional Dialogue Analysis and Sentence Generation”. In *International Conference on Mobile Data Management*, pp. 371–376, 2017.

[4] N. Kumar, L. Chandar, V. Prasad, and K. Sumangali. “Android Based Educational Chatbot for Visually Impaired People”. In *International Conference on Computational Intelligence and Computing Research*, 2016.

[5] D. Lee, K. Oh, and H. Choi. “The ChatBot Feels You – A Counseling Service Using Emotional Response Generation”. In *International Conference on Big Data and Smart Computing*, pp. 437–440, 2017.

[6] B. Setiaji and F. W. Wibowo. “Chatbot Using A Knowledge in Database Human-to-Machine Conversation Modeling”. In *International Conference on Intelligent Systems, Modelling and Simulation*, 2016.

[7] N. T. Thomas. “An E-business Chatbot using AIML and LSA”. In *International Conference on Advances in Computing, Communications and Informatics*, pp. 2740–2742, 2016.

[8] A. Angga, E. Fachri, A. Elevanita, and D. Agushinta. “Design of Chatbot with 3D Avatar , Voice Interface , and Facial Expression”. In *International Conference on Science in Information Technology*, pp. 326–330, 2015.

[9] W. Liu, J. Zhang, and S. Feng. “An Ergonomics Evaluation to Chatbot Equipped with Knowledge-Rich Mind”. In *International Symposium on Computational and Business Intelligence*, 2015.

[10] S. Ghose and J. J. Barua. “Toward the implementation of a Topic specific Dialogue based Natural Language Chatbot as an Undergraduate Advisor”. In *International Conference on Informatics, Electronics and Vision*, 2013.

[11] A. Augello, G. Pilato, A. Machi, and S. Gaglio. “An Approach to Enhance Chatbot Semantic Power and Maintainability : Experiences within the FRASI Project”. In *International Conference on Semantic Computing*, 2012.

[12] R. Kincaid and G. Pollock. “Nicky : Toward A Virtual Assistant for Test and Measurement Instrument Recommendations”. In *International Conference on Semantic Computing*, pp. 196–203, 2017.

[13] F. A. Mikic, J. C. Burguillo, M. Llamas, D. A. Rodr ǵuez, and E. Rodr ǵuez. “CHARLIE : An AIML-based Chatterbot which Works as an Interface among INES and Humans”. In *EAEEIE Annual Conference*, 2009.

[14] L. Bradeško and D. Mladenić. “A Survey of Chabot Systems through a Loebner Prize Competition”, vol. 2, pp. 2–5, 1975.

[15] L. Floridi. “Informational Realism”. In *Computing and Philosophy Conference*, vol. 37, pp. 7–12, 2004.

[16] S. L. Lim and O. S. Goh. “Intelligent Conversational Bot for Massive Online Open Courses (MOOCs)”, 2016.

[17] O. S. Goh, C. Ardil, W. Wong, and C. C. Fung. “A Black-box Approach for Response Quality Evaluation of Conversational Agent Systems”, *International Journal of Computational Intelligence*., pp. 195–203, 2007.

[18] S. A, Abdul-Kader, and D. J. Woods. “Survey on Chatbot Design Techniques in Speech Conversation Systems”, *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 7, pp. 72–80, 2015.

[19] Home Page of The Loebner Prize in Artificial Intelligence. [online] Available at: http://www.loebner.net/Prizef/loebner-prize.html [Accessed on 5 November 2017].

[20] O. S. Goh, C. C. Fung, A. Depickere, K. W. Wong, and W. Wong. “Domain Knowledge Model for Embodied Conversation Agent”. In *International Conference on Computational Intelligence, Robotics and Autonomous Systems*, pp. 1–2, 2005.

[21] Home Page of Alan Turing: The Enigma. [online] Available at: http://www.turing.org.uk/ [Accessed on 15 June 2018].

[22] Home Page of Free A.L.I.C.E. AIML Set. [online] Available at: https://code.google.com/archive/p/aiml-en-us-foundation-alice/ [Accessed on 15 June 2018].

[23] Home Page of Lessons from a Restricted Turing Test. [online] Available at: https://www.eecs.harvard.edu/shieber/Biblio/Papers/loebner-rev-html/loebner-rev-html.html/ [Accessed on 15 June 2018].

[24] A. Pradana, O. S. Goh, and Y. J. Kumar. “SamBot - Intelligent Conversational Bot for Interactive Marketing with Consumer-centric Approach”, *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 6, no. 2014, pp. 265–275, 2017.

[25] Home Page of The Society of The Study of Artificial Intelligence and Simulation of Behaviour. [online] Available at: http://www.aisb.org.uk/events/loebner-prize [Accessed on 15 June 2018].

[26] O. S. Goh, C. C. Fung, K. W. Wong, and A. Depickere. "An Embodied Conversational Agent for Intelligent Web Interaction on Pandemic Crisis Communication". In *International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 3–6, 2006.

Author Biographies

Leong Pui Huang is a Programme Leader in the Department of Computing and Information Technology at Tunku Abdul Rahman University College (TAR UC). She earned both her bachelor degree and master from Universiti Teknikal Malaysia Melaka (UTeM), in the field of Computer Science in year 2013 and 2015. She is currently pursuing her PhD and her research involves intelligent agent and AI applications.

Prof. Ts. Dr. Goh Ong Sing started his career as an academician since 1990 at Universiti Sains Malaysia (USM). In August 2002, he joined Universiti Teknikal Malaysia Melaka (UTeM) and in 2015 he was appointed as a Director of Industry Liaison Centre and Lead Samsung IoT Academy. Now he is working as Assistant Vice Chancellor at Office of Industry and Community Network. 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 100 peer-eviewed scientific journal, books, conference papers, and book chapters. He has led and worked on research grants funded by Malaysian Government's Intensified Research, Malaysia Technology Development Corporation, Big Data Technology Pty Ltd, Murdoch University, and Australian Research Centre for Medical Engineering at University of Western Australia

Dr. Yogan Jaya Kumar is a Senior Lecturer at the Department of Intelligent Computing and Analytic in the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM). He earned both his bachelor degree and master degree from Universiti Sains Malaysia (USM), in the field of Mathematical Science in year 2003 and 2005. He completed his PhD studies at Universiti Teknologi Malaysia, in 2014 in the field of Computer Science. Currently, his research involves in the field of Text Mining,Information Extraction and AI applications.

Aditya Pradana received his Master degree in Software Engineering and Intelligence from Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia in 2017. His research interests are in intelligent agent and blockchain technology. Currently he is pursuing his PhD while working at Samsung IoT Academy UTeM as a web application developer.