A Preliminary Evaluation of a Privacy-Preserving Dialogue System

Bettina Fazzinga[⋆1], Andrea Galassi[⋆2(✉)], and Paolo Torroni[⋆2]

[1] ICAR CNR, Rende, Italy; DICES, University of Calabria, Rende, Italy
bettina.fazzinga@unical.it
[2] DISI, University of Bologna, Bologna, Italy
a.galassi@unibo.it paolo.torroni@unibo.it

Abstract. Dialogue systems are AI applications widely used in many contexts requiring user interaction. However, unconstrained interaction may lead to users communicating sensitive data. This raises concerns about how these systems handle personal data, and about their compliance with relevant laws, regulations, and ethical principles. We propose to integrate advanced natural language processing techniques in a dialogue system architecture based on computational argumentation, ensuring that user data are ethically managed and regulations are respected. A preliminary experimental evaluation of our proposal over a COVID-19 vaccine information case study shows promising results.

Keywords: Dialogue systems · Sentence embeddings · Data Protection · Expert systems · Chatbots · COVID-19.

1 Introduction

The idea of an artificial agent capable of communicating with the user through natural language has inspired researchers since the early days of artificial intelligence. The recent development in language technologies has nourished this ambition further and now the full maturation of intelligent dialogue systems does not seem a far dream any longer. Their adoption allows immediate support to any user, making them incredibly valuable for companies and public administrations alike. In fact, they are being used by public administrations to help citizens to request services,[3] but also to provide updates and information on pressing matters, such as COVID-19 [23].[4]

⋆ Equal contribution.

[3] https://www.canada.ca/en/employment-social-development/services/my-account/terms-use-chatbot.html

[4] https://government.economictimes.indiatimes.com/news/digital-india/covid-19-govt-launches-facebook-and-messenger-chatbot/74843125

The pervasive presence of information-providing chatbots and assistive dialogue systems in many delicate context raises the need for *trustworthy* AI methods, which can guarantee citizens protection against possible misuses of technology. We believe that trustworthiness demands transparency, explainability, correctness, and that it requires architectural choices that take data access into account from the very beginning. In fact, chatbots should not only process data through transparent and verifiable methods following appropriate regulations, but also provide explanations of their outputs in a manner adapted to the intended (human) user. This is especially true in the public sector and when the interaction among different legal entities is involved.

In our earlier work [11], we identified a combination of computational argumentation and language technology as a possible answer to some of these challenges. We described an architecture for AI dialogue systems where user interaction is carried out *in natural language*, both for providing information to the user and to answer user queries about the *reasons* leading to the system output (explainability). We proposed to use computational argumentation techniques to realize a *transparent reasoning module* with a *rigorous, verifiable semantics* (transparency, auditability). We also underlined the importance of modularity in the architecture's design, to decouple the natural language interface, where user data is processed, and the reasoning module, where expert knowledge is used to generate outputs (privacy and data governance). In [11], we focused on the computational argumentation module, describing how the system works to compute answers.

The focus of this work, instead, is on the language module's design and its initial evaluation. The main idea is simple: to use *sentence embeddings* and a similarity function to match user inputs with a set of natural language sentences describing relevant facts. In order to evaluate whether this concept may work in practice, we constructed a tiny dataset of sentences describing user information in possible dialogues regarding COVID-19 vaccines. For example, we encoded different ways users may express whether they suffer from drug allergies, or are immunosuppressed. We run a preliminary experimentation to compare different sentence embeddings and hyperparameters, obtaining encouraging results.

Our presentation starts by discussing related approaches (Section 2). Section 3 gives a high-level description of the system architecture, while we illustrate the implementation in more detail in Section 4. In Section 5 we offer an initial empirical evaluation of the language module, pointing to the feasibility of the approach in real-world contexts. We conclude and address future developments in Section 6.

2 Related Work

Our work is positioned at the intersection of two areas: computational argumentation and natural language understanding. While computational argumentation has had significant applications in the context of automated dialogues

among software agents, its combination with systems able to interact in natural language in socio-technical systems has been more recent [5].

