Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1234–1241
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Towards Building a Spoken Dialogue System for Argument Exploration

Annalena Aicher[1], Nadine Gerstenlauer[1], Isabel Feustel[1], Wolfgang Minker[1], Stefan Ultes[2]
[1]Ulm University, Albert-Einstein-Allee 43, 89075 Ulm
[2]Mercedes Benz AG, Stuttgart, Germany
annalena.aicher@uni-ulm,de

Abstract

Speech interfaces for argumentative dialogue systems (ADS) are rather scarce. The complex task they pursue hinders the application of common natural language understanding (NLU) approaches in this domain. To address this issue we include an adaption of a recently introduced NLU framework tailored to argumentative tasks into a complete ADS. We evaluate the likeability and motivation of users to interact with the new system in a user study. Therefore, we compare it to a solid baseline utilizing a drop-down menu. The results indicate that the integration of a flexible NLU framework enables a far more natural and satisfying interaction with human users in real-time. Even though the drop-down menu convinces regarding its robustness, the willingness to use the new system is significantly higher. Hence, the featured NLU framework provides a sound basis to build an intuitive interface which can be extended to adapt its behavior to the individual user.

Keywords: Natural Language Understanding (NLU), Argumentative Dialogue Systems (ADS), Conversational Systems, Speech Interaction/Recognition, User Usability/Satisfaction, HCI, Preference Modelling

1. Introduction

Building dialogue systems that can converse with humans via natural language is a challenging yet intriguing problem of artificial intelligence. Most of the popular and accessible virtual agents (VA) are adequately trained to handle simple conversations, e.g. inquiry for hotels, restaurants etc. (Saha et al., 2020). However, such VAs are inept in managing complex dialogues especially with regard to demanding conversations, like discussing a controversial topic and providing logically consistent arguments. For example, if the user requests for a supporting argument on a certain topic, they might more likely engage in a discussion if they are able to ask “Could you please be more specific on this argument. I do not see why...” instead of simple, short commands like “Why?” or “More information!”. Furthermore, a natural language system response like “An argument in favor of the previously discussed statement is, that... which is underpinned...” may be more appealing than a written system response.

Such complex tasks demand for a flexible natural language understanding (NLU), an intuitive dialogue structure and the integration of commonsense knowledge. These components are combined in the speech-driven argumentative dialogue system (ADS) BEA (’Building Engaging Argumentation’ (Aicher et al., 2021)) we introduce in this paper. To the best of our knowledge this is the first ADS which tries to cooperatively engage the user to explore arguments in natural language. Therefore, we investigate to what extent a spoken interaction increases the willingness to explore arguments with the aid of this system in a user study. In order to examine the perception of the system with regard to dimensions such as likeability, naturalness, engagement etc. we compare it to a robust baseline consisting of the same ADS but with menu input via mouse click. Furthermore, we want to get an insight into how human users interact with such a system, what they request and how they formulate it, as well as what their expectations are. We choose a limited set of possible requests, to on the one hand ensure the comparability to the baseline system and on the other give the users the chance to suggest extensions and possible advances.

The remainder of the paper is as follows: Section 2 gives an overview over related work and the Section 3 describes the architecture with a focus on the integration of the NLU framework. Section 4 introduces and discusses the experimental setting of the study. Subsequently the evaluation results are discussed in Section 5. We close with a conclusion and a brief discussion of future work in Section 6.

2. Related Work

A natural way of resolving different points of view or forming an opinion for humans is through spoken conversation, i.e., through the exchange of arguments and knowledge. Filter algorithms aim to reduce the numerous opinions and information on almost every topic available online by taking previous user requests into account to serve the users’ interest. As a result, people tend to focus on a biased subset of sources that repeat or strengthen an already established or convenient opinion (Pariser, 2011). In order to avoid this (often unconscious) process of intellectual isolation, we propose an approach to explore a vast amount of different information in a natural and intuitive way using natural language. To this end the envisioned system engages in a deliberative dialogue with a human user in order to support a fair and unbiased opinion building process. Thus, we pursue a cooperative exchange of arguments via spoken language without trying to persuade or winning a debate against the user unlike most approaches to human-machine argumentation in the literature which are embedded in a competitive scenario. Those approaches utilize different models to structure the interaction. Slonim et al. (2021)

