Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 513–522
Marseille, 11–16 May 2020

Evaluation of Argument Search Approaches in the Context of Argumentative Dialogue Systems

Niklas Rach[1,2], Yuki Matsuda[2], Johannes Daxenberger[3], Stefan Ultes[4], Keiichi Yasumoto[2], Wolfgang Minker[1]
[1]Institute of Communications Engineering, Ulm University, Ulm, Germany
[2]Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, Japan,
[3]Ubiquitous Knowledge Processing Lab, TU Darmstadt, Darmstadt, Germany
[4]Mercedes-Benz Research & Development, Sindelfingen, Germany
niklas.rach@uni-ulm.de

Abstract

We present an approach to evaluate argument search techniques in view of their use in argumentative dialogue systems by assessing quality aspects of the retrieved arguments. To this end, we introduce a dialogue system that presents arguments by means of a virtual avatar and synthetic speech to users and allows them to rate the presented content in four different categories (*Interesting*, *Convincing*, *Comprehensible*, *Related*). The approach is applied in a user study in order to compare two state of the art argument search engines to each other and with a system based on web search. The results show a significant advantage of the two search engines over the baseline. Moreover, the two search engines show significant advantages over each other in different categories, thereby reflecting strengths and weaknesses of the different underlying techniques.

Keywords: Argument Search, Argument Quality, Dialogue System Evaluation

1. Introduction

Computational argumentation and related applications have gained a lot of interest recently, with the most prominent example being the IBM Project Debater[1] engaging in live debate with a human. Argumentative dialogue systems and chat bots are applications concerned with tasks that require the exchange of arguments as, for example, persuasion (Chalaguine et al., 2019; Rosenfeld and Kraus, 2016), discussing controversial topics (Rakshit et al., 2019) or responding to customer reports (Galitsky, 2019). In order to address these tasks, the systems depend on knowledge about existing arguments regarding the discussed topic. Most systems so far operate on a carefully but also strictly designed database of arguments that perfectly matches their requirements. However, in order to increase their flexibility in view of the range of discussed topics, automatized and topic independent approaches to acquire arguments are required. Argument search engines (Ajjour et al., 2019) on the other hand have recently emerged from the field of argument mining (Lippi and Torroni, 2016) and provide users with a ranked list of arguments corresponding to a given search query. Hence, they are of particular interest for argumentative dialogue systems as they allow to search a wide variety of sources for arguments and are not restricted to specific topics. However, in order to utilize argument search techniques for argumentative dialogue systems, certain quality standards for the retrieved arguments are crucial.

Within this work we investigate these quality aspects by means of an argumentative dialogue system that evaluates arguments retrieved by different search approaches directly in the interaction with users. This is realized by allowing the user to give specific ratings in the categories *Interesting*, *Convincing*, *Comprehensible* and *Related* as direct feedback to each system utterance. In order to ensure a setting that is representative for dialogue system applications, the arguments are presented by means of a virtual avatar and synthetic speech. The approach is motivated by the difficulty of argument quality assessment from a purely logical perspective (Habernal and Gurevych, 2016b; Wachsmuth et al., 2017a) as well as the common approach to evaluate dialogue systems from the user perspective (Deriu et al., 2019). Especially the subjective nature of our addressed application scenarios (and argumentation itself) and the effects of system modalities (virtual avatar and synthetic speech) in the present scenario render approaches that do not explicitly consider the user perception impractical.

We apply our system in a user study in order to compare two state of the art argument search engines, namely ArgumenText (Stab et al., 2018a) and args.me (Wachsmuth et al., 2017c), to each other. In addition, we introduce an argument retrieval system based on conventional web search to provide a suitable baseline. In order to exclude topic dependencies, the comparison is done over three different controversial topics. The results show significant differences between the investigated approaches for three of the four categories and both search engines outperform the baseline in one category. In addition, both search engines outperform each other in a different category, thereby reflecting the different strengths and drawbacks of the underlying technological approaches. The main contributions of this paper can hence be summarized as follows:

- Providing a general evaluation setup for the comparison of argument retrieval systems in view of their suitability for dialogue system applications.
- A comparison of two state of the art argument search approaches and a baseline approach in an extensive user study by means of the proposed evaluation setup.

[1]https://www.research.ibm.com/artificial-intelligence/project-debater/

The remainder of this paper is as follows: We discuss related work from the field of dialogue system evaluation and argument quality in Section 2. and introduce the utilized evaluation criteria in Section 3. Section 4. includes a discussion of existing argument search engines with an emphasis on the approaches utilized in this work. Section 5. provides an overview over the architecture of the proposed evaluation system, whereas Section 6. covers the experiment and results. The discussion of the results is included in Section 7., followed by a conclusion and an outlook on future work in Section 8.

2. Related Work

In this section, we discuss related work from the fields of dialogue system evaluation and argument quality assessment. For dialogue systems, different general evaluation approaches exist based on the different types of system (Deriu et al., 2019). For task oriented systems, which are the most relevant in view of argumentation, the task success rate, i.e. the rate at which a system successfully carries out the assigned function is a common evaluation criterion (Schatzmann et al., 2007; Laroche et al., 2011). It was combined with a measure of the dialogue cost and the subjective satisfaction of the user with the interaction in the PARADISE framework (Walker et al., 1997; Walker et al., 2000) in order to enable comparisons of different systems. In addition, Ultes et al. (2013) introduced the Interaction Quality as an expert rating based approach to model user satisfaction.

