A Survey of Efficient Ways of Interaction between Humans and Conversational Agent for General-Purpose Conversation

Bauyrzhan Zhakanov[1]

[1]Department of Robotics Engineering,
University of Genoa, Italy,
Research Track II: Report
Email: bauyrzhan.zhakanov@gmail.com

***Abstract*—The objective of the article is the investigation of state-of-the-art papers that evaluate the conversational agents using the various methods for general dialogue. First study has given a more productive approach to interact with different robotic platforms using Cloud services rather than using the same system in Local connection. Second study indicated the performance of baseline neural networks and GPTs, whereas the last study has shown the productive and entertaining interaction between agents and people using various methods and experiments. In general, all studies have provided efficient ways of user interaction that supplement each other for implementing future conversational agents.**

I. Introduction

As the worldwide demand for chatbots increases, especially in tech companies, the evaluation of user interaction must be in the first place [1]. According to Hussain et. al, there are two main categories of the chatbots: task oriented and non-task oriented [2]. The task oriented chatbots can be used for certain purposes by users. For instance, there could be different scenarios such as the address of the place, the booking information, the order of a particular product, etc. According to the researchers, there are two classes for non-task oriented chatbots: Generative-Based model, which is also called a sequence-to-sequence model. This model is able to generate a proper answer during dialogue with a human. Retrieval-Based model, which is capable of learning from the current dialogue from his database repository. The study also represents the employment of various design techniques that can be used in any chatbot approaches: Parsing: the method that indicates only the specific keywords from text or sentences. This technique could be used as the storage of particular words that can be easily manipulated for matching texts and determine the grammatical structure of it. Pattern-Matching: the method is aimed to classify the input as a pattern and generate the pattern as an output. The advantage of this method is being flexible to create sentences, meanwhile the disadvantage of this method is the scaling and the limitation of extraction capability. AIML: the technique that is designed for creating conversational flow in dialogue. This technique is easy to use and flexible in designing chatbots. Chatscript: this is a rule-based agent that has a combination of natural language and a dialogue management system [6]. An existing database of some 2000 predefined concepts and scripters can also easily write their own concepts [7]. Ontologies: this technique is to substitute the hand crafted domain knowledge with ontological domain knowledge. The advantage of method is being able to find the concept nodes of an ontology to build a relationship between concepts. Markov Chain Model: the technique is widely used because it is capable of creating simple human speech, easy to use and the model can be summarized by matrix. Artificial Neural Networks Models: one of the most common approaches for modeling are Seq2Seq (Sequence to Sequence) models and LSTM (Long Short Term Memory) networks that are based on the RNN (Recurrent Neural Networks) architecture. Nowadays, Seq2Seq is considered to be widely used by industries, despite the fact that LSTM is considered to be more efficient in terms of classification than other available techniques, but it refers to distant data. In Chapter II, there are three main state-of-the-art papers that are investigated in this research theme: A Feasibility Study of Culture Aware Cloud Services for Conversational Robots by Recchuito et al [3], and Towards the Use of Pre-Trained Language Models for Task-Oriented Dialogue Systems by Budzianowski et al [4], and A rapport-building 3D embodied conversational agent for the Human Support Robot by Pasternak et al [5]. The objective of the report is to find an efficient way to implement the conversational agent using different approaches. In Chapter III, each of these articles would be assessed by its efficiency and the complement in order to complement the future conversational agent to increase its interaction and entertainment with people.

II. Related work

The first research article is provided by Recchiuto et al [3], where the main purpose of investigation is the summarization of the underlying software architecture of the conversational framework and its functionalities using Cloud services which is called CARESSES. The conversational architecture of this project consists of two main elements: Three layered ontology for storing the concepts, cultural relevance, specific informa-

tion and preferences. An algorithm of building a dialogue that is based on these layers. This algorithm is a keyword based language process that is applied to be triggered if one of the topics in the tree. Those three layers where the Ontology is clearly divided are: TBox (Terminology box) that demonstrates all data that have a significant role in cultural-aware conversation: beliefs, habits, preferences, objects ABox I (Assertion box I) that stores the assertions to store culture-specific data. ABox II (Assertion box II) that stores the assertions to store culture-specific data for unique cultural identity, physical environment, etc. The work analyzes have been evaluated in two scenarios: Local (LAN) and Cloud Connections. From the analysis of the paper, the average latency in Cloud Services is a bit higher, but the standard deviation is smaller than Local connection. According to the results, CL (Communication Latency) of Cloud is higher than LAN rather than PL (Processing Latency) where both are similar. In addition, the Cloud service studied that the PL between Directive-Question and Direction-Open speeches is similar, while the CL has more delay. Nevertheless, researchers had found that there is a linear connection between the size of Ontology such as the generation of dialogue trees (750, 1500, and 2250 topics), and PL (Processing Latency). Moreover, the CARESSES project is compared with Google Cloud Services, where the average latency of the author's project is less than Google's one. However, the CARESSES project is considered the only culturally competent reply of a user, while Google's technology is the conversion of audio-to-text.

