Instructions for ACL-2016 Proceedings

Tom Byars, Cale Clark, Roman Fenlon, Charlie Lyttle, Katie McAskill, Jack Miller, Zein Said, Abdullah Sayed, Laura Schauer, Jason Sweeney, Aron Szeles, Xander Wickham
School of Mathematics and Computer Science, Heriot-Watt University, Edinburgh
tjb10, cc164, rf104, cl157, klm12, jjm7, zs2008,
as512, lms9, js418, as472, aw127@hw.ac.uk

Abstract

This should be a 6-8 page conference paper with appendices, if relevant. Good reports from last year: 1 and 7

1 Introduction

from coursework spec: main research or technical question addressed
Socially assistive robots (SARs) are a crucial part of the future of many sectors, for example, in education or healthcare (Gunson et al., 2022). Especially the latter depends on technology advancements as it is facing numerous obstacles in the future, such as increasing spendings and a growing percentage of older people. A serious lack of healthcare workers is already occurring, with 10 million more healthworkers needed worldwide by 2030 (Cooper et al., 2020; WHO, 2023). SARs can pose a solution to the problem, as they are able to support healthcare in various ways, such as encouraging older people to keep living independently for longer or reducing caregiver burden (Cooper et al., 2020).

These scenarios require SARs to be able to handle multi-party interactions as it is likely that more than one person will interact with the system. Compared to handling dyadic interactions, handling multi-party conversations includes more complex challenges, such as Speaker Recognition, Addressee Recognition, Response Selection (summarised in “who says what to whom“) and turn-taking (Addlesee et al., 2023; Johansson and Skantze, 2015).

Include here what exactly we examined about turn-taking

In this work, we propose a model trained on multi-party human-human conversation data. We collected the data from recordings of special “Who wants to be millionaire?“ episodes where two candidates collaborated to answer the host’s questions.

Include results here.

2 Background

from coursework spec: literature review / related work, including a critical analysis of the field, and commentary on applicability of the technologies and methods used in emerging technologies and application areas

2.1 Socially Assistive Robots

For healthcare, as well as for any other sector, the difficulty of successfully designing SARs lies in creating robots that can effectively converse with humans and adhere to social norms (Moujahid et al., 2022). The more expressive a robot is, the more it will be perceived as intelligent, conscious and polite (Moujahid et al., 2022). To achieve such a positive perception, multiple parts need to be combined into one conversational system, such as the ability to carry out visually grounded as well as task-based dialogues, to perceive and discuss its environment and to chit-chat (Gunson et al., 2022).

The SPRING project conducts research on a SAR robot that is deployed in an eldercare hospital reception area (Addlesee et al., 2020a). The conversational system is deployed on humanoid ARI robot produced by Pal Robotics (Robotics, 2023). ARIs capabilities can be extended with custom AI algorithms, in the case of SPRING-ARI a visual perception system, a dialogue system, and a social interaction planner (Addlesee et al., 2020a). While the SPRING-ARI system successfully demonstrates that task-based, social and visually grounded dialogue can be combined with physical actions, it still lacks the ability to handle conversations with more than one person simultaneously (Addlesee et al., 2020a).

2.2 Multi-party Human Robot Interaction

As stated above, the endeavour to create conversational systems becomes considerably more difficult when dealing with multi-party interactions (Addlesee et al., 2023). Especially turn-taking poses a central problem. It is defined as follows:

> The rules of turn-taking organize the conversation into turns, during which one of the participants has the right to speak while the others agree to listen (Żarkowski, 2019)

In dyadic conversations, there are only two roles a participant can take: speaker or listener, hence it is clear when and to whom the turn is yielded. In multi-party conversations, participants can take multiple roles, therefore turn-taking needs to be coordinated (Johansson and Skantze, 2015). Humans signal their intents mostly through gaze, but also through pauses, prosody, and body positioning (Żarkowski, 2019). To copy this behaviour, earlier models for conversational systems relied on silence time-outs to coordinate turn-taking, however, this approach is found to be too simplistic (Skantze, 2021). Instead, mimicking human turn-taking behaviour better by using a combination of verbal and non-verbal cues leads to robots that are better perceived (Moujahid et al., 2022).