Dialogue systems are typically divided between conversational agents, which support open-domain dialogues, and task-oriented agents, which assist the user in a specific task [7,8]. Our proposal falls in the second category. The task is to obtain information on a specific topic. The advancement of deep learning techniques and their successful application in many Natural Language Processing tasks has lead researchers to investigate the use of neural architectures for end-to-end dialogue systems [27,22]. However, these architectures have downsides too. Their training phase usually has a heavy computational footprint, and it requires the construction of large corpora for the specific use cases. Moreover, they are often vulnerable to biases, privacy violations, adversarial attacks, and safety concerns [1,10,14,21]. Finally, reuse and adaptation to a different domain typically require building a new training corpus and a complete retrain. Given our focus on user protection and our aim to develop a general, data-independent approach, our system is modular (as opposed to end-to-end), and does not involve any training phase. Since it only uses off-the-shelf tools, it can be applied to new contexts without having to construct new training corpora.

The protection of users' identity and personal information is usually addressed through redaction [34] or sanitization [4] methods. Such techniques are usually built on large but domain-specific datasets [24,31]. Nonetheless, these techniques are still far from guaranteeing zero-risk to the user [19], and often focus only predefined categories of entities, ignoring elements that may play a role in re-identifying the individual [20].

Our proposal is general-purpose and aims at maximizing user protection. The main idea is that user information is neither shared nor stored in the application. Instead, it is replaced by a collection of general, "sanitized" information elements that are pertinent to the case at hand. Our approach is akin to Information Retrieval-based chatbots, where dialogue agents retrieve their answer from a knowledge base made of dialogues, treating the user's sentences as queries. In the same vein, Charras et al. [6] use sentence similarity to retrieve the desired answer from a knowledge base made of dialogues, while Chalaguine and Hunter [5] retrieve an answer from a graph. Both work compare sentences through the cosine similarity between the TF-IDF representation of the sentences, but Charras et al. explore also the use of *doc-to-vec* [18] representation. However, the design of these approaches does not include a history of conversation, nor the possibility to retrieve multiple information elements within a single interaction. This is a strong limitation in real-world scenarios, where information cannot be considered in isolation, but on the contrary, multiple pieces must be considered at the same time, independently whether they have been communicated in a single sentence or at different points in the dialogue. Another limitation of previous approaches is their relying on *lexical*, instead of *semantic* similarity. Conversely, we use sentence-level embeddings, which enables semantic similarity measures. Moreover, we consider the possibility of retrieving multiple information in a

