Automatic Generation of Structured Domain Knowledge for Dialogue-based XAI Systems

Carolin Schindler[1], Isabel Feustel[1], Niklas Rach[2], Wolfgang Minker[1]

[1]Institute of Communications Engineering, Ulm University, Germany
[2]Tensor AI Solutions GmbH, Germany

Correspondence: carolin.schindler@uni-ulm.de

Abstract

Explanatory dialogue systems serve as intuitive interface between non-expert users and explainable AI (XAI) systems. The interaction with these kind of systems benefits especially from the integration of structured domain knowledge, e. g., by means of bipolar argumentation trees. So far, these domain-specific structures need to be created manually, therewith impairing the flexibility of the system with respect to the domain. We address this limitation by adapting an existing pipeline for topic-independent acquisition of argumentation trees in the field of persuasive, argumentative dialogue to the area of explanatory dialogue. This shift is achieved by *a)* introducing and investigating different formulations of auxiliary claims per feature of the explanation of the AI model, *b)* exploring the influence of pre-grouping of the arguments with respect to the feature they address, *c)* suggesting adaptions to the existing algorithm of the pipeline for obtaining a tree structure, and *d)* utilizing a new approach for determining the type of the relationship between the arguments. Through a step-wise expert evaluation for the domain *titanic survival*, we identify the best performing variant of our pipeline. With this variant we conduct a user study comparing the automatically generated argumentation trees against their manually created counterpart in the domains *titanic survival* and *credit acquisition*. This assessment of the suitability of the generated argumentation trees for a later integration into dialogue-based XAI systems as domain knowledge yields promising results.

1 Introduction

Explainable artificial intelligence (XAI) is recently gaining considerable attention as a means to improve the transparency of AI models and therewith enabling humans to understand the decisions made by them (Adadi and Berrada, 2018). However, due to the complexity of AI-based systems, it can be challenging to provide XAI explanations that are comprehensible also to non-expert users. By integrating XAI explanations into human-machine dialogue, users can ask clarifying questions and receive tailored explanations (Miller, 2019). In addition, the combination with domain knowledge has the potential to foster a deeper understanding of the behavior of the AI system (Feustel et al., 2024). We follow this line of research by introducing an automatized approach for the retrieval of the required domain knowledge from arbitrary documents. Viewing explanatory reasoning as argumentative (Mercier and Sperber, 2011), we encode the domain knowledge as bipolar argumentation trees (Stab and Gurevych, 2014) for the use in explanatory dialogue systems. Within these tree structures, the domain knowledge is encoded as arguments with supporting or attacking relationships among each other.

While the integration of domain knowledge can be beneficial for explanatory systems, the manual effort for creating structured domain knowledge impairs the flexibility of a corresponding system with respect to the domains it can provide meaningful explanations for. To overcome this limitation and therewith make the integration of domain-specific knowledge more feasible, we propose a modular pipeline based on argument search (Ajjour et al., 2019) for automatically generating argumentation trees modeling the domain knowledge.

Given a domain, a set of features that are utilized in the XAI explanations, and a collection of document that contains the information for the domain knowledge, we automatically generate domain-specific argumentation trees for XAI dialogues by adapting the pipeline proposed by Rach et al. (2021) to the field of explanatory dialogue. Through an expert evaluation, we identify the best configuration of our pipeline. In addition, we evaluate our overall approach by manually generating explanatory dialogues according to the formal model by Madumal et al. (2019) with human- as well as

Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, pages 1–11
May 27–30, 2025.

automatically generated domain knowledge. A user study assessing the coherence of the generated dialogues, yields promising results for including the automatically generated tree structures into actual dialogue-based XAI systems. Additionally, we discuss the dependence of the results on the given collection of documents and the way the structured domain knowledge is utilized in the dialogue model.

The remainder of this work is organized as follows: Section 2 gives an overview over related work and Section 3 details our approach to the automatic generation of structured domain knowledge. After identifying the best performing configuration of our pipeline in Section 4, Section 5 evaluates our approach in a user study. We discuss our results in Section 6, before concluding in Section 7.

2 Related Work

Current dialogue-based XAI systems primarily function as question-and-answer (Q&A) systems that provide explicit verbalizations of the explanations generated by XAI methods (e. g., Slack et al. (2023); Feldhus et al. (2023)). While these systems are effective in providing direct insights, they lack the integration of additional domain-specific information, which has the potential to enhance the context and relevance of the explanations.