The second state-of-the-art article is about the usage of pre-trained models for task oriented dialogue provided by Budzianowski et al [4]. The study represents the progress in language modeling, pre-training, and transfer learning that operates on given text input. The dataset used in this study is provided by MultiWOZ dataset. The frameworks that authors used to build a model was provided by Wolf et al [8], and the generative model from Radford et al [9]. The main evaluation is focused on the comparison between two models: The Baseline of Neural Response Generation model with an oracle belief by Budzianowski et al [10]. The framework for modeling task-oriented conversations that works only with text format. All models pretrained on the GPT model and two GPT-2 models (small GPT-2 and medium GPT-2). There were three measures for evaluation: Inform as an appropriate entity, Success as an answered response, and the fluency by BLEU score [11]. Although, there were two ways of evaluation using the MultiWOZ dataset: greedy sampling and nucleus sampling. Based on the greedy sampling results, GPT-2 medium has shown better value in BLEU score, while other measures were higher in baseline. Meanwhile, GPT-2 medium has improved its scores with the nucleus sampling in every measurement. Moreover, human evaluation showed that there are no differences between the baseline and GPT-based models.

The last scientific article written by Pasternak et al [5] shows the mirroring behaviors of agents during conversation with humans. The objective of research is to study the user's interaction with 3D ECA (embodied conversational agents) based on facial expressions and head movements. The authors' approach towards implementation is the integration of Toyota Human Support Robot (HSR) with ECA via ROS (Robot Operating Systems) topics. To detect and analyze facial expressions, researchers used an adapted version of DLib [12]. Nevertheless, authors classified emotions using EmoPy [13] to ROS topic, and rendered the facial expression on an avatar [14]. The study was completed after three experiments: The assessment of posture-mimicking to test the user experience on a robot. The assessment of mirroring facial expressions with and without emotions. The assessment of the combination of facial mirroring and posture mimicking Based on the result of all three experiments, the combined posture mimicking and facial expression provides better user interaction than one or neither enabled.

III. Evaluation and Results

All scientific papers have shown different experimental results depending on the user's purpose. According to the first article, it is evaluated that there is not much difference between LAN and Cloud services in terms of speed and time latency. However, the implementation of Cloud services could provide usability due to its distant access to the system. In other words, if the system crashes in the local network, it would be able to create issues to the conversational agent since the model would be saved in Cloud. Based on the evaluation of the second scientific paper, it would be concluded that GPT-2 provides better performance in pre-trained models that produce human-like text. Moreover, OpenAI said that the newest version of GPT-2 is its third generation GPT-3 which consists of 175 billion parameters, whereas GPT-2 has only 1.5 billion parameters [15]. In general, the study indicated the approach of the task oriented dialogue generation using the MultiWOZ dataset.

The last study has shown how the 3D embodied avatars on robots could result in better performance with participants. Users are more engaged with the robot which considers facial expression as well as posture movements like nodding rather than having either facial expression or posture mimicking. Authors used a default deep learning toolkit with emotion classification, however it would be better to test the model using different approaches and frameworks or the pre-training the model by adding fully connected layers in order to observe how the newest deep learning frameworks could increase the performance of emotion classification.

IV. Conclusion

In conclusion, three main scientific articles have been investigated and evaluated depending on the authors experimental results. The first paper has provided a more efficient way of interaction with various robotic platforms using Cloud services than LAN. Second paper demonstrated that the newest pre-trained models could provide a proper result than old ones, meanwhile the third paper announced that the communication between agents and people using posture and facial movement

can increase the interaction and the performance of them. In my opinion, the combination of all three components do not oppose, but complement each other.

REFERENCES

1. Ren, R., Castro, J. W., Acuña, S. T., de Lara, J. (2019). Usability of chatbots: A systematic mapping study. In Proc. 31st Int. Conf. Software Engineering and Knowledge Engineering (pp. 479-484)

2. Hussain, S., Ameri Sianaki, O., Ababneh, N. (2019, March). A survey on conversational agents/chatbots classification and design techniques. In Workshops of the International Conference on Advanced Information Networking and Applications (pp. 946-956). Springer, Cham

3. Recchiuto, C. T., Sgorbissa, A. (2020). A feasibility study of culture-aware cloud services for conversational robots. IEEE Robotics and Automation Letters, 5(4), 6559-6566. 4. Budzianowski, P., Vulić, I. (2019). Hello, it's GPT-2–how can I help you? towards the use of pre-trained language models for task-oriented dialogue systems. arXiv preprint arXiv:1907.05774.

5. Pasternak, K., Wu, Z., Visser, U., Lisetti, C. (2021). Let's be friends! A rapport-building 3D embodied conversational agent for the Human Support Robot. arXiv preprint arXiv:2103.04498.

6. Wikipedia contributors. ChatScript. Wikipedia, The Free Encyclopedia. 4 September 2018, 19:19UTC. https://en.wikipedia.org/w/index.php?title=ChatScriptoldid=858055386. Accessed 22 Jan 2019

7. Robino, G.: How to build your first chatbot using Chatscript (2018). (cited 18 December 2018). https://medium.freecodecamp.org/chatscript-for-beginners-chatbots-developers-c58b b591da8

8. Wolf, T., Sanh, V., Chaumond, J., Delangue, C. (2019). Transfertransfo: A transfer learning approach for neural network based conversational agents. arXiv preprint arXiv:1901.08149.

9. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9

10. Budzianowski, P., Casanueva, I., Tseng, B. H., Gasic, M. (2018). Towards end-to-end multi-domain dialogue modeling.

11. Papineni, K., Roukos, S., Ward, T., Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

12. Davis E King. 2009. Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research 10 (2009), 1755–1758.

13. Perez Angelica. [n.d.]. EmoPy: A Machine Learning Toolkit for Emotional Expression, 2018.

14. Mihai Polceanu and Christine Lisetti. 2019. Time to Go ONLINE! A Modular Framework for Building Internet-Based Socially Interactive Agents. In Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (Paris, France) (IVA '19). Association for Computing Machinery, New York, NY, USA, 227–229. https://doi.org/10.1145/3308532.3329452

15. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.