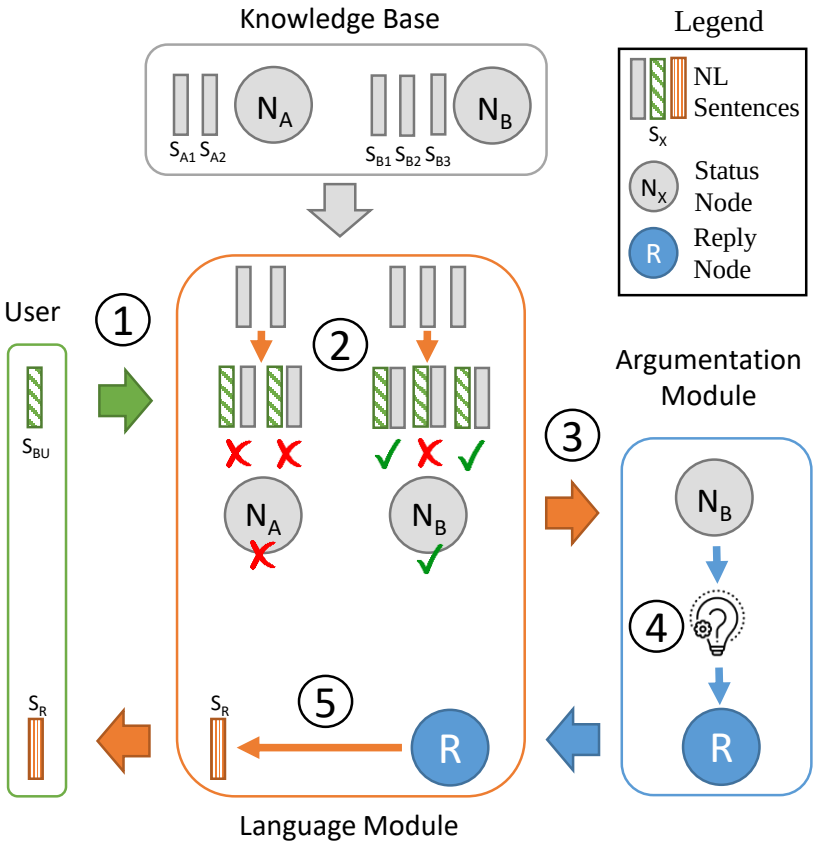

Fig. 1. System architecture and example of interaction with the user. Natural language sentences are represented as rectangles and indicated with S, while circles are used for status and reply nodes (indicated respectively with N and R). We represent a case where nodes and sentences refers to two concepts, A and B, and the user sentence regards B. The information provided by the user is represented with the green color and by diagonal stripes. It is easy to see that such information does not reach the argumentation module.

single interaction, and to maintain a history of retrieved concept thanks to a reasoning module based on argumentation.

3 System Architecture

Our architecture consists of two main modules: the *language module* and the *argumentation module*. The former, which is the focus of this work, is responsible for user interaction: it processes user input and generates answers, all in natural

language. The latter, described in our previous work [11], receives the processed information and reasons over it, so as to find the appropriate answer, according to a *knowledge base* (KB) of the domain of interest. The KB is built by domain experts and consists of an argumentation graph, having a node for each possible relevant piece of information that could be communicated to the system by users, called *status* nodes (that thus contain factual information about possible users conditions/statuses), and a node for each possible answer, called *reply* nodes. A set of natural language sentences is also associated with each status node. In this way, we have a natural language representation of possible ways a user would express what a node is meant to encode. These different representations of facts could be produced by domain experts or crowd-sourced as proposed by Chalaguine and Hunter [5].

The interaction with the user is represented in Figure 1 and it is structured as follows:

1. The user inputs one or more sentences.
2. The language module compares each sentence with the KB sentences associated with status nodes, obtaining a set of "matched" sentences, corresponding to a set of "activated" status nodes. These nodes, collectively, represent the specific use case described by the user.
3. The list of activated status nodes is sent to the argumentation module.
4. The argumentation module performs reasoning over the activated nodes, resulting in an answer, or in the request for more information. Either way, a node is selected and this selection is communicated to the language module.
5. The language module elaborates the output of the argumentation module and produces a natural language reply to the user.

Since the sentence provided by the user may match multiple KB sentences, multiple nodes may be activated in a single interaction. At the same time, the system allows to provide information over multiple interaction, since the argumentation module stores the history of activated nodes in its "memory".

We shall remark that such an architecture protects the privacy of its users on two levels. First of all, the system ignores any information that does not match a KB sentence. The KB is not user-specific: it only represents general knowledge, in a user-independent way. Therefore, any information not strictly relevant to the scenario is filtered out. Then, during step 3, the argumentation module receives only the list of activated nodes, not the sentences as the user has formulated it. The outcome of this procedure is similar to the one produced by a "sanitization" process [4], since the all the relevant information are kept, but in a form that is general and does not contain any information that may lead to the identification of the user.

Our proposal is therefore suited with any scenarios where the language module is authorized and entrusted to manage the user's data, but the argumentation module is not. One of them is a client-server implementation, where the client side includes the language module, while the argumentation module resides on the server side. In this case, all the personal information of the user will remain on the client side, and only the sanitized version of them will reach the

server. Another possibility would be a multi-agent system, where the two modules are managed by different organizations, e.g. a service provided jointly by a government and by a private company.

Last but not least, the reasoning module is transparent, rigorous, and verifiable, allowing the users to request for more information regarding the provided answer. More details about this process are discussed in Fazzinga et al. [11].

4 Language Module

One of the main objectives of our proposal is keeping the approach as general as possible. While many scenarios may gain advantage from tailored NLP solutions, their construction may be too costly, even impossible. We have therefore decided to follow previous works [6,5] and to assess which KB sentences match the user's ones by computing the similarity between their embedded representation. But instead of relying on simple syntactical representation, we propose to use state-of-the-art techniques, apt to capture the semantic content of the sentences.