Incorporating domain-specific information into XAI itself is not a new idea. Pesquita (2021) demonstrated how knowledge graphs derived from ontologies can be utilized to create semantic explanations. Similarly, Bove et al. (2021) integrated domain-specific information into visual explanations, with annotations provided by domain experts. These approaches illustrate the potential of leveraging domain knowledge to enhance the interpretability of AI systems.

While knowledge-based dialogue is a well-established research field encompassing a variety of approaches and applications (Flycht-Eriksson, 1999; Chen et al., 2017), the connection between such knowledge-based dialogue systems and XAI so far remains mostly unexplored. To the best of our knowledge, the only work exploring this connection is Feustel et al. (2024). They employ bipolar argumentation trees within a dialogue-based XAI system to provide access to domain knowledge during conversational exchange. Their pilot study shows that incorporating domain knowledge not only improves the overall dialogue experience but also enriches the accessibility and utility of the explanations within the system. Since they created the structured domain knowledge through manual annotation, their system can benefit from the herein presented work.

3 From Documents to Structured Domain Knowledge

The pipeline by Rach et al. (2021), in the following referred to as the *existing pipeline*, offers a solution to automatically generating topic-specific argumentation trees for persuasive, argumentative dialogues. There, per dialogue, a single argumentation tree is created where all arguments are having a positive or negative stance towards the topic of the dialogue. To allow the explanatory dialogue system to link feature-based XAI explanations to the respective domain knowledge in an argumentative manner, multiple argumentation trees per XAI feature are required (Feustel et al., 2024), where each tree is entailing arguments for a different feature-outcome relation. Since we need to create multiple argumentation trees per XAI feature and not a single tree for the domain, the existing pipeline cannot be applied to our scenario directly. Nevertheless, being successfully evaluated in an argumentative dialogue context, the existing pipeline constitutes a promising basis for our work. The procedure of the existing pipeline is as follows: After utilizing argument search (Ajjour et al., 2019) to retrieve arguments along with their stance towards the topic from a web crawl, the arguments are optionally getting pre-grouped, before performing argumentative relation classification and determining the type of the relations between the arguments through stance propagation. Thereby, the argumentative relation classification entails the tasks of predicting the probability for a relationship between the arguments and then creating a tree structure out of these probabilities.

In the following, we first define the target structure, i. e., we describe how the domain knowledge is structured when modeled through argumentation trees. Afterwards, we detail our pipeline for the automatic generation of this structured domain knowledge. An overview over the processing steps of our pipeline is provided in Figure 1.

3.1 Target Structure

When modeling the domain knowledge of a dialogue-based XAI system with bipolar argumentation structures (Stab and Gurevych, 2014), the

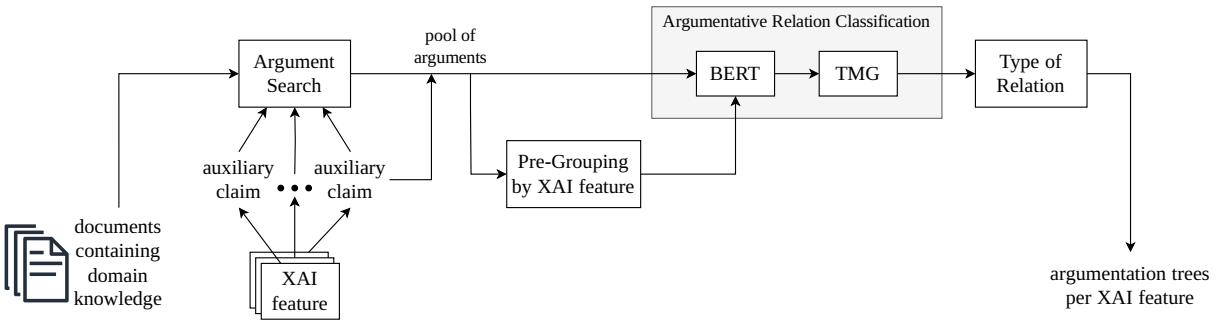

Figure 1: Pipeline for automatic generation of structured domain knowledge for dialogue-based XAI systems.

nodes of the structure represent the arguments, which function as the domain knowledge, and the directed edges between them indicate a supporting or attacking relationship. Throughout this work, an argument is a sentence that can target, i. e., support or attack, exactly one other argument, resulting in a tree structure (Stab and Gurevych, 2014). Following Feustel et al. (2024), we aim for at least one argumentation tree per feature of the explanation of the XAI system, where each root represents a feature-outcome relation. To not lose the relationship between the arguments representing a feature-outcome relation and the XAI explanation of the system, we introduce an auxiliary claim per feature stating that the respective feature is relevant for the domain. These auxiliary claims group together all argumentation trees that are addressing the respective feature. Therefore, we not only create multiple argumentation trees but also need to detect the XAI feature that they are addressing. An example of the targeted structure for a single feature of a domain is depicted in Figure 2.