use a classical debating setting. Their IBM Debater is an autonomous debating system that can engage in a competitive debate with humans via natural language. The opponent utterances are analyzed automatically and a suitable response with a fixed length is generated. Another speech-based approach was introduced by Rosenfeld and Kraus (2016) presenting a system based on weighted Bipolar Argumentation Frameworks (wBAG). Arguing chatbots such as Debbie, which used a similarity algorithm to retrieve counterarguments (Rakshit et al., 2017) and Dave, that used retrieval- and generative-based models (Le et al., 2018) interacted via text with the user. A menu-based framework that incorporates the beliefs and concerns of the opponent was presented by Hadoux and Hunter (2021). In the same line, Chalaguine and Hunter (2020) used a previously crowd-sourced argument graph and considered concerns of the user to persuade them.

In contrast to our system, none of the aforementioned ADS tried to cooperatively engage the user to explore arguments and stating their preferences in natural language. Thus, we extend the menu-based ADS we introduced in (Aicher et al., 2021) which provides a non-competitive setting. We undertake their approach to use explicit feedback to estimate the (overall) preference considering wBAGs (Amgoud and Ben-Naim, 2016; Amgoud and Ben-Naim, 2018a). Our system is able to react to user utterances (opinions, preferences, requests and questions) and enables the user to conduct a deeper research to find diverging views and evidence of the pros or cons of an argument. The transparent presentation of pros and cons towards the discussed topic contributes to explainability and deep understanding.

3. ADS Architecture

In the following, we outline the main components of the architecture of our ADS. After explaining the user and dialogue model, the interface and NLU framework are introduced. A schema of the entire architecture of BEA is given in Figure 1.

3.1. User Model and Knowledge Base

The herein used user model estimates user preferences[1] based on bipolar argument structures. The utilized annotation scheme was introduced for annotating argumentative discourse structures and relations in persuasive essays by Stab and Gurevych (2014). They structure arguments in three components: *major claim*, *claim* and *premise*. The overall topic of the debate is formulated as the *major claim* representing the root node in the graph. *Claims* on the other hand are assertions which formulate a certain opinion targeting the *major claim* but still need to be justified by further arguments, *premises* respectively. We consider two relations between these argument components (nodes), *support* or *attack*. We choose a non-cyclic tree structure, where each node ("parent") is supported or attacked by its "children". If no children exist, the node is a leaf and marks the end of a branch. An example of such a bipolar argument tree is shown in Figure 2. Following the previous approach (Aicher et al., 2021), we utilize wBAGs in which a *weight* is assigned to each argument. Likewise we utilize the Euler-based restricted semantics introduced by Amgoud and Ben-Naim (Amgoud and Ben-Naim, 2018b), which aggregates the strength of arguments in a linear fashion. The energy E_i (also Euler's number) at an argument i is de-

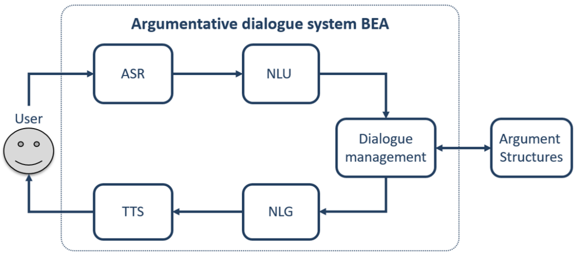


Figure 1: Architecture of BEA. After the user's spoken input is processed by the automated speech recognition module (ASR), it is passed to the Natural Language Understanding unit, which extracts the respective information. This abstractly represented information can be processed by the dialogue management, which decides a suitable corresponding system response by interacting with an argument structure. Once an appropriate response is selected it is processed by a Natural Language Generation (NLG) module which formulates its textual representation and finally presented to the user in natural language by Text-to-Speech (TTS) module. In case of the baseline system the ASR and TTS modules were omitted.

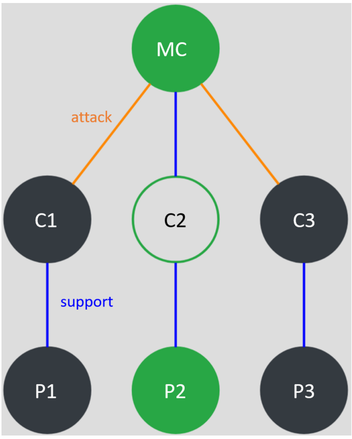


Figure 2: Visualization of argument tree structure. MC denotes the major claim (root node), which is attacked by the claim C1 and C3 (denoted by a orange line) and supported by claim C2 (denoted by a blue line). The respective leaf nodes are the premises P1, P2 and P3

---

[1] Here preferences includes the following opinion statements: preference, indifference and rejection.