In particular, we encode both the user sentences and the KB sentences using *sentence embeddings*. These are high-dimensional numerical representations of textual sentences that can be computed using (pre-trained) neural architectures. Many embeddings have been proposed along the years [26,25], and modern attention-based [13] sentence embeddings such as BERT [9] do not only model the syntactic content and structure of a sentence, but also capture its meaning. Ideally, if two sentences have a similar meaning, they will be mapped onto similar sentence embeddings. Sentence embeddings have been used successfully in a variety of NLP tasks, including hard ones such as understanding negations and speculations, and have shown to outperform traditional rule-based systems [30].

Among the many possible models, we have decided to focus on Sentence-BERT models [28], which are specifically trained to perform well on tasks of sentence similarity. While it is possible to train new models for specific domains or tasks, many pre-trained models are already available and can be used as off-the-shelf tools without the need of creating a corpus, nor to perform a training or fine-tuning steps.

The similarity between two embeddings can be computed using any similarity function that operates on high-dimensional numerical vectors. We use the Bray-Curtis similarity [2] since it has led to satisfactory results in related settings before [12], but other measures, such as cosine similarity [16], may be equally valid options. A possible alternative to the use of sentence embeddings combined with a similarity measure may be the use of neural architectures specifically trained to perform this task, such as cross-encoders [28]. However, the computational footprint of these techniques may be too heavy in most contexts, since they require to encode and process any possible pair at any step of iteration.

Given a measure of similarity between two sentences, we transform it to a Boolean value by applying a threshold, which is an hyper-parameter of the architecture. In this way, we discriminate between the pairs of sentences that are similar enough to be considered "a match", and those that are not.

Table 1. Sentences used in our case study and the status node they are associated with.

Node ID	Sent. ID	Sentence
N1	S1	I am celiac
N1	S2	I suffer from the celiac disease
N1	S3	I am afflicted with the celiac disease
N1	S4	I have the celiac disease
N1	S5	I recently found out to be celiac
N1	S6	I have suffered from celiac disease since birth
N2	S7	I do not have the celiac disease
N2	S8	I am not celiac
N2	S9	I do not suffer from the celiac disease
N2	S10	I am not afflicted with the celiac disease
N3	S11	I am not immunosuppressed
N3	S12	I do not suffer from immunosuppression
N3	S13	I am not afflicted with immunosuppression
N4	S14	I am immunosuppressed
N4	S15	I suffer from immunosuppression
N4	S16	I am afflicted with immunosuppression
N4	S17	I do suffer from immunosuppression
N4	S18	I indeed suffer from immunosuppression
N4	S19	I recently found out to be immunosuppressed
N5	S20	I do not have any drug allergy
N5	S21	I do not suffer from drug allergies
N5	S22	I do not suffer from any drug allergy
N5	S23	I am not afflicted with any drug allergy
N5	S24	I do not have medication allergies
N5	S25	I do not have any medication allergy
N6	S26	I have a drug allergy
N6	S27	I do have a drug allergy
N6	S28	I have a serious drug allergy
N6	S29	I suffer from drug allergy
N6	S30	I am afflicted with drug allergies
N6	S31	I suffer from medication allergies

5 Experimental Evaluation

To assess the effectiveness of our language module based on sentence embeddings and similarity measures, we run a preliminary experimentation on a small-sized dataset built around the use case of vaccines for COVID-19. We are especially interested in evaluating our method on sentences with a similar syntactic structure, but different meaning (e.g., a sentence and its negation).

5.1 Setting

In the context of COVID-19 vaccines, our dialogue system helps the users to understand whether or not, and eventually where (hospital or generic site), they

can get vaccinated, depending on their health status. Our KB has been built from the information published by the Italian Medicines Agency (AIFA) on their website (https://www.aifa.gov.it/en/). For example, for people suffering of diabetes, no special recommendation is given, so our system will tell users that they can be vaccinated at any site (without the need of going to the hospital), while in the case that they suffer from bronchial asthma, our system will tell them to get vaccinated at the hospital. In this context, it is essential for our system to perfectly understand users health conditions, so, in the following, we focus on the matching phase between user sentences and the information stored in KB.

