Evaluation of a Spoken Argumentative Dialogue System for Opinion-Building

Annalena Aicher and **Wolfgang Minker**
Institute for Communications Engineering
Ulm University, Germany
annalena.aicher@uni-ulm.de

Stefan Hillmann and **Thilo Michael** and **Sebastian Möller**
Quality and Usability Lab
TU Berlin, Germany

Stefan Ultes
Mercedes Research & Development
Sindelfingen, Germany

Abstract

Speech interfaces for argumentative dialogue systems (ADS) are rather scarce and quite complex. To provide a more natural and intuitive interface, we include an adaption of a recently introduced natural language understanding (NLU) framework tailored to argumentative tasks into a complete end-to-end ADS. Within this paper, we investigate the influence of two different I/O modalities and discuss issues and problems we encountered in a user study with 202 participants using our ADS.

1 Introduction

The exchange of arguments and conversation with humans via natural language demand for a flexible natural language understanding (NLU), an argumentative dialogue structure, and the integration of commonsense knowledge. The speech-driven argumentative dialogue system (ADS) we introduce in this paper combines these components and enables the user to scrutinize arguments on both sides of a controversial topic. Unlike most approaches to human-machine argumentation (Slonim et al., 2021; Rosenfeld and Kraus, 2016; Le et al., 2018; Rakshit et al., 2017; Chalaguine and Hunter, 2020; Fazzinga et al., 2021) we pursue a cooperative exchange of arguments. Our aim is a system that cooperatively engages the users to explore arguments and to state their preferences in natural language. Therefore, we modified and extended our previously introduced menu-based ADS (Aicher et al., 2021). The speech-based system is evaluated and compared to the robust baseline in terms of naturalness and usability aspects in a crowd-sourcing study with 202 participants.

2 ADS Interface and NLU Framework

The system's graphical user interface (GUI) is illustrated in Figure 1. The interface can either provide a drop-down menu or speech input. In the drop-down system, users can choose their action by clicking. In the speech system, our formerly introduced NLU framework (Abro et al., 2022) processes the spoken user utterance. Its intent classifier uses the BERT Transformer Encoder presented by Devlin et al. (2018) and a bidirectional LSTM classifier.

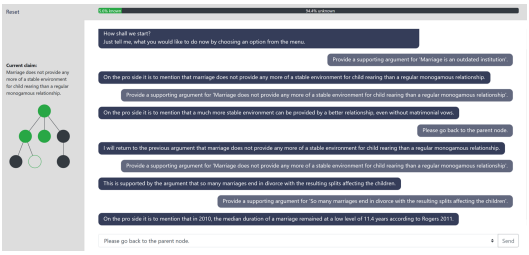

Figure 1: GUI of the menu system with folded drop-down menu. Above the drop-down menu the dialogue history is shown.

In the speech-based system, instead of the drop-down menu, a button labelled "Start Talking" is shown. The button is pressed to start and stop the speech recording. Except for this difference, both systems share the same architecture. The system's generated output is based upon the original textual representation of the argument components on the sample debate topic *Marriage is an outdated institution* taken from *Debatabase* of the idebate.org[1] website.

3 User Study Setting

The study was conducted online via the crowd-sourcing platform "Crowdee" (www.crowdee.com, 12-29th November 2021) with participants from the UK, US and Australia. All 202 participants (menu: 104 [50 female, 54 male], speech: 98 [39 female, 59 male]) were non-experts without a

[1] https://idebate.org/debatabase (last accessed 23th February 2022). Material reproduced from www.iedebate.org with the permission of the International Debating Education Association. Copyright © 2005 International Debate Education Association. All Rights Reserved.

topic-specific background. After an introduction to the system (short text and demo video), the users had to listen to enough arguments to build a well-founded opinion. As soon as ten arguments were heard, the end of the interaction could be chosen freely. Afterwards, the participants had to rate the interaction in 40 statements[2] on a 5-point Likert scale (1 = totally disagree, 5 = totally agree).