Move	Description	Determiners
$stance$	Stance on current argument	Always
$stance\ overall$	Stance on overall topic	Always
$level\ up$	Switches to parent argument	Always (except for root)
$exit$	Terminates conversation	Always
why_{pro}	Information-seeking for a pro argument	If supporting child node exists
why_{con}	Information-seeking for a con argument	If attacking child node exists
$jump\ to\ argument$	Switches to referred argument	If argument is skipped without preference
$prefer$	Preference of presented argument; update weights	Always (except for root)
$prefer\ old/$ $new/equally$	Compare preference between current and former argument(s); update weights	If sibling(s) are preferred
$indifferent$	Indifference towards presented argument; update weights	Always (except for root)
$reject$	Current argument and children are abandoned; weights updated	Always (except for root)
$number\ visited$	Shows stats on number of seen and unseen arguments	Always
$available\ moves$	Shows all available moves	Only in speech-based system

Table 1: Explanation for the available moves the user can choose from in each turn of the interaction.

fined as (Amgoud and Ben-Naim, 2018b):

$$E_i = \sum_i s_{i,\text{pro}} - \sum_i s_{i,\text{con}} \tag{1}$$

where "i,pro" and "i,con" are the sets of supporters and attackers and s_i denotes the corresponding strength value. Hence, the stronger or more-numerous the supporting argument components are, the greater and more-likely-positive is that exponent (and vice versa for attackers). The aggregated strength of an argument component i is a function of its initial weight ω_i and the energy in (1) (Amgoud and Ben-Naim, 2018b):

$$s_i = 1 - \frac{1-\omega_i^2}{1+\omega_i e^{E_i}}. \tag{2}$$

Consequently, s_i considers the weight of the component itself as well as the influence of its children. If a component is leaf node, its strength equals its weight. After a preference is expressed, the corresponding weight is adjusted according to an update function and iterated through all parent nodes following Equation 2. If the user states a preference/indifference/rejection towards component i, its strength is updated such that:

$$\text{preference:} \quad s'_i = s_{i,\text{max}} + \frac{n_\text{v}}{n_\text{a}}\left(1 - s_{i,\text{max}}\right) \tag{3}$$

$$\text{rejection:} \quad s'_i = 0 \tag{4}$$

$$\text{indifference:} \quad s'_i = 0.01, \tag{5}$$

where $s_{i,\text{max}}$ denotes the maximum strength of all siblings of argument i. In case no sibling argument has been preferred yet $s_{i,\text{max}} := 0.5$. n_v describes the number of sibling arguments of argument i which have already been presented to the user and n_a denotes the total number of all sibling arguments. Hence, the preference update takes into account how many sibling arguments have already been heard in relation to the ones available. In particular, this means that the more sibling arguments are heard before one of them is preferred, the greater is the impact of this preference on the weight. According to Wilcock and Jokinen (2021) in scenarios that do not adhere to a clear structure regarding speaking time and turn taking (like debates), extensive utterances presented by synthetic speech are hard to follow and understand. To prevent the user from being overwhelmed by the amount of information, in contrast to our previous work (Aicher et al., 2021) we introduce the available arguments incrementally depending on the user's request. Thus, the users can state their opinion before all sibling arguments have been presented. As we formally update only weights in order to consider later preferences, we determine the new weight (solution of Equation 2) as

$$\omega'_i = \frac{e^{E_i}\left(1-s'_i\right)}{2}\left(-1 \pm \sqrt{1 + \frac{4s'_i}{e^{2E_i}\left(1-s'_i\right)^2}}\right). \tag{6}$$

Due to the square root, there are two solutions to this equation but only one meets this in the required interval $[0,1]$. If an argument is rejected, its weight (and thus its strength as well) is set to 0. The whole opinion model is updated after each expressed preference according to the following scheme:

1. According to the expressed user's preference regarding a certain argument i (node), its new strength s'_i is calculated.
2. The new energy of the parent node of i is determined by applying Eq. (1). Using the new energy of the parent of i calculated in the previous step and the update formula in Eq. (2), the new strength of the parent node is calculated.
3. Step two is repeated until all related values are updated.