Since argumentation is a comparatively new domain for dialogue systems, assessment of argumentative systems is currently done in a very system specific way: The Project Debater introduced by IBM was evaluated by engaging with a human in a live debate. The outcome was determined by a comparison of the audience's stance before and after the debate, showing an advantage of the human debater over Project Debater. Rosenfeld and Kraus (2016) evaluated the persuasive effect of their introduced persuasive agent in a user study with an argument structure specifically collected for this task. In addition, the system introduced by Yuan et al. (2008) was evaluated in an expert evaluation (Yuan, 2004) and a user study (Ævarsson, 2006). The effect of different types of arguments presented by an argumentative chat bot were investigated in the behaviour change domain and also by means of a user study by Chalaguine et al. (2019), showing that arguments that address the concerns of the user were preferred over others.

The assessment from a technical perspective was considered by Rakshit et al. (2019), where a comparison of response times for the different underlying techniques was utilized as evaluation criterion. Moreover, a retrieval-based approach and a generative approach to generate the system response were discussed (Le et al., 2018) and separately evaluated on established metrics for the underlying technological task.

Finally, Sakai et al. (2018) and Rach et al. (2019) evaluated argument structures acquired specifically for the use in dialogue systems also by means of user studies. However, the evaluation in these cases is focused on specific systems and/or data and the effect of different acquisition techniques as desired in the present work was therefore not included. In order to provide a detailed evaluation that is not tailored to one specific task, we combine the generalized approach of task success rate as evaluation criterion with aspects related to the field of argument quality assessment.

As for computational argumentation, Wachsmuth et al. (2017b) presented a unified taxonomy for the theoretical assessment of argument quality in different sources and for different argument granularities. They divide argument quality in the broad categories of *logical*, *rhetorical* and *dialectical* quality and introduce 15 fine-grained sub-dimensions as well as a corpus annotated with these dimensions. Habernal and Gurevych (2016a) introduced an approach to assess the convincingness of arguments in which arguments were rated in direct comparison to each other in a crowd-sourcing experiment. The correlations between the theoretical and the crowd-sourcing based approach were also investigated (Wachsmuth et al., 2017a) and a corpus for the comparison of the convincingness of evidences was introduced (Gleize et al., 2019). The overall quality of single arguments as well as argument pairs was discussed by Toledo et al. (2019) together with automatized approaches for argument ranking and argument-pair classification. Finally, Potthast et al. (2019) utilized expert ratings of the above mentioned categories *logical*, *rhetorical* and *dialectical* quality to assess different retrieval approaches for argument search in combination with the information retrieval notion of relevance. To the best of our knowledge, no studies have been carried out which explicitly focus on argument quality assessment in the context of dialogue systems.

3. Evaluation Criteria

As a first step in developing the actual evaluation setup, we introduce the criteria utilized throughout the evaluation process to assess the presented arguments. Their definition is driven by the goal of providing a set of evaluation dimensions suitable for an assessment during the interaction as well as general enough to provide valuable insights for the application of the compared approaches in argumentative dialogue systems. Therefore, we introduce four categories that cover the following quality aspects:

a) The structural properties of the arguments that are influenced by the different technological approaches (i.e. identification of arguments and stance) of the search engines.

b) The suitability of the retrieved arguments for the different tasks of an argumentative system.

In order to address aspect a), we rely on dimensions of *argument quality* that are strongly influenced by the technological differences between the search engines. For aspect b), we start from the general notion of *task success* as a common approach to assess task oriented dialogue systems. Since the success of an argumentative dialogue system depends on the individual user and is therefore hard to measure objectively, we identify properties of the retrieved arguments that facilitate the completion of possible tasks

and assess them in separate categories. The following subsections provide a detailed discussion of both aspects and include the notion of the categories utilized within our system. Throughout this work, we denote a single search result from an argument search engine as *argument* and its polarity in view of the overall topic as *stance* (support/PRO or attack/CON).

3.1. Argument Quality Related Criteria

Following the work of Wachsmuth et al. (2017b), overall argument quality includes the three dimensions of *logical*, *rhetorical*, and *dialectical* quality. *Logical* quality is related to the structure of an argument, i.e. the question if an argument is logically sound. *Rhetorical* quality is reflected in the persuasive effect of an argument and hence also depends on aspects like the presentation of the argument and the credibility of the person presenting it. Lastly, *dialectical* quality is related to the contribution of an argument to the overall resolution of an issue or different opinions on a topic.

The task of an argument retrieval system includes the mining of arguments from relevant sources as well as the recognition of the respective stance (Ajjour et al., 2019). Errors result in a system output that is not perceived as argumentative, not related to the topic or presented with the wrong relation to the topic, which is all part of the logical quality dimension. In order to distinguish between the different errors, we assess the logical quality of the retrieved arguments with the two evaluation categories *Comprehensible* (Does the argument make sense by itself?) and *Related* (Is the presented argument related to the topic and is the presented relation correct?). The first one is binary, whereas the second one allows a choice among one positive and two negative options (related, not related, wrong relation) in order to enable a distinction between the different errors that can occur.

