BoTest: a Framework to Test the Quality of Conversational Agents Using Divergent Input Examples

Elayne Ruane[1,2], **Théo Faure**[1,2], **Ross Smith**[3], **Dan Bean**[3],
Julie Carson-Berndsen[1], **Anthony Ventresque**[1,2]
[1]School of Computer Science, University College Dublin, Ireland.
[2]Lero - the Irish Software Research Centre
[3]Microsoft Corporation, Skype Division, Seattle, USA.
{elayne.ruane,theo.faure}@ucdconnect.ie, {julie.berndsen,anthony.ventresque}@ucd.ie

ABSTRACT

Quality of conversational agents is important as users have high expectations. Consequently, poor interactions may lead to the user abandoning the system. In this paper, we propose a framework to test the quality of conversational agents. Our solution transforms working input that the conversational agent accurately recognises to generate divergent input examples that introduce complexity and stress the agent. As the divergent inputs are based on known utterances for which we have the 'normal' outputs, we can assess how robust the conversational agent is to variations in the input. To demonstrate our framework we built ChitChatBot, a simple conversational agent capable of making casual conversation.

Author Keywords

Conversational Agent Testing; Conversational Agent Quality Assessment; Chatbot.

ACM Classification Keywords

Human-centered computing: Human computer interaction (HCI): HCI design and evaluation methods

INTRODUCTION

Conversational Agents (CAs) have recently gained a lot of popularity both in industry and academia. Many large companies have developed their own CAs in the past 3-5 years such as Apple Siri, Google Now, Microsoft Cortana and Amazon Alexa. Such interest and success is reflected in a growing market size which is expected to reach more than $12B in 2024 with an annual growth rate of 38% [1].This is supported by a rich body of work in academia addressing questions such as embodiment of conversational agents [2] and user expectations [3]. Chatbots, a type of conversational agent, have become increasingly popular [4, 5] with many widely used platforms such as Skype and Facebook Messenger supporting bot integration.

Previous work around chatbot quality has focused on user quality of experience and has shown that users have high expectations and low tolerance for poor quality [3]. However, little work has been done to test the linguistic quality of the chatbot in a way that does not require user feedback. This paper[1] describes our first attempt to bridge this gap. Our objective is to understand how to ensure CAs rolled out to users are robust enough to meet the expectations and needs of a heterogeneous multitude of users.

We propose a framework (see Figure 1) to test the quality of CAs using *divergent input examples*. The framework takes a textual user utterance as input. Divergent examples of this input are generated that introduce language complexity for the CA to deal with but which maintain the intent of the original input. As such, the CA should produce the same or a similar output. Divergent examples can represent various challenges from grammatical errors to register changes. The CAs performance on these divergent examples can be compared to its response to the original input utterance and analysed to highlight areas in which the quality of the agent is reduced. This can paint a picture of the CAs robustness and meaningful feedback can be provided to the developers.

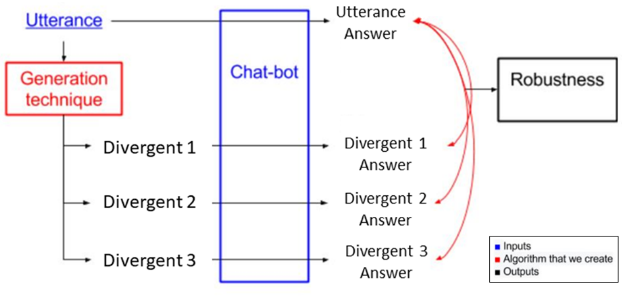


Figure 1. Overview of our testing framework (BoTest).

OUR CONVERSATIONAL AGENT: CHITCHATBOT

We have developed our own CA: ChitChatBot, which makes 'chit chat' with the user by discussing topics such as the weather, holidays, tv, movies, music, and celebrity gossip.

[1] This work was supported with the financial support of the Science Foundation Ireland grant 13/RC/2094 to Lero.

We designed ChitChatBot to discuss these topics to introduce a level of linguistic complexity not found with agents designed to complete simple tasks with a limited scope. Instead, ChitChatBot must deal with multiple and varied intents. It also deals with 'loose' linguistic models, in the sense that the expected utterances reflect casual English. We implemented ChitChatBot using the Language Understanding Intelligent Service[2] from Microsoft Azure Cognitive Services. ChitChatBot was trained using over 400 manually labeled utterances to learn to recognise 12 different user intents representing the topics mentioned above. Our bot achieved a 95% accuracy over all intents – i.e. when presented with an utterance it could give the right intent in 95% of cases.