The overall stance of the user can then be determined by calculating the energy of the major claim. As no information about the user's preferences is known before the interaction, we initialize all weights with the same value ($\omega_0 = 0.01$, representing indifference).

The nodes are updated recursively until the root node is reached. The difference between the sum of the strengths of the supporting children and the sum of the strengths of the attacking children of the root node displays the final "stance" of the user (Aicher et al., 2021). If it is greater than 0, the user supports the major claim, if it is smaller than 0, the user rejects it, and if it equals 0, the user is indifferent.

A suiting argument structure is provided by a sample debate on the topic *Marriage is an outdated institution* which was thoroughly discussed by Rach et al. (2018). It serves as knowledge base for the arguments and is taken from the *Debatabase* of the idebate.org[2] website. It consists of a total of 72 argument components (1 *major claim*, 10 *claims* and 61 *premises*) and their corresponding relations and encoded in an OWL ontology (Bechhofer, 2009) for further use. In each "why pro/con" move a single argument component is presented to the user. The depth of a branch varies from 5 up to 10 argument components. Due to the generality of the annotation scheme, the system is not restricted to the herein considered data. In general, every argument structure that can be mapped into the applied scheme can be processed by the system.

3.2. Dialogue Model

The interaction between the system and the user is separated in turns, consisting of a user action and corresponding natural language answer of the system. Table 1 shows the possible actions (moves) the user is able to choose from. One can navigate through the argument tree (explore argument branches), express preferences and enquire more information. The determiners show which moves are available depending on the position of the current argument (root / parent node / "leaf" node).

After the system's greeting the resulting dialogue is determined only by the user and their choices. The system response is based on the original textual representation of the argument components, which is embedded in moderating utterances. To support the impression of a natural conversation and to engage the user, personalized system responses are used, e.g. "I understand you like.../ I think it is interesting that/ So just try...".

In Table 2 an exemplary dialogue shows how an interaction with a user looks like and Figure 2 illustrates the respective bipolar argument tree. By preferring the strengths and weights of $P2$ and its respective parent nodes $C2$ and MC are iteratively updated as described in Subsection 3.1. Note, that only heard arguments are included in the calculation, and thus the unheard nodes ($C1, C3, P1, P2$) are not considered. Following the update algorithm described in Subsection 3.1 it follows for the new strength of $P2$ that $s'_{P2} = 0.5 = \omega_{C2}$ and thus, with $E'_{C2} = 0.5$, that $s'_{C2} = 0.589$ and $\omega'_{C2} = 0.5$. For MC it follows that $E_{MC} = 0.5 > 0$ and thus, the user's stance towards the major claim is supported by the user.

3.3. Interface and NLU Framework

Our aim is to investigate the effect of spoken language interaction when compared to a graphical I/O. Therefore, we introduce a NLU framework for speech input as well as a clickable menu for the baseline system.

Instead of the drop-down menu displayed in Figure 3, a button with ´´Talk with BEA" is shown in the speech-based system, which is pressed to start and stop the speech recording, see Figure 4. Except for this difference both systems share the same architecture. The dialogue history is shown above the menu/button, where the system's responses are shown on the left and the user's chosen moves on the right side. To provide the user a visual cue to the length of the ongoing conversation, a progress bar above the dialogue history shows the number of arguments that were already discussed and how many are still unknown at each stage of the interaction.

To provide a suitable system response during the speech interaction, it is crucial to understand the user's intent and if an argument is addressed. Therefore, we integrate the previously introduced NLU framework (Abro et al., 2021). It uses an intent classifier model consisting of two main components: the BERT Transformer Encoder (Devlin et al., 2018) and a bidirectional LSTM classifier. Furthermore, to identify the arguments a user refers to, a similarity model is applied which is based on Sentence-BERT (SBERT) (Reimers and Gurevych, 2019). We (Abro et al., 2021) showed that this NLU framework outperforms state-of-the-art baselines and thus, is the most suitable for our purpose. The system-specific intents have to be trained once but no pre-training for new topics or user groups