3.2. Task Related Criteria

The task success rate is a common evaluation criterion for task oriented dialogue systems (Deriu et al., 2019) and measures to which extend and how often a system provides correct information/responses to user requests. However, in contrast to conventional setups (like providing information in the restaurant domain (Schatzmann et al., 2007)), the system output of an argumentative dialogue system cannot clearly be divided into right and wrong responses. Nevertheless, the capability of a dialogue system to solve an argumentative task depends on whether it is able to select suitable arguments and hence on the output of the respective retrieval system. Based on the types of argumentative dialogue (Reed and Norman, 2003), the different tasks of argumentative dialogue systems can be broadly divided into competitive and cooperative tasks.

In competitive tasks like persuasion or negotiation, the overall goal is to *convince* the opposite site of for example a certain point of view (persuasion) or to accept a specific offer (negotiation). Consequently, the relevant property of the utilized arguments is their overall likeliness to convince the opponent, in short their *convincingness*. The respective evaluation category in our setup is hence *Convincing* (Does the argument convince me?) for competitive setups.

Cooperative setups (for example deliberation) on the other hand aim for a mutual solution of an issue by exchanging arguments that contribute to this task. In contrast to competitive setups, the goal of the involved parties is not to convince the other participants of a certain point of view but to find the best common ground. Therefore, the suitability of an argument for these tasks depends on its ability to contribute to this solution. However, in an argumentative application, this common ground should satisfy the user's needs and hence depends the user perspective on the presented arguments. Consequently, we condensate this property in the question if an argument is *interesting* for the user in view of the discussed topic and assess it by means of a category with the same name.

It should be noted that both of these categories are also related to argument quality: the *convincingness* of an argument is an aspect of rhetorical quality whereas *Interesting* reflects the user's personal view on the overall relevance of an argument and is hence related to dialectical quality. Nevertheless, each category only covers a part of the respective quality dimension, as both dimensions can be further influenced by other modules of the argumentative dialogue system (behavior of the avatar, natural language generation). As the focus of this work is on the evaluation of different retrieval systems, we do not explicitly evaluate different approaches for these modules and consequently focus on the above discussed aspects instead of the complete quality dimension.

4. Argument Search Approaches

In the next step, we discuss the argument search approaches that are compared throughout this work. In order to be included into the evaluation, the respective search engine has to meet the following two requirements: It has to be accessible by an API and provide information about the stance of the retrieved arguments. The first requirement is necessary for the technical applicability of the search engine within argumentative applications whereas the second requirement is motivated by the need for stance information in the majority of the desired tasks of an argumentative system.

The aim of an argument search engine is to retrieve a ranked list of arguments related to a given search query. Different systems introduced so far follow different paradigms in order to accomplish this goal (Ajjour et al., 2019) and include the IBM Project Debater (Levy et al., 2018), TARGER (Chernodub et al., 2019), PerspectroScope (Chen et al., 2019), args.me (Wachsmuth et al., 2017c) and ArgumenText (Stab et al., 2018a). Out of this list, only ArgumenText, args.me and TARGER provide an API to access retrieved arguments and only the first two also include information about their stance. Consequently, we focus our evaluation on these two and discuss the underlying approaches in detail in the following subsections. In addition, we propose a novel system utilizing a conventional web search approach in order to generate baseline results for the evaluation.

Information	System Return
Premises	*If marriage's main function is to protect against bereavement and divorce then it is essentially protecting against harms that it itself brings. Without marriage, bereavement and divorce would cease to be as serious harms as they currently are.*
Stance	con
Conclusion	Marriage represents a legal bond which protects both parties in a relationship.

Table 1: Exemplary search result of args.me for the search query *marriage*, including all information utilized throughout this work (premises, stance, conclusion).

4.1. args.me

The args.me search engine (Wachsmuth et al., 2017c) allows users to search arguments related to a search query from a corpus with over 300k arguments retrieved from different debating websites. The indexing of arguments is done offline and independently of the search query. In contrast to the other approaches discussed herein, the underlying algorithm exploits the specific debate setup of the source pages in order to identify argument stance and boundaries. Consequently, arguments are defined as a set of premises related to a conclusion, as shown in the example in Table 1[2]. Although the arguments therefore include more contextual information than sentential arguments, it is difficult to use this output directly in dialogue systems. Especially in scenarios that do not adhere to a clear structure in view of speaking time and turn taking (like debates), extensive utterances are hard to follow and understand if presented only by synthetic speech (Wilcock and Jokinen, 2019).

In order to determine a reasonable maximum number of words for an argument in our setup, we investigate the length of manually annotated arguments from an online debate. We use the argument structure from Rach et al. (2019) since a) it is based on data from one of the source pages of args.me (idebate.org) and b) was annotated for the use in dialogue systems. We find that over 97% of the manually annotated arguments consist of less than 60 words, and we therefore set this value as the maximum number of words included in the system output. Based on this threshold, we further re-rank the arguments retrieved by args.me in order to prefer arguments with an overall length of the premises smaller than 60, given that the complete search query is present in the conclusion of the argument. In addition, the premises of longer arguments are truncated to include only the first m sentences, whereas m is the maximum number of sentences that lead to an overall number of words ≤ 60.