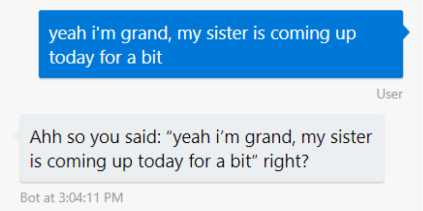
TESTING FRAMEWORK

Our framework[3] is modular and can easily integrate divergent techniques such as syntactic divergents (e.g. word order errors), morphological divergents (e.g., incorrect verb tense), semantic divergents (e.g., use of synonyms), etc. For illustrative purposes, we demonstrate our framework using two divergent techniques: non-native preposition error and native colloquial phrasing. The non-native preposition error occurs when a second language learner uses an incorrect preposition such as "arrive to the airport" instead of "arrive at the airport". These errors are very common and something we would expect the CA to be able to deal with. The second divergent, native colloquial phrasing, occurs when a native speaker of a language uses informal or local expressions. In this case, we used Irish colloquial language such as "Are ya well?" instead of "How are you?". These divergents may be harder for the CA to deal with as they significantly alter the structure of the utterance. We selected these divergents to represent both simple and complex divergent types.

We evaluated ChitChatBot on a selected set of 8 out of the 12 trained intents. We did not test the simple intents such as "Greet.Hello" and "Greet.Goodbye". For each intent we selected 6 utterances for which the ChitChatBot correctly identified the intent. We produced an equivalent utterance by altering nouns and adjectives but maintained the sentence structure. From these equivalent utterances we generated either a preposition error divergent or a colloquial phrasing divergent. As such we had 3 divergents of each type for each intent, leading to 48 divergents in total.This was done manually for the purposes of this experiment.

The bot failed to identify the user intent for 2 out of 48 divergents. Both were native colloquial phrasing divergents such as in Figure 2 where the conversation is not successful.

The bot correctly identified the user intent for 20 our of 48 divergents but failed to understand the true meaning of the utterance and produces a generic response. For instance, when given the divergent input "there was a huge time difference at India but it was beautiful", ChitChatBot found the correct intent ("Vacation") but failed to recognise India as a country due to the wrong preposition.The bot correctly identified the user intent and specific meaning of the utterance for the remaining 26 out of 48 divergents. These 26 correctly handled divergents were comprised of 13 of each divergent indicating that ChitChatBot had similar performance on both divergents.



yeah i'm grand, my sister is coming up today for a bit

Ahh so you said: "yeah i'm grand, my sister is coming up today for a bit" right?

Figure 2. The bot failed to identify the intent (native colloquial phrasing divergent).

However, when we group results by intent we do not see such balanced performance. Our bot was better able to handle preposition error divergents over colloquial phrasing divergents for some intents but the inverse was true for other intents. Interestingly, the average length of training utterances was 40 characters for the intents on which the bot was better able to handle colloquial phrasing divergents but only 25 characters where performance was better on preposition error divergents. This makes sense because colloquial phrasing is verbose in nature and so does not have as much of an effect on intents that are naturally invoked using longer utterances. There was no significant correlation between the bots performance on the divergents for a particular intent and the number of utterances used to train the bot for that intent.

CONCLUSION

The potential applications and benefits of conversational agents are clear. Users have demonstrated a willingness to use these agents for an array of tasks of varying complexity. When these CAs are rolled out to users they should be fit for purpose and provide the level of quality of experience that will encourage the user to engage in repeated and frequent use. We propose BoTest, a testing framework that finds types of user input the agent is unable to handle, allowing developers to address quality issues before deployment. Our validation showed how this approach works and the kind of information the framework can provide.

REFERENCES

1. Grand View Research Report: IVA Market Size Projected To Reach $12.28 Billion By 2024. Accessed: 2017-12-21.
2. Cassell, J. *Embodied conversational agents*. MIT press, 2000.
3. Luger, E., and Sellen, A. Like having a really bad pa: the gulf between user expectation and experience of conversational agents. In *CHI* (2016), 5286–5297.
4. Nguyen, M.-H. Business Insider: The latest market research, trends & landscape in the growing AI chatbot industry. Accessed: 2017-12-21.
5. Weizenbaum, J. Eliza a computer program for the study of natural language communication between man and machine. *Communications of the ACM 9*, 1 (1966), 36–45.

[2] `https://www.luis.ai/`

[3] `https://www.github.com/elayneruane1/BoTest/`