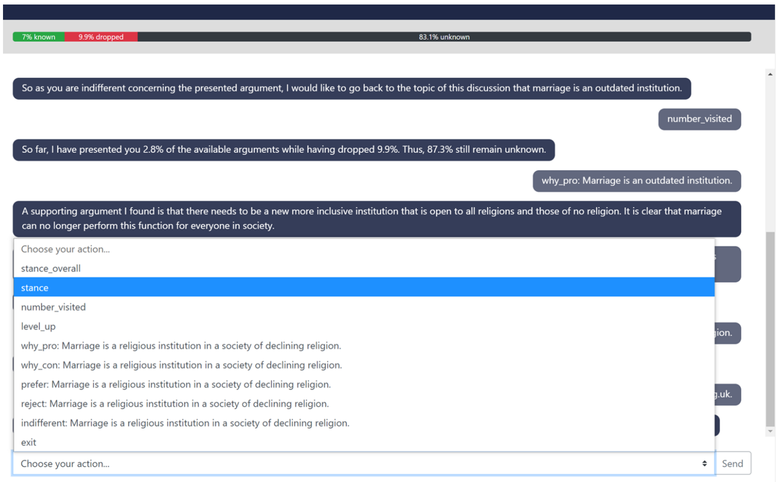
---

[2] `https://idebate.org/debatabase` (last accessed 23th July 2020).

Material reproduced from `www.iedebate.org` with the permission of the International Debating Education Association. Copyright © 2005 International Debate Education Association. All Rights Reserved.

Actor	Utterance	Move
System	*Let's discuss the topic that "Marriage is an outdated institution."*	
User	*Please elaborate this further.*	$why_{con}(MC)$
System	*On the pro side it is to mention that there needs to be a new more inclusive institution that is open to all religions and those of no religion. It is clear that marriage can no longer perform this function for everyone in society.*	
User	*Could you explain why?*	$why_{pro}(C2)$
System	*This is supported by the argument that marriage is a religious institution in a society of declining religion.*	
User	*This sounds reasonable to me*	$prefer(P2)$
System	*I think it is interesting that you prefer this argument.*	
User	*What is my calculated stance on the discussion topic.*	$stance(MC)$
System	*At the moment you seem to support that marriage is an outdated institution.*	

Table 2: Exemplary dialogue between BEA and the user about the topic *"Marriage is an outdated institution"*



So as you are indifferent concerning the presented argument, I would like to go back to the topic of this discussion that marriage is an outdated institution.

number_visited

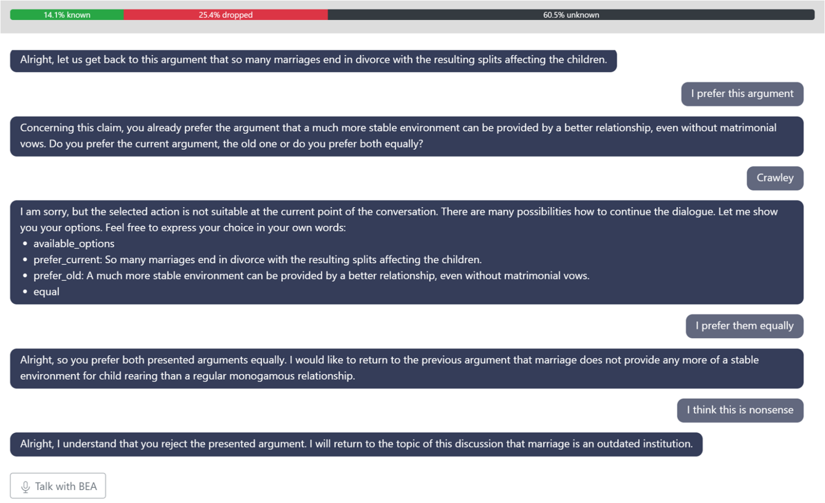
So far, I have presented you 2.8% of the available arguments while having dropped 9.9%. Thus, 87.3% still remain unknown.

why_pro: Marriage is an outdated institution.

A supporting argument I found is that there needs to be a new more inclusive institution that is open to all religions and those of no religion. It is clear that marriage can no longer perform this function for everyone in society.

Choose your action...
stance_overall
stance
number_visited
level_up
why_pro: Marriage is a religious institution in a society of declining religion.
why_con: Marriage is a religious institution in a society of declining religion.
prefer: Marriage is a religious institution in a society of declining religion.
reject: Marriage is a religious institution in a society of declining religion.
indifferent: Marriage is a religious institution in a society of declining religion.
exit
Choose your action...
Send

Figure 3: GUI of the baseline system with unfolded drop-down menu.



Alright, let us get back to this argument that so many marriages end in divorce with the resulting splits affecting the children.

I prefer this argument

Concerning this claim, you already prefer the argument that a much more stable environment can be provided by a better relationship, even without matrimonial vows. Do you prefer the current argument, the old one or do you prefer both equally?