State exactly the gap that we will fill - whatever that will be

3 Data Collection

- (Laura) Talk about multi-party data collection
- (Aron and Katie) How we collected our data, describe the intents we used to label the data
- (Aron and Katie) Cohen's Kappa for our data collection method: probably need to annotate a couple of transcripts twice for reporting on this

Data collection was performed in-house, by the team. We first collected all available recordings of Who Wants to Be a Millionaire with two participants. Using the YouTube API, we managed to get the transcripts for the vast majority of recordings. To annotate these transcripts, we came up with the following unified set of annotations that would allow us to capture as much information as possible, without saturating the data.

- question - The system presents the question
- options - The system presents the options
- chit-chat - Speech not related to the quiz
- offer-answer() - A player presents an answer to the other player
- offer-to-answer - A player signals that he/she knows the answer
- check-answer - A player checking if the other player knows the answer
- agreement - Agreement between players about the answer
- ask-agreement - A player asks the other player for confirmation about their answer
- accept-answer System considers answer the final answer
- final-answer() - Players give final answer
- confirm-agreement - System tries to confirm the final answer with participants
- confirm-final-answer - Participants confirm their answer is final

3.1 Cohen's Kappa Coefficient

As the system uses the annotations to train the system, ensuring that they are reliable and consistent is essential for accurate training. Cohen's kappa was used on a sample of the completed transcripts to evaluate their accuracy. Cohen's kappa measures the reliability between raters on categorical data. It is said to be more accurate than calculating the agreement as it accounts for agreements happening by chance.

add source

To calculate Cohen's kappa, a sample of four transcripts were re-annotated by a team member, without them seeing the original annotations to remove any possible bias. This amounts to approximately 15% of the total transcripts.

Every occurrence by both raters of each annotation label was gathered. An overall agreement rate was calculated using -

$$Agreement = \frac{total\ Agreed}{total} \tag{1}$$
$$Agreement = 0.9684$$

source

This alone is not reliable enough to trust. To calculate the chance of an agreement happening between the two raters, first the probability of a rater choosing a label was calculated. This was done for each rater for every label using -

$$Prob(R_x[label]) = \frac{R_x\ Total\ for\ [label]}{total} \tag{2}$$

Then the chance of agreement for each annotation label was calculated using -

$$\begin{gathered} ChanceAgreement[label] = \\ Prob(R_1[label]) * Prob(R_2[label]) \end{gathered} \tag{3}$$

The overall chance of agreement was calculated by summing the chance of agreements for each annotation label.

Finally, the kappa score was calculated.

$$KappaScore = \frac{Agree - ChanceAgree}{1 - ChanceAgree} \tag{4}$$
$$KappaScore = 0.9601$$

add source

According to Cohen , a Kappa score of 0.9601 is interpreted as almost perfect agreement. From this, the annotation of transcripts can be concluded as reliable and therefore the training model will accurately learn from this data.

4 Design and Implementation

- *from coursework spec:* design and implementation of the system: components and architecture
- Jack's System Graph

4.1 Automatic Speech Recognition

To enable interaction between users and the conversational system, the first step is to transform user's speech into text, in order to pass text further onto intent recognition (NLU). Transforming audio into text works through Speech-To-Text (STT)software. In recent years, STT systems are becoming more accurate and fast, however, research found that none of the existing systems can yet reliably handle conversations in real time (Addlesee et al., 2020b).