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

Information	System Return
Argument	*Countries that have implemented strict gun control laws have been able to reduce the incidence of gun death.*
Stance	pro
Confidence	$c_a = 0.999, c_s = 0.997$

Table 2: Exemplary search result of ArgumenText for the search query *gun control*, including all information utilized throughout this work (argument, stance, confidence scores).

4.2. ArgumenText

The ArgumenText search engine (Stab et al., 2018a) extracts sentential arguments (Stab et al., 2018b) on arbitrary topics. To that end, it retrieves relevant documents from a large web crawl (CommonCrawl[3]), and detects for each sentence extracted from the most relevant documents whether it constitutes a supporting or opposing argument with regard to the query (i.e. topic under consideration), or no argument at all. The underlying algorithm uses an attention-based neural network as described in Stab et al. (2018b), trained on annotated sentences from web documents for more than 40 topics. ArgumenText has been shown to yield a coverage of almost 90% compared to expert-curated collections of arguments on given controversial topics (Stab et al., 2018a). While the system's drawback is its lower precision compared to expert annotations (argument and stance detection), it can detect arguments on virtually any topic of interest and it retrieves content from many different sources. For a given search query, the API returns a list of arguments and their corresponding stances. In addition, confidence scores of both argument (c_a) and stance (c_s) detection are provided for each argument. Throughout this work, we compute the overall confidence as $c = c_a \times c_s$ and rank the retrieved arguments according to this score in order to equally take into account both aspects of the search. Table 2 shows an exemplary search result including the utilized information [4].

4.3. Baseline

In order to compare the discussed argument search approaches to a suitable baseline, we introduce an architecture that utilizes the results of a conventional web search in order to find web pages that contain arguments related to the search query. To ensure reproducible results, we employ the chatnoir search engine (Bevendorff et al., 2018) on CommonCrawl data with the search query *arguments <TOPIC>*. The text blocks of the websites with the highest ranking for each topic are then searched for sentences that contain topic specific key words. The list of key words consists of sub strings of the topic (for example *nuclear* and *energy*) as well as a WordNet (Bird et al., 2009; Miller et al., 1990) synonym either of the complete search query or (if none was found) the sub strings. The final arguments are then sampled from this list with a fixed random seed, also

[3]http://commoncrawl.org/

[4]https://guncontrolfacts.org/category/gun-control-pros-and-cons/

to ensure reproducibility.

In order to determine the stance of the retrieved argument, we train a classifier on the IBM stance classification data (Bar-Haim et al., 2017). The corpus consists of 2394 claims annotated with an overall sentiment label $s_c \in \{1, -1\}$, the stance towards the related topic and a sentiment label for the topic $s_t \in \{1, -1\}$. Similar to the baseline approaches in the original work, we train a model to estimate the sentiment of each claim and assume that the target towards which the estimated sentiment is expressed is consistent with the target of the topic. The corresponding stance can then be derived as $s_c \times s_t$.

Our approach utilizes a pre-trained BERT (Devlin et al., 2019) model[5] to get sentence embeddings for all claims in the corpus which are then used as feature vectors for a support vector machine (SVM) classification. The parameters of the SVM are optimized in a systematic grid search in order to match the specific task. The performance of our model is evaluated by averaging the results for five different random train/test splittings of the data with the same characteristics provided in the original work (training: 25 topics and 1039 claims, test: 30 topics and 1355 claims). Since the overall system requires an estimate of the stance for all retrieved arguments, we do not consider different coverage rates (as in the original work) and only investigate predictions on the complete test set, independently of the confidence of the classifier. For the sentiment classification, we report an average accuracy of 0.80 and an F1 score of 0.77. The final stance classification results in an average accuracy of 0.68 (F1 = 0.70), which clearly outperforms the baseline in the original work and is slightly higher than the values provided for other therein discussed methods.

5. System

The final component of the complete evaluation setup is a dialogue system that allows users to apply the evaluation criteria discussed throughout Section 3. in an intuitive way and during the ongoing interaction. The system utilized within this work was designed specifically for this task and allows the user to select his or her ratings as a direct response to the system utterance. The rating for each category can be given once for each argument and cannot be changed. In addition, the user is able to start the conversation, request the next argument, go to the previous one and repeat the latest utterance that includes an argument. If requested, the system selects the next argument randomly from the pool of available ones but each argument can occur only once during the interaction. It is important to note that the system requirements in view of the utilized arguments are as liberal as possible in order to enable the comparison of multiple different search approaches. The only information that has to be provided is the content of the argument as well as a notion of the respective stance towards the main topic. The overall interaction is stopped by the system after a fixed time in order to ensure the same conditions for each user.

Figure 1: Screenshot of the system, including avatar and navigation buttons.

5.1. Interface

The interface is adapted from the systems introduced by Rach et al. (2018) and Weber et al. (2020). It is based on the Charamel™ avatar[6] which presents the system utterance via synthetic speech by utilizing Nuance TTS and Amazon Polly Voices[7]. Besides the avatar, the interface also includes buttons for the ratings in each category as well as the remaining user options (repeat, next, previous, start). If an option is not available in the current state of the interaction, this is indicated by the appearance of the respective buttons. A screenshot of the interface including buttons and avatar is shown in Figure 1.