We consider a case study with a KB made of only 6 status nodes, corresponding to the presence/absence of 3 particular medical conditions, i.e., celiac disease, immunosuppression, and drug allergy. For each node, our KB contains from 3 to 7 sentences that can be used to express the same concept (see Table 1).

Instead of using an additional set of sentences to simulate the user input, we compare the KB sentences between each other and verify whether sentences belonging to the same node do match. To evaluate our method quantitatively, we treat it as a binary classification task on every possible pair of (different) sentences. If the two sentences belong to the same status node, their pair is considered a positive instance, otherwise it is considered negative.

In our experiment we compare different models of sentence embeddings and different threshold criteria. For sentence embeddings we evaluate the following Sentence-BERT [28] models:[5]

- `stsb-mpnet`: based on MPNet [33] and pre-trained for semantic similarity on the STSbenchmark [3].
- `paraphrase-mpnet`: based on MPNet and pre-trained for paraphrase mining.
- `paraphrase-TinyBERT-L6`: based on TinyBERT [15] and pre-trained for paraphrase mining.
- `paraphrase-MiniLM-L3`: based on MiniLM [35] and pre-trained for paraphrase mining.
- `nq-distilbert`: based on DistilBERT [32] and pre-trained for question answering on Google's Natural Questions dataset [17].
- `paraphrase-multilingual-mpnet`: multilingual extension [29] of the monolingual model. We have decided to include this model in the perspective of future multi-lingual applications.

We also include TF-IDF representation as in Charras et al. [6], Chalaguine and Hunter [5], using the entire set of sentences to create the vocabulary. As thresholds, we use three arbitrary values (0.75, 0.70, 0.65), plus two values based on the distribution of the similarity scores: one is given by the average of the similarities (mean), and the other one is given by the sum between the average similarity and the standard deviation (mean+std).

For each combination of models and thresholds, we measure precision, recall, and F1 score of the positive class (see Table 2). Precision is especially important:

[5] All the implementations of the models are taken from http://www.sbert.net/.

false positives can be seen as cases where the system "misunderstands" the input of the user, and therefore precision can be seen as a measure of *correctness*. Recall instead can be seen as a measure of the ability of the system to not "miss" information communicated by the user. For the purposes of our system, poor recall is a less serious problem than poor precision, since the argumentation module proactively asks the user for missing bits of information that would influence the final result. In our perspective, the priority must be to guarantee the correctness of the final answer, even if this means that the system will, in some cases, ask for information that the user has already submitted. For this reason, we use precision as the main evaluation metric.

5.2 Results and Discussion

Our results clearly show that the `stsb-mpnet` and the `paraphrase-mpnet` models are the best ones, with the former achieving perfect precision with all the fixed similary scores and the latter achieving equivalent or even better F1 scores with every threshold. In particular, they both achieve an almost perfect result (only one false positive, no false negatives) using the mean+std threshold. The `paraphrase-multilingual-mpnet` model perform slightly worse than the monolingual version, providing encouraging results in the perspective of future multilingual applications. The TF-IDF model is the one that performs worse with all the threshold values, in part probably due to the small size of the vocabulary.

Table 3 shows an example of matching using sentences from S1 to S19, which are those related to the status nodes "Has celiac disease", "Has not celiac disease", "Is immunosuppressed", "Is not immunosuppressed". The matches are computed by the `stsb-mpnet` and the `paraphrase-mpnet` models using a threshold value of 0.65. The former achieves perfect precision but not perfect recall, and indeed we can see that it misses some matches, such as S8 and S10. The latter reaches perfect recall but not precision, which indicates the presence of false positives e.g. the pair S1 and S8. Some of these false positives might be particularly dangerous in a real application since they mean that the system has misunderstood a sentence for its negation, e.g. the sentence "I am not celiac" as "I am celiac". The argumentation module would be able to detect such conflicts and in future works we plan to include conflict resolution modules and procedures. A careful user experience design may also be able to mitigate the issue, for instance by displaying relevant pieces of information interactively as they are understood by the system.

These results are encouraging and motivate us to continue along this research direction. Nonetheless, our research is still in its early stages and we are aware that a proper and sound evaluation of the whole proposal would require to include more nodes, a rigorous split between calibration and test sentences, and should eventually be validated by human testers.