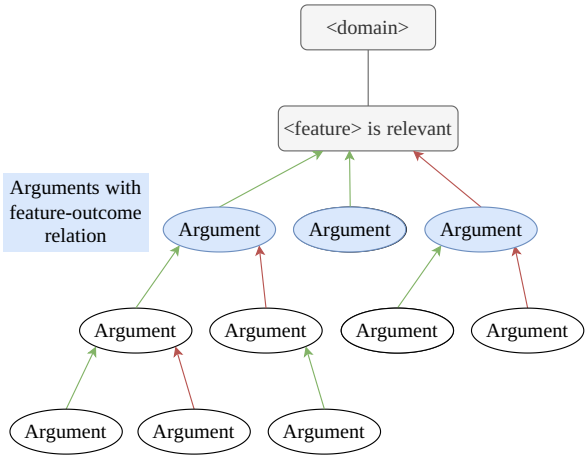

Figure 2: Exemplary depiction of the argumentation trees for a single feature within a domain. Every argument with a feature-outcome relation is the root of an individual argumentation tree. Green arrows indicate supporting relationships, while red arrows indicate attacking relationships.

3.2 Pipeline for Automatic Generation of Structured Domain Knowledge

Below, we describe the individual steps of our pipeline (see Figure 1) for the generation of structured domain knowledge for dialogue-based XAI systems. Namely, these are: argument acquisition through argument search, an optional pre-grouping of the arguments with respect to the features of the XAI system, argumentative relation classification transforming the pool of argumentative sentences into structured knowledge, and determining the type of the relationships between the arguments.

3.2.1 Argument Acquisition

By applying methods from the field of argument mining (Lawrence and Reed, 2019), argument search engines (Ajjour et al., 2019) allow to retrieve a ranked list of arguments with positive or negative stance towards a given search query. Based on an assessment of the suitability of different argument search engines for argumentative dialogues (Rach et al., 2020), the existing pipeline utilizes ArgumenText (Stab et al., 2018) with a web crawl as an argument search engine. Since our argumentation trees are representing domain knowledge which should be of high quality and may not be publicly accessible, we are utilizing the Classify API of summetix[1], which is the successor of ArgumenText, for the argument acquisition. Differently to the web crawl, we do not only input a query but also our own collection of relevant documents into the Classify API. Therewith, the documents that serve as a basis for the argument extraction are controllable and already tailored towards the targeted domain. Hence, instead of using the domain as

[1] https://www.summetix.com/

a query for the argument extraction, we can also utilize the auxiliary claims as queries. We then retrieve a collection of arguments per auxiliary claim which is equivalent to retrieving a collection of arguments per XAI feature. This additionally has the advantage that we are explicitly querying for arguments that address certain features and hence it might be more likely that the extracted arguments are representing the intended domain knowledge.

3.2.2 Pre-grouping of the Arguments

When arguments are pre-grouped, the existing pipeline restricts the allowed relationships between the arguments: Without pre-grouping, each argument can have a relationship to every other argument, whereas with pre-grouping, each argument can only have a relationship to arguments within the same group. Our target structure entails a grouping of the arguments according to the XAI features. While the grouping could be achieved implicitly through the argumentative relation classification itself, it might be desirable to a priori group the arguments by the feature they are addressing. To determine the XAI feature addressed by an argumentative sentence, we query the Classify API of summetix for every auxiliary claim and record the score for the sentence being an argument for the feature represented by the claim. We then assign the feature with the highest score to the sentence.

3.2.3 Argumentative Relation Classification

The core of the pipeline is the argumentative relation classification transforming the pool of arguments into the desired target structure.

To be able to apply the procedure of the existing pipeline without any major adaptions, we introduce an auxiliary node functioning as the root of our target structure and therefore can be viewed to represent the domain. This auxiliary node is having a relation with all auxiliary claims but with none of the retrieved arguments. This formalization of the problem allows to treat the process of creating multiple argumentation trees per feature as the process of creating one argumentation tree with the auxiliary node and the auxiliary claims being arranged in the tree in advance. Moreover, when not already determined through pre-grouping, the inclusion of the auxiliary claims into the pool of arguments allows to propagate the XAI feature of the auxiliary claim to the arguments targeting it. Therewith, it is also clear which feature is addressed by the individual argumentation trees.