Get rid of pronoun

The system we have built required two things specifically from the ASR that are not available in all ASR systems. The first being real time transcription. As our system is going to be used on a robot, it must transcribe what the user is saying in real-time to avoid long delays before the system responds. For a usable system in this context the time taken for a system to respond can be no longer than the delay a human would leave before responding or only marginally longer. In addition, our system requires diarization. Diarization is the process of taking a single audio file, in which two or more people are speaking, and determining who is speaking at what time. As two users would be speaking to each other and to our system, our system must recognise when each individual is talking. This would allow us to track the intents of each user and determine when they have agreed with each other. We tried many of the most popular STT systems including Amazon's Transcribe, IBM's Watson and locally running Pyannote. These are all valid options for transcription but lack real-time diarization, that is, diarising the speech as it is spoken. Finally, we settled with Google's Cloud Speech-to-Text API due to its high accuracy, real-time transcription capabilities, and customisability. Also, as it is widely used, resources available for troubleshooting and integration were readily available.

Those two sentences have the same content

Get rid of pronoun

As well as these reasons, Google's API boasted diarization capabilities, along with its real-time transcription capabilities seemed to fit our use case. However, in use, diarization was inaccurate and often grouped two separate speakers together or split sentences up seemingly at random. This became even more apparent when two users were speaking over each other. This is in accordance with the findings of (Addlesee et al., 2020a). Diarization is quite resource-heavy meaning that running a diarization algorithm in real time takes an extreme amount of resources and, unfortunately, is not feasible on normal, consumer-grade hardware. And although diarization APIs exist, they do not respond fast enough for our use case. This made Google's diarization unusable for our system, therefore we decided to handle the diarization separately and integrate it with the real time Google transcription.

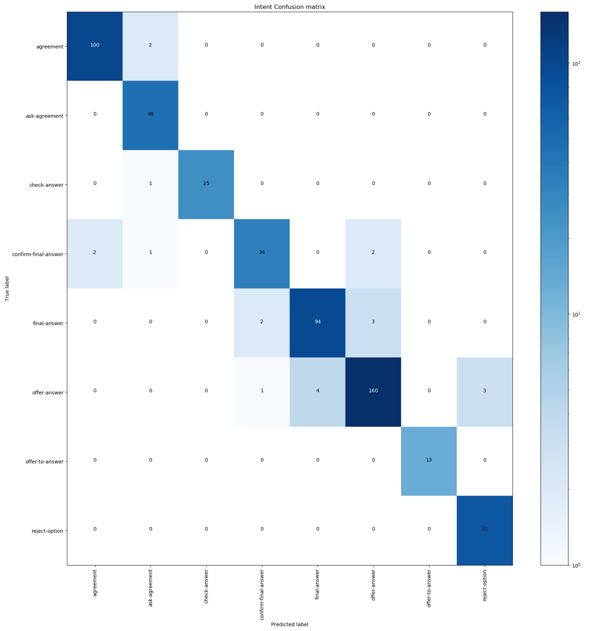

Figure 1: Intent Recognition Confusion Matrix

4.2 Natural Language Understanding

- Short description of RASA
- Explain the aim (detecting intents)
- Refer to the intents described in 3 Data Collection
- Evaluation - show different versions of the model: difference in F-Score, Confusion Matrix, ...

4.3 Dialogue Management

- Clearly explain 2 parts (State-Machine and NN)
- State-Machine: high-level control, handles things we have no data for, eg. pauses
- NN: Report on differences between RNN and LSTM and why the choice for the LSTM has been made
- (optional): Comparison to RASA rule-policy

4.4 Natural Language Generation

To make the system feel more unique and less robotic, we made use of Natural Language Generation (NLG) for the phrases that the "host" says to the participants. We used OpenAI's API, specifically "gpt-3.5-turbo", the same version that is used in ChatGPT. This allowed us to prompt for different outputs for the system that convey the same information. For example, when receiving a correct answer, "Yes, that's it! Well done!" and "You got it! Great job!" are both possible outputs. There are 50 different options for each response the "host" can say.

5 Evaluation

from coursework spec: evaluation of the system and presentation of the results

5.1 Methodology

In this study, we performed extrinsic and intrinsic evaluations. The extrinsic evaluation focused on both subjective and objective measures of the system's performance. The subjective measures included the user's enjoyment and perception of the system's natural behaviour, while the objective measures included the number of turns taken and the agreement rate. Additionally, the correlation between correct answers and enjoyment was also examined. Overall, the evaluation aimed to assess the effectiveness of the system in engaging users and providing accurate responses.