Table 2. Experimental results of the embedding models and the threshold criterion on the sentence matching task.

Embedding Model	Threshold	P	R	F1
stsb-mpnet	mean	0.33	**1.00**	0.50
	mean+std	0.99	**1.00**	**0.99**
	0.75	**1.00**	0.67	0.80
	0.70	**1.00**	0.86	0.92
	0.65	**1.00**	0.97	**0.99**
paraphrase-mpnet	mean	0.32	**1.00**	0.49
	mean+std	0.99	**1.00**	**0.99**
	0.75	**1.00**	0.86	0.92
	0.70	**1.00**	0.94	0.97
	0.65	0.96	**1.00**	0.98
paraphrase-TinyBERT-L6	mean	0.40	**1.00**	0.57
	mean+std	0.72	0.99	0.83
	0.75	**1.00**	0.46	0.63
	0.70	0.94	0.70	0.80
	0.65	0.81	0.94	0.87
paraphrase-MiniLM-L3	mean	0.43	**1.00**	0.60
	mean+std	0.55	0.96	0.70
	0.75	0.81	0.43	0.57
	0.70	0.66	0.61	0.63
	0.65	0.57	0.87	0.69
nq-distilbert	mean	0.37	**1.00**	0.54
	mean+std	0.50	0.75	0.60
	0.75	0.96	0.33	0.49
	0.70	0.64	0.46	0.54
	0.65	0.58	0.64	0.61
paraphrase-multilingual-mpnet	mean	0.31	**1.00**	0.47
	mean+std	0.99	0.97	0.98
	0.75	**1.00**	0.81	0.90
	0.70	0.98	0.93	0.96
	0.65	0.90	**1.00**	0.95
TF-IDF	mean	0.27	0.71	0.39
	mean+std	0.34	0.39	0.36
	0.75	0.38	0.07	0.12
	0.70	0.33	0.07	0.12
	0.65	0.50	0.14	0.22

6 Conclusion

We proposed the integration of advanced sentence embeddings into a modular dialogue system architecture based on argumentation, so as to support privacy by design.

In particular, the language module is the only module that processes user input, and its output to the argumentation module is devoid of any sensitive,

Table 3. Matches computed by the models using the 0.65 threshold value on sentences from S1 to S19. The + symbol indicates the correct matches. The • symbol indicates the matches computed using the `stsb-mpnet model`. The ∘ symbol indicates the matches computed using the `paraphrase-mpnet` model.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19
S1	+•∘	+•∘	+•∘	+•∘	+•∘	+•∘		∘											
S2	+•∘	+•∘	+•∘	+•∘	+•∘	+•∘													
S3	+•∘	+•∘	+•∘	+•∘	+•∘	+•∘													
S4	+•∘	+•∘	+•∘	+•∘	+•∘	+•∘	∘												
S5	+•∘	+•∘	+•∘	+•∘	+•∘	+•∘													
S6	+•∘	+•∘	+•∘	+•∘	+•∘	+•∘													
S7				∘			+•∘	+•∘	+•∘	+•∘									
S8	∘						+•∘	+•∘	+•∘	+∘									
S9							+•∘	+•∘	+•∘	+•∘									
S10							+•∘	+∘	+•∘	+•∘									
S11											+•∘	+•∘	+∘	∘					
S12											+•∘	+•∘	+•∘						
S13											+∘	+•∘	+•∘						
S14											∘			+•∘	+•∘	+•∘	+•∘	+•∘	+•∘
S15														+•∘	+•∘	+•∘	+•∘	+•∘	+•∘
S16														+•∘	+•∘	+•∘	+•∘	+•∘	+•∘
S17														+•∘	+•∘	+•∘	+•∘	+•∘	+•∘
S18														+•∘	+•∘	+•∘	+•∘	+•∘	+•∘
S19														+•∘	+•∘	+•∘	+•∘	+•∘	+•∘

personal, or irrelevant piece of information the user may have written. The output of this module can therefore be seen as the anonymized and sanitized version of the user's sentences. This makes the system amenable to distributed, multi-party implementations, where domain knowledge representation and reasoning may be left to third parties, and the user interface completely decouples the user input from the arguments used in the reasoning. We shall point out that guaranteeing the anonymization of user data, may not only a desirable feature, but even a legal requirement in some contexts, such as those regulated by EU's GDPR[6]. Importantly, the architecture is general-purpose and does not require domain-specific training or reference corpora.