Following, the existing pipeline, the confidence score of a pairwise BERT (Devlin et al., 2019) classification model is utilized to estimate the probability of a directed relationship between the ordered pairs of arguments. The model is fine-tuned with a balanced subset of the dataset by Carstens and Toni (2015) on predicting the labels *relation*, entailing supporting and attacking relations, and *no relation*. The type of the relation is identified post-hoc and is described in more detail in the next subsection. Given, the probability for a relationship, we apply their algorithm traversing and modifying graphs (TMG) (Schindler, 2020) to create the argumentation trees. To this end, TMG selects the most probable outgoing relationship for every argument and subsequently searches for circular graphs in the resulting structure, which are by default not attached to the argumentation tree with the domain as the root. In their formulation, which we name TMG_{all}, these circular graphs are connected to the argumentation tree by selecting the node with the most probable relation to any node outside the circular graph and change its outgoing relationship respectively. Due to the different nature of our target structure, we are adding an alternative variant of TMG, $\text{TMG}_{\text{feature}}$, which connects the circular graphs to the argumentation tree in a different way. More precisely, $\text{TMG}_{\text{feature}}$ is only considering the auxiliary claims and not any argument outside the circular graph as a potential target. This way, every circular graph becomes an individual argumentation tree for an XAI feature.

3.2.4 Determining the Type of Relationship

In the existing pipeline, the type of the relationships is determined by propagating the stance of the arguments towards the topic of the discussion through the argumentation tree. For the structured domain knowledge, we instead propagate the stance of the arguments towards the argument with the feature-outcome relation through the tree. The type of relation between the roots of each argumentation tree and the respective auxiliary claim is determined by the stance of the root towards the auxiliary claim. Simply propagating the stance towards the auxiliary claims through the trees is not sufficient since the dialogue-based XAI system will utilize every argumentation tree on its own and the representation of the auxiliary claims is to coarse-grained compared to a feature-outcome relation.

4 Configuration of the Pipeline through Expert Evaluation

The above description of our pipeline gives rise to the following possibilities for configuration: We can query the Classify API of the search engine either with the domain or the auxiliary claims, apply a pre-grouping of the arguments by XAI feature or not, and employ TMG_{all} or $\text{TMG}_{\text{feature}}$. For the auxiliary claims, we moreover test four different formulations

- *feature*: <feature> is relevant.
- $\textit{feature}_{\textit{inclDomain}}$: <feature> is relevant for <domain>.
- *feature+*: <feature>, which is related to <list of terms>, is relevant.
- $\textit{feature+}_{\textit{inclDomain}}$: <feature>, which is related to <list of terms>, is relevant for <domain>.

where <*feature*> is the respective name of the XAI feature, <*domain*> describes the domain (e. g., surviving the titanic), and <*list of terms*> are additional terms related to the feature. These additional terms are a combination of the related concepts, synonyms and types of the feature extracted from ConceptNet (Speer et al., 2017), the values of the feature if it is a categorical feature, and a further description of the feature if provided in the implementation of the dialogue-based XAI system, in our case in the one by Feustel et al. (2024).

To identify the best configuration of our pipeline, we perform an expert evaluation in the domain *titanic survival* with the XAI features *age*, *fare*, *gender*, and *passenger class* and make use of the same collection of documents utilized by Feustel et al. (2024) for creating the structured domain knowledge manually. After querying the argument search engine with the domain and all formulations of the auxiliary claims, we annotate the retrieved sentences to determine the best method for the retrieval and whether to include pre-grouping as a step in the pipeline. The decision on the variant of the TMG algorithm is driven by the comparison of the depth and width of the resulting argumentation trees. The expert evaluation is conducted by the authors of the paper. Since we are not performing a hypothesis test but merely identify the best configuration of our pipeline, we see no conflict of interest.

4.1 Annotation Study

For every sentence retrieved through the argument search, we perform an annotation regarding the following criteria:

- valid: Is the sentence an argument that can be used in a debate about <domain>?
- suitable: Is this argument suitable as a domain knowledge for dialogue-based XAI about <domain>? When the sentence is not valid it is also not suitable.
- feature(s): Which XAI feature is mainly addressed by the argument? If the argument addresses multiple XAI features and you cannot decide which is the main one, you may list the features. When none of the XAI features are addressed, state this as well.

The first two authors of the paper performed the annotation for 63 different sentences retrieved through the possible configurations of the argument search. They agreed in 100% of the cases for the criterion valid, in 84.13% of the cases for the criterion suitable, and again in 84.13% of the cases for assigning the exact same set of features in the criterion feature(s). To resolve the cases of disagreement, the third author of the paper was asked to perform the corresponding annotations, as well. Subsequently, we applied a majority vote for the criterion suitable and utilized the intersection of the assigned sets of features for the criterion feature(s). Through this procedure, a conclusive annotation could be created per sentence and criterion.