5.2. Natural Language Generation

The system utterance is generated by a modified version of the template based Natural Language Generation (NLG) used in Rach et al. (2018) that receives the list of retrieved arguments from a certain argument search approach as input. Each argument is presented with an explicit notification of the assigned stance regarding the discussed topic. The specific formulation is selected randomly from a list of available options. In addition, the system starts each interaction by giving a topic specific introduction including a short repetition of the task and the general claim *<TOPIC> is/are good*. Moreover, statements to notify the user if no further arguments are left and if the time for the discussed topic is over are also included. Table 3 shows the transcript of an interaction with the system, including the different utterance types of the dialogue system, the corresponding system utterance and exemplary user ratings. The arguments in this example were created manually for demonstration.

6. Experiment and Results

We applied the complete evaluation setup in a user study with 19 participants at the Nara Institute of Science and

[5]Within this work, we utilize the *base* model, available at https://github.com/hanxiao/bert-as-service

[6]https://www.charamel.com/competence/avatare

[7]https://docs.aws.amazon.com/polly/latest/dg/voicelist.html

	System Utterance	Ratings
Intro	[...] The topic I want to discuss first is: *Veganism is good.*	-
Arg_1	In contradiction to my initiating claim I found the following opinion: *Eating animals is in our nature, therefore it cannot be wrong.*	interesting not convincing comprehensible related
Arg_2	Okay, let's continue with a new aspect. The next argument is a supporting one. The content is: *There is a nice vegan restaurant, just around the corner.*	not interesting not convincing comprehensible not related
Arg_3	All right, let me see what else I found. Next in line is a support argument with the following content: *A key motivation for veganism is to prevent the exploitation of animals.*	interesting convincing comprehensible related
End	Thank you for your ratings. Unfortunately, we reached the time limit for this topic. [...]	-

Table 3: Dialogue transcript for the topic *Veganism*, including excerpts of the introduction and the closing statement, three arguments and the corresponding exemplary user ratings.

Technology (NAIST). In order to exclude topic dependencies, we included the three common controversial topics *nuclear energy*, *self-driving cars* and *animal testing* into the evaluation and retrieved the best ranked 20 arguments for each topic using the three argument search approaches described in Section 4. In the course of the experiment, participants were able to interact with the system via the introduced interface and to rate the presented arguments in the discussed four categories *Interesting*, *Convincing*, *Comprehensible* and *Related*.

6.1. Setup

The experiment was divided into the three stages *introduction*, *interaction/rating* and *feedback*. During the introduction, each participant received written and oral instructions including an explanation of the interface, meaning of the categories and purpose of the experiment. In addition, a test trial with a separate small argument pool was offered in order to clarify the task. Since the experiment was conducted with non-native speakers, participants were not obliged to rate each argument in each category and instructed to skip a rating, if undecided. In addition, each participant rated the following statements/question on a five point Likert scale before starting the experiment:

- I'm in favour of <*TOPIC*>.
- How often do you use speech based devices/applications?

During the interaction phase, participants only interacted with the system and were not allowed to ask additional questions. Each participant listened to arguments for the topics *nuclear energy*, *self-driving cars* and *animal testing* retrieved with one of the three compared argument search approaches. In order to investigate the agreement between participants, a fourth topic (*death penalty*) was added. The pool of arguments for this topic was the same for each participant and included the top 8 arguments retrieved with both args.me and ArgumenText. For each of the four topics, the system stopped the interaction after a fixed time of five minutes.

After the interaction, participants were asked to anonymously provide feedback about their own English proficiency in view of the task, if they could understand the synthetic speech and the overall understandability of the system on a five point Likert scale. Moreover, the opportunity to give written and/or oral feedback was provided. In case the language barrier hindered a completion of the task, respective ratings were excluded from the evaluation (this case occurred only once).

6.2. Results

The experiment resulted in a total of 2407 ratings distributed over all four topics. We start the evaluation by assessing the participant responses regarding the understandability of the system provided after the interaction in order to rule out errors related to the presentation of the arguments. Responses were given on a five point Likert scale from strongly disagree (1) to strongly agree (5) for the statements

- The synthetic speech was easy to understand.
- All in all I had no problems understanding the system utterances.

Figure 2 shows the distribution of the participant responses. We see that only two participants disagreed with the statement that the synthetic speech was easy to understand and no participant explicitly reported problems with understanding the system utterances.

In the next step, we evaluate the agreement between the participants by means of Krippendorf's alpha (Hayes and Krippendorff, 2007) for the topic *death penalty* rated by all participants. This method is chosen as it allows to compare multiple raters, is able to handle missing data, and allows for a comparison with existing work. We analyse the ratings given for the same arguments in the same category, resulting in a maximum alpha of 0.15 for the agreement between all participants in the category *Comprehensible*. Since participants were instructed to rate based on their own opinion and not according to objective guidelines, these results are as expected and emphasise the highly subjective nature of the task. In addition, we investigated the agreement between participants with the same personal stance towards the discussed topic and found that no increased agreement (consistent for PRO or CON) can be observed.

Consequently, we proceed with a statistical evaluation of the complete ratings for each search approach and category (rather than an evaluation on the base of individual arguments) in order to derive conclusions from the results.