Crawley

I am sorry, but the selected action is not suitable at the current point of the conversation. There are many possibilities how to continue the dialogue. Let me show you your options. Feel free to express your choice in your own words:

- available_options
- prefer_current: So many marriages end in divorce with the resulting splits affecting the children.
- prefer_old: A much more stable environment can be provided by a better relationship, even without matrimonial vows.
- equal

I prefer them equally

Alright, so you prefer both presented arguments equally. I would like to return to the previous argument that marriage does not provide any more of a stable environment for child rearing than a regular monogamous relationship.

I think this is nonsense

Alright, I understand that you reject the presented argument. I will return to the topic of this discussion that marriage is an outdated institution.

Figure 4: GUI of the speech-based system.

(especially non-native English speakers) is required. We trained the model for the moves in Table 1 using a set of circa 15 self-generated utterances for each move type. The spoken user input is captured with a browser-based audio recording that is further processed by the `Python` library `SpeechRecognition` using Google Speech Recognition. After a user intent is recognized and a corresponding response chosen, the spoken system response is presented using the Speech Synthesis of Web Speech API.

4. Experimental Setting

To evaluate the speech interaction, we designed an online user study where the participants had two conversations, one with the baseline system and one with the speech based one. In order to prevent sequence effects, a random half of the participants started with the baseline system and finished with the speech based one and the other half in reversed order. Before rating the systems the participants were asked to provide information on their age and their gender. After each conversation the participants had to rate 43 statements regarding the interaction. The first 33 were taken from the SASSI questionnaire (Hone and Graham, 2000)[3], which are divided in 6 categories: *Accuracy, Likeability, Cognitive Demand, Annoyance, Habitability* and *Speed*. The last 10 statements were added to assess the *Engagement* the user felt by the system and the *Presentation* of the arguments. Finally, the participants received a short questionnaire concerning which system they preferred and why. As all participants were Germans, they were asked to state if their English was sufficient to understand the system and the answers of the questionnaire. Before each interaction, the participants received a detailed, personal introduction what to do and explanation how to interact with the system. Each participant was asked to confirm if they understood what to do and to report any problems.

The participants could decide on their own how long they wanted to interact with each system. The only restriction was to use every move type at least once.

5. Evaluation Results

The study was conducted online under personal supervision with 20 participants consisting of 8 females and 12 males. The age range was between 22 and 68 with an average age of 32.73 (median 29). All participants were non-experts without a topic-specific background. The average interaction time was 33.35 minutes for the speech system and 25.67 minutes for the menu systems. Since the average number of heard arguments was higher (speech:18.91, menu:14.48) in the speech than in the menu system, this indicates that the former seems to be more engaging for the users. As described in Section 4, the participants had to rate 43 statements about the dialogues on a five-point Likert scale (1 = totally disagree, 5 = totally agree) for each system. Before taking the average over each category we inverted the rating scale for the negatively formulated statements such that the optimal rating is 5.

Category	M_{menu}	M_{speech}	p **value**
Accuracy	**4,07**	3,23	< 0.001*
Likeability	3,31	**3,82**	0.001*
Cog. Demand	**3,41**	3,35	0.797
Annoyance	2,70	**3,42**	< 0.001*
Habitability	**3,68**	2,83	< 0.001*
Speed	**3.15**	3.10	0.79
Engagement	2,61	**3,66**	< 0.001*
Presentation	3,61	**3,77**	0.147
q35 Article	2,30	**3,65**	0.005*
q36 Naturalness	1,80	**3,55**	0.001*

Table 3: Results of the participants' ratings. M_{menu} denotes the mean of the menu system and M_{speech} of the speech-based system. The differences that are statistically significant ($\alpha < 0.05$) are marked with *. The results for single statements are shown below the line.

The results are shown in Table 3. Below the category means, the results for single statements of special interest are shown. To determine whether the difference between the systems is significant, we used the non-parametric Wilcoxon signed rank test (Woolson, 2007) for paired samples. The *Cognitive Demand*[4] and the *Argument Presentation*[5] were rated without significant differences, which is plausible as the systems do not differ in these aspects. Interestingly, also the *Speed* of both systems was rated to be almost optimal, as none of them was considered “too slow” or “too fast”. Thus, the speech-based system is applicable in a real-time scenario.