4.2 Results

Acquisition of Arguments The best performance for retrieving arguments was achieved by querying the Classify API with the auxiliary claims in the *feature* formulation. We excluded the formulations *feature+* and $\textit{feature+}_{\textit{inclDomain}}$ of the auxiliary claims from further analysis since with these we only retrieved six arguments and no arguments with the features *fare* and *gender*. When querying the API with the domain, only 89% of the 27 retrieved sentences are valid and from those only 92% are suitable. Moreover, we did not retrieve any arguments addressing the features *age* and *passenger class*. Utilizing the auxiliary claims in the formulation *feature* or $\textit{feature}_{\textit{inclDomain}}$, the pool of retrieved arguments has a size of 37 and 39 respectively, is valid to 97%, all valid arguments

are also suitable, and all XAI features are covered. By retrieving less argument that are not addressing any of the XAI features (5% vs. 10%), the *feature* formulation is performing better than the $feature_{inclDomain}$ formulation.

Pre-grouping and Argumentative Relation Classification Since the best auxiliary claim formulation for acquiring the arguments is *feature*, we run our pipeline with the arguments retrieved this way and also utilize this formulation of the auxiliary claims throughout the pipeline including the pre-grouping of the arguments. In this setup, we find pre-grouping outperforming the variant of our pipeline without pre-grouping and the $TMG_{feature}$ algorithm being better suited than TMG_{all}. When applying pre-grouping, 91% of the arguments annotated to be addressing an XAI feature are assigned to a correct feature, whereas without pre-grouping this is only the case for 37%. Comparing TMG_{all} and $TMG_{feature}$ both with pre-grouping, there are no differences in the generated argumentation trees in terms of maximum depth and the amount of trees for the features *age* and *fare*. For *gender* and *passenger class*, TMG_{all} generates a single argumentation tree per feature with a maximum depth of 6 compared to $TMG_{feature}$ which generates three trees with a maximum depth of 4 and four trees with a maximum depth of 2, respectively. The generation of a single argumentation tree by TMG_{all} leads to a restriction for the dialogue system: When the user asks, why the feature was relevant, the dialogue system has to select the only available feature-outcome relation and cannot adapt its response to the feature values input into the AI model and the user's needs. Following this line of reasoning, we identify $TMG_{feature}$ as the better variant.

5 User Study

With the following user study, we aim to assess the feasibility of our approach for automatically generating structured domain knowledge for dialogue-based XAI systems. To this end, we manually generate explanatory dialogues with human annotated and automatically generated domain knowledge and compare the coherence of the resulting dialogues. After presenting how structured domain knowledge in the form of argumentation trees can be utilized in an existing explanation dialogue game, we detail the study setup and present our results.

5.1 Generation of Explanatory Dialogues

The explanatory dialogues for our user study are created by manually applying the explanation dialogue game model by Madumal et al. (2019) to the respective structured domain knowledge. We create one dialogue per XAI feature with the two interlocutors *questioner* and *explainer*. The questioner, who needs an explanation, starts the interaction by asking why the respective XAI feature was relevant for the decision in the domain. The explainer now tries to explain why the XAI feature was having an influence. Therefore, the first move of the explainer is to select the best suited argument with a feature-outcome relation that is supporting the respective auxiliary claim "<feature> is relevant.".

Whenever, there is an attacking relationship for the argument presented by the explainer, the questioner will start the argumentation by presenting this counter-argument. Following the explanation dialogue game, the explainer then has to affirm the argument, optionally present an argument attacking the one by the questioner, and then continue the interaction by providing a further explanation. To ensure a better coverage of the domain knowledge, our explainer always presents an attacking argument if available. For continuing the explanation, the explainer traverses the argumentation tree backwards and selects a yet unused argument that is supporting one of its previous statements. The explainer is also allowed to select a new argument with feature-outcome relation if there are no more suitable arguments left within the current tree. When formulating the selected argument as a further explanation, the explainer has to make clear to which previous statement the argument is relating to.

When the questioner cannot attack the explanation by the explainer, it asks a return question like "Why?", "Can you elaborate on this?", or "Can you further explain why ... ?". The explainer then continues the explanation by selecting the best suited argument out of those that are supporting its previous statement. If there are no supporting arguments, we introduce an apologize move by the explainer into the dialogue game. To continue the dialogue, the explainer then proceeds as above by traversing the domain knowledge backwards.