The COVID-19 vaccines case study has given the context for a preliminary experimental evaluation. Our results indicate that the use of sentence embeddings computed by pre-trained neural architectures greatly outperforms the TF-IDF model used in other approaches, leading to *precise* matches. We also emphasized the importance of precision and correctness over recall.

In future developments we aim at extending our experimental evaluation, including human testers in the loop. We also want to investigate additional case studies, potentially involving languages other than English. It would also be

[6] See https://eur-lex.europa.eu/eli/reg/2016/679/oj.

interesting to extend our architecture with techniques for the detection and the resolution of conflicts, especially false positives, both in the reasoning module and in the language module. Finally, we would like to provide the user the possibility to directly correct matches. That could further improve the transparency of our architecture and reduce the number of false positives. However, that would require redesigning user interaction, which is now intentionally simple, possibly making it more complicated and less intuitive.

Acknowledgments

The research reported in this work was partially supported by the EU H2020 ICT48 project "Humane AI Net" under contract #952026.

References

1. Barikeri, S., Lauscher, A., Vulic, I., Glavas, G.: Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. In: ACL/IJCNLP (1). pp. 1941–1955. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.acl-long.151
2. Bray, J.R., Curtis, J.T.: An ordination of the upland forest communities of Southern Wisconsin. Ecological Monographs **27**(4), 325–349 (1957). https://doi.org/10.2307/1942268
3. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 1–14. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). https://doi.org/10.18653/v1/S17-2001
4. Chakaravarthy, V.T., Gupta, H., Roy, P., Mohania, M.K.: Efficient techniques for document sanitization. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. p. 843–852. CIKM '08, Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1458082.1458194
5. Chalaguine, L.A., Hunter, A.: A persuasive chatbot using a crowd-sourced argument graph and concerns. In: Prakken, H., Bistarelli, S., Santini, F., Taticchi, C. (eds.) COMMA. Frontiers in Artificial Intelligence and Applications, vol. 326, pp. 9–20. IOS Press (2020). https://doi.org/10.3233/FAIA200487
6. Charras, F., Dubuisson Duplessis, G., Letard, V., Ligozat, A.L., Rosset, S.: Comparing System-response Retrieval Models for Open-domain and Casual Conversational Agent. In: WOCHAT. Los Angeles, United States (2016), https://hal.archives-ouvertes.fr/hal-01782262
7. Chen, H., Liu, X., Yin, D., Tang, J.: A survey on dialogue systems: Recent advances and new frontiers. SIGKDD Explor. Newsl. **19**(2), 25–35 (Nov 2017). https://doi.org/10.1145/3166054.3166058
8. Deriu, J., Rodrigo, Á., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., Cieliebak, M.: Survey on evaluation methods for dialogue systems. Artif. Intell. Rev. **54**(1), 755–810 (2021). https://doi.org/10.1007/s10462-020-09866-x

9. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) NAACL-HLT (1). pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423
10. Dinan, E., Fan, A., Williams, A., Urbanek, J., Kiela, D., Weston, J.: Queens are powerful too: Mitigating gender bias in dialogue generation. In: EMNLP (1). pp. 8173–8188. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.656
11. Fazzinga, B., Galassi, A., Torroni, P.: An argumentative dialogue system for COVID-19 vaccine information. In: Baroni, P., Benzmüller, C., Wáng, Y.N. (eds.) Logic and Argumentation - 4th International Conference, CLAR 2021, Hangzhou, China, October 20-22, 2021, Proceedings. Lecture Notes in Computer Science, vol. 13040, pp. 477–485. Springer (2021). https://doi.org/10.1007/978-3-030-89391-0_27
12. Galassi, A., Drazewski, K., Lippi, M., Torroni, P.: Cross-lingual annotation projection in legal texts. In: COLING. pp. 915–926. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.79
13. Galassi, A., Lippi, M., Torroni, P.: Attention in natural language processing. IEEE Trans. Neural Networks Learn. Syst. **32**(10), 4291–4308 (2021). https://doi.org/10.1109/TNNLS.2020.3019893
14. Henderson, P., Sinha, K., Angelard-Gontier, N., Ke, N.R., Fried, G., Lowe, R., Pineau, J.: Ethical challenges in data-driven dialogue systems. In: AIES. p. 123–129. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3278721.3278777
15. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: TinyBERT: Distilling BERT for natural language understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 4163–4174. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.372
16. Kenter, T., de Rijke, M.: Short text similarity with word embeddings. In: CIKM. p. 1411–1420. CIKM '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2806416.2806475
17. Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A.P., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M., Dai, A.M., Uszkoreit, J., Le, Q., Petrov, S.: Natural questions: a benchmark for question answering research. Trans. Assoc. Comput. Linguistics **7**, 452–466 (2019). https://doi.org/https://doi.org/10.1162/tacl_a_00276
18. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: ICML. JMLR Workshop and Conference Proceedings, vol. 32, pp. 1188–1196. JMLR.org (2014), http://proceedings.mlr.press/v32/le14.html
19. Li, B., Vorobeychik, Y., Li, M., Malin, B.A.: Scalable iterative classification for sanitizing large-scale datasets. IEEE Trans. Knowl. Data Eng. **29**(3), 698–711 (2017). https://doi.org/10.1109/TKDE.2016.2628180
20. Lison, P., Pilán, I., Sánchez, D., Batet, M., Øvrelid, L.: Anonymisation models for text data: State of the art, challenges and future directions. In: ACL/IJCNLP (1). pp. 4188–4203. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.acl-long.323
21. Liu, H., Dacon, J., Fan, W., Liu, H., Liu, Z., Tang, J.: Does gender matter? towards fairness in dialogue systems. In: COLING. pp. 4403–4416. International

Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.390
22. Luo, L., Huang, W., Zeng, Q., Nie, Z., Sun, X.: Learning personalized end-to-end goal-oriented dialog. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 6794–6801 (Jul 2019). https://doi.org/10.1609/aaai.v33i01.33016794
23. Miner, A.S., Laranjo, L., Kocaballi, A.B.: Chatbots in the fight against the COVID-19 pandemic. npj Digital Medicine **3(1)** (2020). https://doi.org/10.1038/s41746-020-0280-0
24. Nguyen, H., Cavallari, S.: Neural multi-task text normalization and sanitization with pointer-generator. In: Proceedings of the First Workshop on Natural Language Interfaces. pp. 37–47. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.nli-1.5
25. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: EMNLP. pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1162
26. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Walker, M.A., Ji, H., Stent, A. (eds.) NAACL-HLT. pp. 2227–2237. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/n18-1202
27. Rajendran, J., Ganhotra, J., Singh, S., Polymenakos, L.: Learning end-to-end goal-oriented dialog with multiple answers. In: EMNLP. pp. 3834–3843. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1418
28. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: EMNLP/IJCNLP (1). pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1410
29. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) EMNLP. pp. 4512–4525. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.365
30. Rivera Zavala, R., Martinez, P.: The impact of pretrained language models on negation and speculation detection in cross-lingual medical text: Comparative study. JMIR Med Inform **8**(12), e18953 (Dec 2020). https://doi.org/10.2196/18953
31. Sánchez, D., Batet, M., Viejo, A.: Utility-preserving privacy protection of textual healthcare documents. J. Biomed. Informatics **52**, 189–198 (2014). https://doi.org/10.1016/j.jbi.2014.06.008
32. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. In: The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS (2019), http://arxiv.org/abs/1910.01108
33. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.: Mpnet: Masked and permuted pre-training for language understanding. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) NeurIPS (2020), https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html
34. Szarvas, G., Farkas, R., Busa-Fekete, R.: Research paper: State-of-the-art anonymization of medical records using an iterative machine learning framework. J. Am. Medical Informatics Assoc. **14**(5), 574–580 (2007). https://doi.org/10.1197/jamia.M2441

35. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) NeurIPS (2020), https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html