Category	args.me	ArgumenText	Baseline	p	args.me/ ArgumenText	ArgumenText/ Baseline	args.me/ Baseline
Interesting	0.81	**0.88**	0.72	≤ 0.01	0.10	$\leq$ **0.01**	0.10
Convincing	0.42	**0.59**	0.48	0.01	**0.01**	0.13	0.30
Comprehensible	**0.85**	0.83	0.78	0.32	-	-	-
Related	**0.89**	0.69	0.66	≤ 0.01	$\leq$ **0.01**	0.70	$\leq$ **0.01**

Table 4: Results of the statistical comparison of the ratings for the topics *nuclear energy*, *self-driving cars* and *animal testing*. Left part: Ratio of positive ratings and overall ratings for each category and architecture including the p-value derived with Fisher's exact test. Right part: Resulting p-values for all three pairings and categories that showed a significant difference derived by Fisher's exact test with Benjamini-Hochberg correction. Values for the category *Comprehensible* are not included since the prior testing showed no significant differences between the three compared approaches.

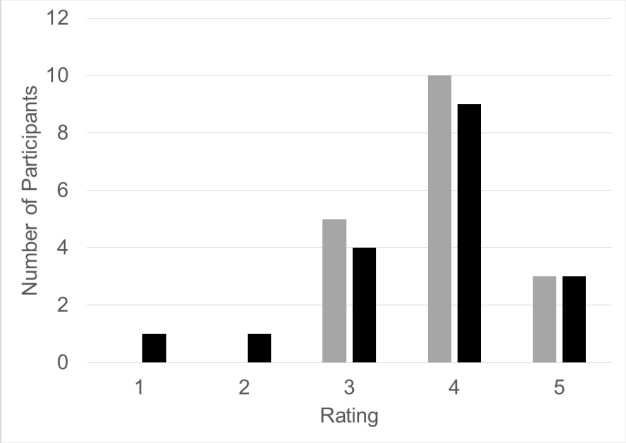

Figure 2: Responses on a five point Likert scale from completely disagree (1) to fully agree (5) for the statements *The synthetic speech was easy to understand* (black bars) and *All in all I had no problems understanding the system utterances* (grey bars).

For the statistical analysis, the ratings for all three approaches corresponding to the topics *nuclear energy*, *self-driving cars* and *animal testing* are compared for each dimension with Fisher's exact test (Sprent, 2011). The resulting p-values and the ratio of positive ratings and all ratings are shown in the left part of Table 4. We see that the categories *Interesting*, *Convincing* and *Related* yield a p-value smaller than $\alpha = 0.05$ whereas for the category *Comprehensible* no statistically significant difference between the investigated approaches is found. For a more detailed discussion, we compare the three approaches pairwise (also with Fisher's exact test), and utilize the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) in order to correct the p-values accordingly for multiple hypothesis testing. The results for all three pairings of the utilized systems are shown in the right part of Table 4.

The baseline shows no significant advantage over the other two approaches in any category. In contrast, the results for the *Related* category indicate a clear advantage of args.me over both baseline and ArgumenText. In addition, ArgumenText shows a statistically significant advantage over args.me in the category *Convincing* and over baseline in the category *Interesting*.

7. Discussion

In this section we discuss the results and implications of our findings.

7.1. Comparison of Search Approaches

The inability of the baseline architecture to outperform the two investigated argument search engines emphasises the need for argument search in general in order to retrieve suitable arguments as it can clearly not be substituted with a conventional web search.

In view of the two different argument search engines, we argue that the advantage of args.me in the category *Related* is due to its different approach to stance detection. args.me utilizes the specific structure of debating websites in order to determine the stance of an argument as well as its boundaries which ensures a very precise estimate of the argument relation to the discussed topic. In contrast, classifier based detection of stance and arguments as utilized in ArgumenText yield, at the current state of the art, a lower precision. On the other hand, classifier based approaches allow for a search in a broader variety of sources and hence for a richer pool of arguments, which is in our opinion reflected in the advantage of ArgumenText in the *Convincing* category. Moreover, the modified selection of arguments from the args.me search results (re-ranking and shorting) is likely to influence the perception of the users — but is (at the current time) unavoidable in order to use the respective arguments in a speech based dialogue system. We hence conclude that this is a current limitation of the system, which could be overcome by a more fine-grained system output that also includes a list of premises for each argument.

7.2. Subjectivity of the Task

In coding and annotation tasks, a low agreement between coders usually indicates misunderstandings in view of the task or guidelines. The low agreement between participants reported earlier hence raises the question if more restrictive guidelines for the participants are required. From our perspective, the goal of an argument-wise comparison of the participant ratings with expert annotations would justify this conclusion. However, we pursue the goal of assessing the user perception of the arguments based on his or her personal opinion, since users of argumentative applications are most likely not instructed on how to interpret

the system output. Consequently, the observed disagreement is in our opinion a result of the different views of participants on the discussed topics and argumentation as a whole, i.e. the subjective nature of the task. The results are also in line with existing studies on argument quality where only slightly higher alpha values were achieved between seven annotators (Wachsmuth et al., 2017b), and no effect of the annotators' stance could be measured (Potthast et al., 2019).