The dialogue ends, when all argumentation trees of the respective XAI feature have been traversed. To create a smoother dialogue experience, we additionally prohibit the questioner from asking a return

Interlocutor	Dialogue Move	Utterance
Questioner	begin question	Why is the duration having an influence when determining whether one would have gotten a loan?
Explainer	further explain	The purpose of the loan can influence the acceptable loan duration.
Questioner	return question	Can you elaborate on this?
Explainer	further explain	Longer durations may be more appropriate for larger purchases like homes or education, while short-term loans are more suitable for smaller purchases or emergencies.
Questioner	affirm	Okay, I understand.
Questioner	return question	But are there any other reasons why the duration is having an influence?
Explainer	further explain	Yes. Shorter loan terms are generally less risky for lenders and therefore lead to a better acceptance rate.
Questioner	begin argument	But longer durations can allow borrowers to spread out repayments, making them more manageable.
Explainer	affirm argument	I get your point.
Explainer	counter argument	However, longer loan durations may increase the risk of unforeseen circumstances.
Explainer	further explain	And shorter loan terms are generally less risky for lenders and therefore lead to a better acceptance rate because shorter repayment periods reduce the overall exposure to potential default.
Questioner	begin argument	However, then lenders have less time to assess potential changes in the borrower's financial situation.
Questioner	begin question	Why is the duration having an influence when determining whether one would have gotten a loan?
Explainer	further explain	Lenders will usually feel more comfortable lending you money for a shorter period because you're more likely to be able to pay it back.
Questioner	return question	Why?
Explainer	further explain	A shorter loan term will also save you more money because you'll pay interest for fewer years.

Table 1: Human generated explanatory dialogues in the domain *credit acquisition* for the XAI feature *duration* with manually generated (upper part) and automatically generated (lower part) domain knowledge. The dialogue move is provided in accordance with the explanation dialogue game model by Madumal et al. (2019).

question when it has already performed a return question in its last two moves and the explainer cannot continue explaining without traversing the domain knowledge backwards. In these cases, the questioner then affirms the explanation and formulates a return question asking for further reasons why the XAI feature was having an influence.

The explanatory dialogues created for our study in the domain *credit acquisition* for the XAI feature *duration* are shown in Table 1.

5.2 Study Setup

We perform the user study within the domains *titanic survival* and *credit acquisition*. The manual creation of the argumentation trees follows the procedure by Feustel et al. (2024). For the automatically generated trees, we employ our pipeline in its previously determined best configuration, i. e., the auxiliary claims are formulated as "<feature> is relevant.", the auxiliary claims are used for retrieving and pre-grouping the arguments, and $\text{TMG}_{\text{feature}}$ is applied for obtaining the tree structure. As the documents for the domain knowledge, we utilize the first ten URLs that are processable by the Classify API of summetix and were retrieved by performing a Google Search[2] with the queries "factors for surviving the titanic" and "factors for acquiring a loan", respectively. For *titanic survival*, we acquired arguments for all four XAI features, namely *age*, *fare*, *gender*, and *passenger class*. For *credit acquisition*, we only consider the XAI features *checking account*, *duration*, and *savings* in our user study since the automatically selected collection of documents did not allow for extracting arguments addressing the XAI features *purpose* and *amount*. To keep the length of the generated explanatory dialogues feasible for the user study, we select the 10 arguments with the highest retrieval score per feature before starting the relation classification of our pipeline. Similarly we restrict the human generated domain knowledge to a maximum of 10 arguments per feature.

Following the evaluation of the existing pipeline (Rach et al., 2021), we asses the coherence (Venkatesh et al., 2018) of the generated dialogues by making use of the following categories with yes/no questions:

- comprehensible: Do you understand what the speaker wants to say?
- reference: Does the utterance address its reference?
- attitude: Does the attitude of the utterance fit the speaker's role?

In the user study, the web interface presents the generated dialogues utterance-wise and asks the participants for an answer to these questions whenever the

[2] https://google.com/

utterance entails an argumentative sentence from the structured domain knowledge. Before starting the study, a textual page explained the above categories in more detail and provided hand-crafted examples in the domain *acceptance as a tenant*. At the end of the study, the participants assessed how clearly they understood the instructions for each of the categories on a five-point Likert scale.

We asked five non-expert users (two females, three males) to take part in our study and presented every user with all of the 14 dialogues, i. e., four dialogues for *titanic survival* and three dialogues for *credit acquisition* and this one time with the human generated and the other time with the automatically generated domain knowledge.