Regarding the *Likeability* which describes the user's opinion and feelings towards the system (e.g. “I enjoyed using the system”, “I would use this system”,...), the speech-based system was significantly preferred. In consistence we observed a significantly better performance of the speech-based system with respect to the category *Annoyance*. It describes to which extent the users perceive the system as being “repetitive/ boring/ irritating”. Furthermore, regarding the *Engagement* which includes the variety, intuitivity and motivation conveyed by the system, the baseline system is outperformed. Regarding these three categories, we observe a considerable preference towards the speech-based system. This indicates that speech interaction leads to a more comfortable, enjoyable and human-like interaction. The highly significant single question results support these findings, especially with regard to the naturalness and preference of exploring the arguments using the speech-based system over reading about the topic in an article.

---

[3] This questionnaire has been developed to measure the subjective assessment of speech-based system interfaces.

[4] The *Cognitive Demand* summarises both, the perceived level of effort needed to use the system and user feelings arising from this effort.

[5] The *Argument Presentation* describes the way the arguments are presented and their content.

Moreover, while talking to the speech-based system the participants heard about 17% more arguments than compared to the baseline system. This indicates that using a speech interface to discuss a topic is significantly more engaging to human users and encourages their willingness to explore diverging views. This is confirmed by the recorded answers of the participants. Instead of simple commands, the participants engaged in a discussion with the system and tried to reason and justify their opinion ("I do not believe this is true. There are still many marriages which show commitment and respect.", "I fully agree as I also believe that divorce is undermining the idea of marriage.", etc.). 80% (16 out of 20) of the participants explained their preference/rejection of the presented arguments ("In my opinion this is not a valid point because...", "I agree with this as ..." etc.). Thus in the speech-based scenario, the participants are motivated to discuss and reflect on presented arguments, requesting for more details and information ("Let us concentrate on the implications of law. What further supporting arguments do you have on that?" etc.) and arguing on their stance. This is also underpinned by the free text explanation given by the participants who preferred the speech-based system.

Still, the advantages of the menu-based system are in contrast described as being "controllable", "predictable" and "easier to use", which can be observed in the following categories. The *Accuracy* describes, whether the system recognizes the user input correctly and hence, follows their intent. Additionally, the *Habitability* is defined by whether the user knows what to say and understands the system's reaction. The ratings in both categories are significantly higher for the menu-based system. This can be explained by the errors occurring in the speech recognition in the speech-based system. As interaction protocols show, although the NLU was able to identify most of the user intents correctly, some were mistakenly matched or could not be identified. Obviously, these errors do not occur when using the menu-based system explaining the better performance.

In both systems participants missed a more flexible navigation, especially the possibility to go back to the major claim directly or to jump to previously mentioned arguments independent of a uttered preference. Even though we were aware of this disadvantage, we chose not to incorporate such options beforehand, as the clarity of selectable options could not be guaranteed in the menu system. As the graphical interface only offers limited space it would not have been possible to display all navigation options, especially after some more arguments were heard. Even though this problem did not concern the speech system we did not incorporate this in neither as we aimed for an unbiased comparison. Interestingly the suggestion of a more flexible navigation and proactive system behaviour ("The system could have stated new arguments on its own", "The system should suggest new arguments, it would be more diversified.") was stressed much more in the speech system. This might explained by the more natural appearance of the speech system which rose higher expectations with regard to its flexibility in contrast to the quite static appearance of the menu system.

6. Conclusion and Future Work

In this work, we have introduced the first speech-based argumentative dialogue system which enables its user to explore arguments in real-time while estimating their preferences. We provide a platform that fosters critical thinking and open-mindedness which can e.g. be used in the area of education or (political) discourse/debates. A user study showed that speech interaction is perceived significantly more natural, intuitive and engaging than a menu-based one. While talking to the speech-based system significantly more arguments have been heard by the participants. Regarding both systems the participants criticized the limited ability to navigate through the arguments and suggested a more flexible navigation to other arguments.

Comparable with a conversation between two human users, the participants did not only state their preferences but justified and explained them to the system. Furthermore, different levels of detail in information was asked individually, which shows that an individual adaption in depth and breadth of the argumentation is needed. In contrast to the menu baseline, much more flexibility was expected from the speech system by e.g. offering arguments proactively and not just on request, likewise to a "normal discussion with a human". Thus, a speech-based interaction is a step towards our aim to provide a system that fosters critical and reflective thinking and open-mindedness (for educational purposes, debate training etc.). The clear preference to use the speech-based system over reading the information in an article indicates that we pursue a promising approach to explore arguments.