7.3. Implications for Dialogue Systems

The results of our study also allow some general conclusions for the development of future argumentative dialogue systems that aim to exploit argument search in order to retrieve arguments. Firstly, the reported subjectivity of the argument perception stresses the need for a careful selection of arguments based on the target audience or user. Consequently, adaptation and user modelling approaches investigated for the use in dialogue systems (Ultes et al., 2019; Mo et al., 2018; Casanueva et al., 2015) are also required in the domain of argumentation, although it is not clear from the results which user traits are the most relevant. In addition to this, the reported high user comprehension of the system's utterances (Figure 2) allows the conclusion that arguments retrieved by argument search engines can generally be understood. However, several participants reported that spelling or grammar errors in the arguments lead to an unnatural system output, which, although generally comprehensible, was not natural and intuitive. Consequently, more advanced approaches to NLG in combination with paraphrasing and grammar correction are also of interest in order to improve the user experience (Kwon et al., 2015; Wen et al., 2015).

Finally it should be noted that many dialogue systems require a more fine grained structure of arguments (Aicher et al., 2019; Sakai et al., 2018; Rosenfeld and Kraus, 2016; Rach et al., 2018) that include not just the general argument stance but also their explicit relations to each other. Hence, additional processing of the search results in order to structure the retrieved arguments as for example clustering of arguments (Reimers et al., 2019) may be required in order to allow systems of these kind to exploit argument search engines.

8. Conclusion

We introduced an evaluation setup for argument search approaches in the context of argumentative dialogue systems. Our approach assesses the users' opinions and perception regarding arguments presented by an avatar with synthetic speech. During the interaction with the system, users are able to rate the arguments presented by the avatar in the categories *Interesting*, *Convincing*, *Comprehensible* and *Related* as a direct response to the system utterance. The approach was applied in a user study in order to compare two argument search engines (ArgumenText and args.me) to each other and to a baseline web search architecture. Our results show a statistically significant advantage of both search engines over this baseline in one of the categories. Moreover, each search engine also shows a significant advantage over the other in a certain category which reflects the strengths and disadvantages of the underlying techniques. In addition to the results, we discussed implications for the general use of argument search engines in the context of dialogue systems.

Future work will focus on an evaluation of the dialogue system components that were not explicitly evaluated in the present study. This includes mainly the selection of arguments, i.e. the dialogue management, and the non-verbal behaviour of the avatar.

9. Acknowledgements

Parts of this work have been funded by the Deutsche Forschungsgemeinschaft (DFG) within the project "How to Win Arguments - Empowering Virtual Agents to Improve their Persuasiveness", Grant Number 376696351, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).

10. Bibliographical References

Ævarsson, K. (2006). *Human-computer debating system evaluation and development*. Ph.D. thesis, University of Akureyri, Iceland.

Aicher, A., Rach, N., Minker, W., and Ultes, S. (2019). Opinion building based on the argumentative dialogue system bea. In *Proceedings of the 10th International Workshop on Spoken Dialog Systems Technology (IWSDS)*.

Ajjour, Y., Wachsmuth, H., Kiesel, J., Potthast, M., Hagen, M., and Stein, B. (2019). Data acquisition for argument search: The args. me corpus. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 48–59. Springer.

Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A., and Slonim, N. (2017). Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261. Association for Computational Linguistics.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Bevendorff, J., Stein, B., Hagen, M., and Potthast, M. (2018). Elastic chatnoir: Search engine for the clueweb and the common crawl. In *European Conference on Information Retrieval*, pages 820–824. Springer.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Casanueva, I., Hain, T., Christensen, H., Marxer, R., and Green, P. (2015). Knowledge transfer between speakers for personalised dialogue management. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 12–21. Association for Computational Linguistics.

Chalaguine, L. A., Hunter, A., Hamilton, F. L., and Potts, H. W. W. (2019). Impact of argument type and concerns

in argumentation with a chatbot. In *31st IEEE International Conference on Tools with Artificial Intelligence*, pages 1557–1562. IEEE.

Chen, S., Khashabi, D., Callison-Burch, C., and Roth, D. (2019). PerspectroScope: A window to the world of diverse perspectives. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 129–134. Association for Computational Linguistics.

Chernodub, A., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C., and Panchenko, A. (2019). Targer: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 195–200. Association for Computational Linguistics.

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., and Cieliebak, M. (2019). Survey on evaluation methods for dialogue systems. *ArXiv*, abs/1905.04071.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Galitsky, B. (2019). Enabling a bot with understanding argumentation and providing arguments. In *Developing Enterprise Chatbots*, pages 465–532. Springer.

Gleize, M., Shnarch, E., Choshen, L., Dankin, L., Moshkowich, G., Aharonov, R., and Slonim, N. (2019). Are you convinced? choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976. Association for Computational Linguistics.

Habernal, I. and Gurevych, I. (2016a). What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223. Association for Computational Linguistics.

Habernal, I. and Gurevych, I. (2016b). Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1589–1599. Association for Computational Linguistics.

Hayes, A. F. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Kwon, O.-W., Lee, K., Roh, Y.-H., Huang, J.-X., Choi, S.-K., Kim, Y.-K., Jeon, H. B., Oh, Y. R., Lee, Y.-K., Kang, B. O., et al. (2015). Genietutor: A computer-assisted second-language learning system based on spoken language understanding. In *Natural language dialog systems and intelligent assistants*, pages 257–262. Springer.