5.3 Results

The assessment of the clarity of the instructions for the categories was rated by the five participants as shown in Table 2. While the categories *comprehensible* and *attitude* were totally clear to the majority of the study participants, understanding the category *reference* was more challenging. Therefore, to eliminate outliers and achieve a result that is as objective as possible, we follow Wachsmuth et al. (2017) by selecting the three most agreeing participants per category and gaining a final answer for each question through majority vote. The category-wise inter annotator agreement is assessed by Randolph's kappa (Randolph, 2005). For the three most agreeing participants, the agreement is substantial (0.78) for comprehensible, moderate (0.58) for reference, and almost perfect (0.89) for attitude (Landis and Koch, 1977), whereas the agreement for all five participants is 0.64, 0.32, and 0.47, respectively.

The dialogue-wise results of the user study for the three most agreeing participants are shown in Table 3. We report the ratio of positive and overall ratings and perform a Boschloo exact test (Boschloo, 1970) to assess the statistical difference between the automatically and manually generated domain knowledge. Following Rach et al. (2021), an utterance is regarded to be coherent, when all of the three categories are rated positively, i. e., with "yes", in the result. For the human-generated domain knowledge, we can see that all categories were rated positively for all utterance besides the reference category for the feature *duration* in the *credit acquisition* domain. With the automatically generated tree structures, no errors in terms of attitude were identified and the percentage of comprehensible argumentative utterances is above 90% for both domains. Moreover, there is no significant difference between the human and automatically generated argumentation trees for the categories attitude and comprehensible. For reference and coherence, however, we observe a statistically significant difference between the manually and automatically generated domain knowledge.

	totally agree	agree	neutral	disagree	totally disagree
comprehensible	3	2	–	–	–
reference	–	3	1	1	–
attitude	4	–	1	–	–

Table 2: Amount of responses on a five-point Likert scale for how clearly the participants have understood the instructions.

6 Discussion

To close the gap between the human and our automatically generated argumentation trees for domain-specific knowledge, our results suggest that only an improvement of the references made between the arguments is required. This room for improvement might be attributed to the following areas: First, our pipeline could be identifying the relationships between the arguments in a non-suited way for domain knowledge. This could be improved by fine-tuning the pairwise BERT model on a dataset that is tailored more towards the modeling of domain knowledge or by further adapting or even exchanging the process of creating the final argumentation trees through TMG. Second, our instantiation of the explanatory dialogue game model could have contributed to the results. While we have utilized the same strategy for generating the dialogues, the underlying argumentation trees are having different characteristics: The human generated domain knowledge shows an almost equal amount of supporting and attacking relations, whereas the automatically generated ones are consisting nearly only of supporting relations. Therefore, with the automatically generated trees, the chains of reasoning within the dialogues became increasingly larger and the interaction between the questioner and the explainer was also more single-sided. This potential cause is also underpinned by a comment from one of the study participants: "When the answer of the explainer [...] didn't really fit the question asked but still fit the topic of the conversation I was a bit unsure if [I should an-

	titanic age	titanic fare	titanic gender	titanic passenger class	titanic overall	credit checking account	credit duration	credit savings	credit overall
comprehensible	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
reference	1.00	1.00	1.00	1.00	1.00	1.00	0.86	1.00	0.95
attitude	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
coherence	1.00	1.00	1.00	1.00	1.00	1.00	0.86	1.00	0.95
comprehensible	0.86	1.00	1.00	1.00	0.96	0.75	1.00	1.00	0.93
reference	1.00	0.6	0.67	0.83	0.78	0.75	1.00	0.5	0.64
attitude	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
coherence	0.86	0.6	0.67	0.83	0.74	0.5	1.00	0.5	0.57
comprehensible					0.86				0.55
reference					**0.02**				**0.03**
attitude					1.00				1.00
coherence					**0.01**				**0.01**

Table 3: Feature-wise and overall results per domain for manually (upper part) and automatically (mid part) generated domain knowledge. We report the ratio of positive and overall ratings. Additionally, we report the p-values of the pairwise Boschloo exact test comparing automatically and manually generated domain knowledge (lower part).

swer with yes or no for the category reference]." This comment directly leads us to another aspect, namely the difficulty of assessing the category reference for the participants. While it might be in general difficult to asses this category, an improved formulation and explanation for the category could improve results in future works. Finally, the underlying data and therewith the documents utilized for extracting the domain knowledge might play a role. We utilized the top results of a web search engine without checking the content of the documents and their suitability for extracting domain knowledge. Hence, the argumentation trees created through our pipeline might also have a general disadvantage compared to the human-generated ones in terms of the available data.