The evaluation focused on intrinsic measures of individual components in a multi-party conversational system. The components were assessed separately, with ASR being evaluated using the word error rate, NLU using precision, recall, accuracy, and F1 score, DM IDK YET, and NLG using n-gram-based overlap with BLEU. The aim was to assess the performance of each component and identify areas of improvement.

Additionally, the evaluation also aimed to test the hypothesis that using verbal cues instead of silence cues in a multi-party conversational system increases user interaction and satisfaction with the system. This hypothesis needed to be statistically proved or disproved through significance testing.

5.2 Experiment Layout

The experiment followed a between-subjects design. Every participant was required to read and sign a consent form before they can play the quiz. This was an in-person experiment, with the quiz running on a laptop, where participants can see the questions and the options. Members of the experiment were required to play the quiz at least once. However, they were encouraged to play as many times as they can. After they no longer wished to play, they were asked to complete a questionnaire about their experience. The questionnaire queried them on their experience using a five-point Likert scale.

5.3 Results

- ASR results
- NLU result
- DM result NLG result
- questionnaire result

6 Conclusion

6.1 Ethical Reflection

7 Future Work

from coursework spec: suggestions for future work

To further improve on NLG within this system, some content moderation could be performed on the generations, to ensure that there are no inappropriate outputs, this can be done entirely within OpenAI's API. Also, the system could be updated to make use of GPT-4, which at this current time is not publicly available.

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowledgments section. Do not include this section when submitting your paper for review.

References

[Addlesee et al.2020a] Angus Addlesee, Yanchao Yu, and Arash Eshghi. 2020a. A comprehensive evaluation of incremental speech recognition and diarization for conversational AI. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3492–3503. International Committee on Computational Linguistics.

[Addlesee et al.2020b] Angus Addlesee, Yanchao Yu, and Arash Eshghi. 2020b. A comprehensive evaluation of incremental speech recognition and diarization for conversational ai. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3492–3503, Barcelona, Spain (Online), Dec. International Committee on Computational Linguistics.

[Addlesee et al.2023] Angus Addlesee, Weronika Sieinska, Nancie Gunson, Daniel Hernández García, Christian Dondrup, and Oliver Lemon. 2023.

[Cooper et al.2020] Sara Cooper, Alessandro Di Fava, Carlos Vivas, Luca Marchionni, and Francesco Ferro. 2020. ARI: the social assistive robot and companion. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 745–751. ISSN: 1944-9437.

[Gunson et al.2022] Nancie Gunson, Daniel Hernandez Garcia, Weronika Sieińska, Angus Addlesee, Christian Dondrup, Oliver Lemon, Jose L. Part, and Yanchao Yu. 2022. A visually-aware conversational robot receptionist. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 645–648. Association for Computational Linguistics.

[Johansson and Skantze2015] Martin Johansson and Gabriel Skantze. 2015. Opportunities and obligations to take turns in collaborative multi-party human-robot interaction. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 305–314, Prague, Czech Republic, Sep. Association for Computational Linguistics.

[Moujahid et al.2022] Meriam Moujahid, Helen Hastie, and Oliver Lemon. 2022. Multi-party interaction with a robot receptionist. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 927–931.

[Robotics2023] Pal Robotics. 2023. Ari - the social and collaborative robot. Accessed on: 2023-02-08.

[Skantze2021] Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech and Language*, 67:101178.

[WHO2023] WHO. 2023. Global health workforce statistics. Accessed on: 2023-02-07.

[Żarkowski2019] Mateusz Żarkowski. 2019. Multiparty turn-taking in repeated human–robot interactions: An interdisciplinary evaluation. *International Journal of Social Robotics*, 11(5):693–707, Dec.

A Supplemental Material, Appendix