Laroche, R., Putois, G., Bretier, P., Aranguren, M., Velkovska, J., Hastie, H., Keizer, S., Yu, K., Jurcicek, F., Lemon, O., et al. (2011). D6. 4: Final evaluation of classic towninfo and appointment scheduling systems. *Report D6*, 4.

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

Levy, R., Bogin, B., Gretz, S., Aharonov, R., and Slonim, N. (2018). Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081. Association for Computational Linguistics.

Lippi, M. and Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Mo, K., Zhang, Y., Li, S., Li, J., and Yang, Q. (2018). Personalizing a dialogue system with transfer reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press.

Potthast, M., Gienapp, L., Euchner, F., Heilenkötter, N., Weidmann, N., Wachsmuth, H., Stein, B., and Hagen, M. (2019). Argument search: Assessing argument relevance. In *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 1117–1120. ACM.

Rach, N., Weber, K., Pragst, L., André, E., Minker, W., and Ultes, S. (2018). Eva: A multimodal argumentative dialogue system. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 551–552. ACM.

Rach, N., Langhammer, S., Minker, W., and Ultes, S. (2019). Utilizing argument mining techniques for argumentative dialogue systems. In *9th International Workshop on Spoken Dialogue System Technology*, pages 131–142. Springer.

Rakshit, G., Bowden, K. K., Reed, L., Misra, A., and Walker, M. (2019). Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents*, pages 45–52. Springer.

Reed, C. and Norman, T. (2003). *Argumentation machines: New frontiers in argument and computation*, volume 9. Springer Science & Business Media.

Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., and Gurevych, I. (2019). Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578. Association for Computational Linguistics.

Rosenfeld, A. and Kraus, S. (2016). Strategical argumentative agent for human persuasion. In *Proceedings of the*

Twenty-second European Conference on Artificial Intelligence, pages 320–328. IOS Press.

Sakai, K., Inago, A., Higashinaka, R., Yoshikawa, Y., Ishiguro, H., and Tomita, J. (2018). Creating large-scale argumentation structures for dialogue systems. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Schatzmann, J., Thomson, B., Weilhammer, K., Ye, H., and Young, S. (2007). Agenda-based user simulation for bootstrapping a pomdp dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152. Association for Computational Linguistics.

Sprent, P. (2011). Fisher exact test. *International encyclopedia of statistical science*, pages 524–525.

Stab, C., Daxenberger, J., Stahlhut, C., Miller, T., Schiller, B., Tauchmann, C., Eger, S., and Gurevych, I. (2018a). Argumentext: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 21–25. Association for Computational Linguistics.

Stab, C., Miller, T., Schiller, B., Rai, P., and Gurevych, I. (2018b). Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, volume Long Papers, pages 3664–3674. Association for Computational Linguistics.

Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., Venezian, E., Lahav, D., Jacovi, M., Aharonov, R., and Slonim, N. (2019). Automatic argument quality assessment - new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5624–5634. Association for Computational Linguistics.

Ultes, S., Schmitt, A., and Minker, W. (2013). On quality ratings for spoken dialogue systems–experts vs. users. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 569–578. Association for Computational Linguistics.

Ultes, S., Miehle, J., and Minker, W. (2019). On the applicability of a user satisfaction-based reward for dialogue policy learning. In *Advanced Social Interaction with Agents*, pages 211–217. Springer.

Wachsmuth, H., Naderi, N., Habernal, I., Hou, Y., Hirst, G., Gurevych, I., and Stein, B. (2017a). Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255. Association for Computational Linguistics.

Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., Hirst, G., and Stein, B. (2017b). Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187. Association for Computational Linguistics.

Wachsmuth, H., Potthast, M., Al Khatib, K., Ajjour, Y., Puschmann, J., Qu, J., Dorsch, J., Morari, V., Bevendorff, J., and Stein, B. (2017c). Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining (ArgMining 2017) at EMNLP*, pages 49–59. Association for Computational Linguistics.

Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280. Association for Computational Linguistics.

Walker, M., Kamm, C., and Litman, D. (2000). Towards developing general models of usability with paradise. *Natural Language Engineering*, 6(3-4):363–377.

Weber, K., Janowski, K., Rach, N., Weitz, K., Minker, W., Ultes, S., and André, E. (2020). Predicting persuasive effectiveness for multimodal behavior adaptation using bipolar weighted argument graphs. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems (forthcoming)*.

Wen, T.-H., Gašić, M., Kim, D., Mrkšić, N., Su, P.-H., Vandyke, D., and Young, S. (2015). Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 275–284. Association for Computational Linguistics.

Wilcock, G. and Jokinen, K. (2019). Towards increasing naturalness and flexibility in human-robot dialogue systems. In *Proceedings of the 10th International Workshop on Spoken Dialog Systems Technology (IWSDS)*.

Yuan, T., Moore, D., and Grierson, A. (2008). A human-computer dialogue system for educational debate: A computational dialectics approach. *International Journal of Artificial Intelligence in Education*, 18(1):3–26.

Yuan, T. (2004). *Human-computer debate, a computational dialectics approach*. Ph.D. thesis, Leeds Metropolitan University.