In future work, we want to increase the robustness and reliability of the speech interaction by including an accurate and stable automated commercial speech recognition. Furthermore, the flexibility of the NLU (as it does not need to be trained on new data corpora) shall be exploited to connect our system to new databases in real-time. This will also enable us to explore other argument structures, e.g. flatter ones to facilitate the navigation within the graph. Moreover, we aim to increase the naturalness of and motivation to interact with our system. Therefore, we will extend the system's flexibility with regard to the navigation within the argument tree and to react directly to the user utterance e.g. by processing possible new arguments the user mentions and updating the system's database during the interaction. Furthermore, we want to estimate user preferences using implicit and explicit feedback. Using this information the system will be able to proactively suggest new arguments and e.g. engage the user in hearing additional arguments. Finally, we will evaluate the above mentioned extensions of the speech-based system in a broad crowd-sourcing study and investigate if we can improve the user satisfaction and motivation to engage in an argumentative discussion.

7. Acknowledgements

This work has been funded by the DFG within the project "How to Win Arguments – Empowering Virtual Agents to Improve their Persuasiveness", Grant no. 376696351, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).

8. Bibliographical References

Abro, W. A., Aicher, A., Rach, N., Ultes, S., Minker, W., and Qi, G. (2021). Natural language understanding for argumentative dialogue systems in the opinion building domain. *arXiv*, arXiv:2103.02691.

Aicher, A., Rach, N., Minker, W., and Ultes, S. (2021). Opinion building based on the argumentative dialogue system bea. *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pages 307–318.

Amgoud, L. and Ben-Naim, J. (2016). Evaluation of arguments from support relations: Axioms and semantics. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI-16*, pages 900—-906.

Amgoud, L. and Ben-Naim, J. (2018a). Weighted bipolar argumentation graphs: Axioms and semantics. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5194–5198.

Amgoud, L. and Ben-Naim, J. (2018b). Weighted bipolar argumentation graphs: Axioms and semantics. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5194–5198. International Joint Conferences on Artificial Intelligence Organization, July.

Bechhofer, S. (2009). Owl: Web ontology language. In *Encyclopedia of Database Systems*, pages 2008–2009. Springer.

Chalaguine, L. A. and Hunter, A. (2020). A persuasive chatbot using a crowd-sourced argument graph and concerns. In *COMMA*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hadoux, E. and Hunter, Anthony, e. a. (2021). Strategic argumentation dialogues for persuasion: Framework and experiments based on modelling the beliefs and concerns of the persuadee. In *arXiv*, volume 2101.11870.

Hone, K. S. and Graham, R. (2000). Towards a tool for the subjective assessment of speech system interfaces (sassi). *Nat. Lang. Eng.*, 6(3–4):287–303, September.

Le, D. T., Nguyen, C.-T., and Nguyen, K. A. (2018). Dave the debater: a retrieval-based and generative argumentative dialogue agent. *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130.

Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Rach, N., Langhammer, S., Minker, W., and Ultes, S. (2018). Utilizing argument mining techniques for argumentative dialogue systems. In *Proceedings of the 9th International Workshop On Spoken Dialogue Systems (IWSDS)*, May.

Rakshit, G., Bowden, K. K., Reed, L., Misra, A., and Walker, M. A. (2017). Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents - 8th International Workshop on Spoken Dialog Systems*, pages 45–52.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rosenfeld, A. and Kraus, S. (2016). Strategical argumentative agent for human persuasion. In *ECAI'16*, pages 320–328.

Saha, T., Saha, S., and Bhattacharyya, P. (2020). Towards sentiment-aware multi-modal dialogue policy learning. *Cognitive Computation*, pages 1–15, 11.

Slonim, N., Bilu, Y., Alzate, C., Bar-Haim, R., Bogin, B., Bonin, F., Choshen, L., Cohen-Karlik, E., Dankin, L., Edelstein, L., et al. (2021). An autonomous debating system. *Nature*, 591(7850):379–384.

Stab, C. and Gurevych, I. (2014). Annotating argument components and relations in persuasive essays. In *COLING*, pages 1501–1510.

Wilcock, G. and Jokinen, K. (2021). Towards increasing naturalness and flexibility in human-robot dialogue systems. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pages 109–114. Springer Singapore.

Woolson, R. (2007). Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3.