While we evaluated our pipeline in the domains of *titanic survival* and *credit acquisition*, it can be applied to any domain and feature-based XAI system as long as reliable documents containing the required domain knowledge are available.

7 Conclusion and Future Work

We have presented an approach to automatically generate structured domain knowledge for dialogue-based XAI systems. To this end, we adapted an existing pipeline (Rach et al., 2021) from the field of persuasive, argumentative dialogue to the field of explanatory dialogue. Our approach combines methods from formal argumentation with data-driven techniques to ensure a flexible, yet reliable knowledge base. Through an expert evaluation, we identified the best configuration of our pipeline. Utilizing this configuration in a user study, we compare the automatically generated argumentation trees to human-generated ones by assessing the coherence of manually generated explanatory dialogues including the respective trees as domain knowledge. The study concludes that the human-generated argumentation trees are performing better than the automatically generated ones since the reference of the arguments leaves room for improvement. However, we discussed that this might be attributed to the instantiation of the employed explanatory dialogue game and the documents utilized for extracting the domain knowledge.

Therefore, besides improving the argumentative relation classification of the pipeline itself, a task for future work could be the optimization of the selection of the documents entailing the domain knowledge when not provided with these documents by a human. Additionally, the pipeline could become more robust by including validations based on established methods from the field of computational argumentation. Last but not least, a more large scale user study evaluating the automatically generated argumentation trees in an actual interaction with a dialogue system providing contextualized XAI explanations would provide further valuable insights.

Acknowledgments

We thank summetix GmbH for supporting our research with access to their Classify API.

References

Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160.

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data acquisition for argument search: The args.me corpus. In *KI 2019: Advances in Artificial Intelligence - 42nd German Conference on AI, Kassel, Germany, September 23-26, 2019, Proceedings*, volume 11793 of *Lecture Notes in Computer Science*, pages 48–59. Springer.

R. D. Boschloo. 1970. Raised conditional level of significance for the 2 × 2-table when testing the equality of two probabilities. *Statistica Neerlandica*, 24(1):1–9.

Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2021. Contextualising local explanations for non-expert users: an XAI pricing interface for insurance. In *Joint Proceedings of the ACM IUI 2021 Workshops co-located with 26th ACM Conference on Intelligent User Interfaces (ACM IUI 2021), College Station, United States, April 13-17, 2021*, volume 2903 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Lucas Carstens and Francesca Toni. 2015. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO. Association for Computational Linguistics.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor.*, 19(2):25–35.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz, and Sebastian Möller. 2023. Interrolang: Exploring NLP models and datasets through dialogue-based explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 5399–5421. Association for Computational Linguistics.

Isabel Feustel, Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2024. Enhancing model transparency: A dialogue system approach to XAI with domain knowledge. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 248–258, Kyoto, Japan. Association for Computational Linguistics.

Annika Flycht-Eriksson. 1999. A survey of knowledge sources in dialogue systems. *Electron. Trans. Artif. Intell.*, 3(D):5–32.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A grounded interaction protocol for explainable artificial intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, pages 1033–1041. International Foundation for Autonomous Agents and Multiagent Systems.

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38.

Catia Pesquita. 2021. Towards semantic integration for explainable artificial intelligence in the biomedical domain. In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2021, Volume 5: HEALTHINF, Online Streaming, February 11-13, 2021*, pages 747–753. SCITEPRESS.

Niklas Rach, Yuki Matsuda, Johannes Daxenberger, Stefan Ultes, Keiichi Yasumoto, and Wolfgang Minker. 2020. Evaluation of argument search approaches in the context of argumentative dialogue systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 513–522, Marseille, France. European Language Resources Association.

Niklas Rach, Carolin Schindler, Isabel Feustel, Johannes Daxenberger, Wolfgang Minker, and Stefan Ultes. 2021. From argument search to argumentative dialogue: A topic-independent approach to argument acquisition for dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 368–379, Singapore and Online. Association for Computational Linguistics.

Justus J. Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa.

Carolin Schindler. 2020. Argumentative relation classification for argumentative dialogue systems. Bachelor's thesis, Institute of Communications Engineering, Ulm University.

Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining machine learning models with interactive natural language conversations using talktomodel. *Nat. Mac. Intell.*, 5(8):873–883.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 21–25, New Orleans, Louisiana. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju. 2018. On evaluating and comparing conversational agents. *CoRR*, abs/1801.03625.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. 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, Valencia, Spain. Association for Computational Linguistics.