4 Results and Discussion

In average the interaction with the system last 31.45 minutes (menu: 27.57 speech: 35.34). This significant difference can be explained by the fact that the spoken interaction (speaking and hearing) inherently takes longer than clicking on an option in the drop-down menu and reading the response. Another significant difference is observable in the number of heard arguments (average menu/speech: 22/15). Even though the average time the users of the menu system interacted with the ADS is lower, the number of provided arguments is significantly higher compared to the speech system. 9.6%/17.3% of the menu/speech system participants quit the conversation after hearing the minimum number of 10 arguments (in total: 13.4%). Most of the participants heard between 20-30 arguments of 72 available arguments. Whereas some participants in the menu system listened to even more arguments, only one participant of the speech system did so. The category "Overall Quality" ("What is your overall impression of the system?") is rated on a specific 5-point Likert scale (5 = Excellent, 4 = Good, 3 = Fair, 2 = Poor, 1 = Bad). We perceive a highly significant[3] ($\alpha < 0.01$) difference between both systems, as the menu system with a rating of 3.49 outperformed the speech system rated with 2.66. Altogether, the speech system is significantly outperformed in all categories of the questionnaire. The biggest differences were perceivable in ratings concerning errors which occurred or whether the system provided the expected information. Clearly, this points to a lack in processing the user utterances (errors in the ASR or NLU module). By checking the dialogue logs of the interactions with users in the speech system, we found that about 15% of all speech utterances were processed erroneously. Even though in 70% of these cases the NLU identified the correct intent, the results show that this has had a considerable impact on user perception of the speech system. Furthermore, we noticed inconsistencies in the user behavior, e.g. repetition of requests multiple times and ignoring the system's answer to choose another action. In contrast to the menu system which only displayed the possible actions, the speech users had to figure out what actions they can perform and formulate them. Even though the speech system offered a "Help" button, as well as the "available options" action, only 1.3% of the participants used them. This might be explained by the fact that only 35% of the users spend enough time on the introduction website to read through the explanation and watch the video properly. This is further underpinned by users' feedback, stating that "It was not possible to do what I wanted to do. I repeated myself many times"/"I was stuck in the argument and could not get back.". The results show that the I/O modalities and respective difficulties/problems decrease the rating of the general impression of the system, even in aspects which have no relation to the former. E.g. the incremental approach to present arguments, the sufficiency of different options or the conclusiveness of arguments which are content- but not modality-dependent, are rated significantly worse in the speech than in the menu system. Therefore, it is crucial to solve the identified issues and to introduce a double-staged study setting, which ensures the participants understood how to interact with the system. Even though the introduced speech system does not outperform the menu baseline, we could show that the menu system provides a robust baseline that tends to be rated positively in almost every question. Thus, it suits as a robust baseline to which enhanced spoken ADS versions can be compared to.

5 Conclusion and Future Work

In this work, we evaluated an ADS in two I/O modalities by conducting a crowdsourcing study. Due to an erroneous ASR module and issues in understanding how to communicate with our ADS via speech, we observed that the latter was outperformed significantly by our strong menu baseline. In future work, we will enhance the system's ASR and NLU robustness by training on larger data-sets and including a request for repetition if the intent prediction accuracy falls below a threshold.

[2]Taken from a questionnaire according to ITU-T Recommendation P.851 (P.851, 2003)

[3]To determine whether the difference between the two system means is significant, we used the non-parametric Mann-Whitney U test (McKnight and Najab, 2010) for two independent samples with no specific distribution.

References

Waheed Ahmed Abro, Annalena Aicher, Niklas Rach, Stefan Ultes, Wolfgang Minker, and Guilin Qi. 2022. Natural language understanding for argumentative dialogue systems in the opinion building domain. *Knowledge-Based Systems*, 242:108318.

Annalena Aicher, Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2021. Opinion building based on the argumentative dialogue system bea. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, pages 307–318. Springer.

Lisa A. Chalaguine and A. Hunter. 2020. A persuasive chatbot using a crowd-sourced argument graph and concerns. In *COMMA*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bettina Fazzinga, Andrea Galassi, and Paolo Torroni. 2021. An argumentative dialogue system for covid-19 vaccine information. In *Logic and Argumentation*, pages 477–485, Cham.

Dieu Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130.

Patrick E. McKnight and Julius Najab. 2010. *Mann-Whitney U Test*, pages 1–1. American Cancer Society.

ITU-T Recommendation P.851. 2003. Subjective quality evaluation of telephone services based on spoken dialogue systems (11/2003). International Telecommunication Union.

Geetanjali Rakshit, Kevin K. Bowden, Lena Reed, Amita Misra, and Marilyn A. Walker. 2017. Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents - 8th International Workshop on Spoken Dialog Systems*, pages 45–52.

Ariel Rosenfeld and Sarit Kraus. 2016. Strategical argumentative agent for human persuasion. In *ECAI'16*, pages 320–328.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, and Lilach Edelstein. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.