Multi-Class versus Means-End: Assessing Classification Approaches for Argument Patterns

Maximilian Heinrich
Bauhaus-Universität Weimar

Khalid Al-Khatib
University of Groningen

Benno Stein
Bauhaus-Universität Weimar

Abstract

In the study of argumentation, the schemes introduced by Walton et al. (2008) represent a significant advancement in understanding and analyzing the structure and function of arguments. Walton's framework is particularly valuable for computational reasoning, as it facilitates the identification of argument patterns and the reconstruction of enthymemes. Despite its practical utility, automatically identifying these schemes remains a challenging problem. To aid human annotators, Visser et al. (2021) developed a decision tree for scheme classification. Building on this foundation, we propose a means-end approach to argument scheme classification that systematically leverages expert knowledge—encoded in a decision tree—to guide language models through a complex classification task. We assess the effectiveness of the means-end approach by conducting a comprehensive comparison with a standard multi-class approach across two datasets, applying both prompting and supervised learning methods to each approach. Our results indicate that the means-end approach, when combined with supervised learning, achieves scores only slightly lower than those of the multi-class classification approach. At the same time, the means-end approach enhances explainability by identifying the specific steps in the decision tree that pose the greatest challenges for each scheme—offering valuable insights for refining the overall means-end classification process.

1 Introduction

Argumentation is a crucial process in shaping our understanding of the world and fostering critical thinking. It plays a vital role in a range of contexts, including debate, decision-making, and the process of informing or changing beliefs. To classify common patterns of argumentation, the schemes developed by Walton et al. (2008) are of particular interest, as these schemes are extremely versatile and allow for a range of use cases. They can identify reasoning patterns within specific domains, such as legal reasoning (Verheij, 2003), help in the selection of argumentation strategies (Wachsmuth et al., 2018), and also uncover patterns in reasoning synthesis applications (Baff et al., 2019). In addition, the schemes can be used to reconstruct missing parts of arguments (Feng and Hirst, 2011), to train argumentation skills, or to enhance existing debate systems such as those described by Rakshit et al. (2017); Le et al. (2018); Slonim et al. (2021). Table 1 illustrates an example of such a scheme, namely the 'Cause to Effect' scheme.[1] Due to their fine nuances, the classification of Walton schemes is very challenging, even for people with a background in linguistics (Macagno et al., 2017). In addition, in a real life argumentation scenario, many parts of the schemes are only hinted and not explicitly mentioned (Dumani et al., 2021). To help people classify arguments based on Walton schemes, Visser et al. (2021) has developed the Argument Scheme Key (ASK) - a decision tree that guides users step by step through the annotation process. This raises the question of whether such a decision tree approach could be applied to language models to improve argument classification.

In this paper, we explore how the ASK decision tree can enhance the effectiveness of argument scheme classification. We refer to this approach as means-end classification. Rather than requiring the model to perform the complex task of scheme detection in a single step, the means-end approach decomposes the process into a guided sequence of simpler subtasks. At each stage, the model executes a straightforward task, such as identifying the presence of a specific argument property. In addition to potentially improving classification scores, this approach also boosts explainability: it enables the analysis of each decision made during the pro-

[1] In Walton's compendium (Walton et al., 2008), scheme names frequently begin with the prefix 'Argument from'. For brevity, we omit this prefix throughout this work.

Proceedings of the 12th Argument Mining Workshop, pages 195–204

Cause to Effect	
Definition	
Premise 1	Generally, if A occurs, then B will (might) occur.
Premise 2	In this case, A occurs (might occur).
Conclusion	Therefore, in this case, B will (might) occur.
Examples	
Dataset	EthiX
Example 1	If entropy leads to universal randomness and randomness is the lack of all deterministic forces, then at least one part of a wholly deterministic universe is false, meaning there must be something else influencing the universe outside of determinism.
Example 2	The sensations felt when consuming marijuana and alcohol are very different. As such, they are not interchangeable, meaning that people may use both.
Dataset	USTV
Example 1	USA is in deep trouble. These countries, especially China, are giving incentives.
Example 2	NRA is protecting the Second Amendment. NRA are very, very good people. TRUMP is very proud of the endorsement of the NRA.

Table 1: Definition of the Cause to Effect' scheme, accompanied by examples from the EthiX (Bezou-Vrakatseli et al., 2024) and USTV (Visser et al., 2021) datasets. The definition used follows the version in Bezou-Vrakatseli et al., which slightly adapts the original formulation by Walton et al. (2008). As shown, most of the arguments are enthymemes, lacking a direct correspondence to the scheme definitions, and the two datasets display distinct styles of argumentation.

cess, making it possible to identify where and why the model's classification succeeds or fails. The contributions of this paper are:

(1) We conduct a comprehensive evaluation of argument scheme classification by comparing the traditional multi-class classification approach with the means-end approach on two separate datasets. Each approach is evaluated using both prompting-based and supervised learning methods. Our results highlight the key strengths and weaknesses of each approach and offer insights into how the means-end approach can be effectively applied to scheme classification tasks.

(2) We assess the effectiveness of ASK decision tree nodes using both prompting-based and supervised learning models. This novel analysis yields valuable insights into the utility—and limitations—of individual nodes in argument scheme classification. It also reveals which schemes can be reliably identified and to what extent. These findings offer a deeper understanding of the classification process, surpassing the explanatory power of traditional multi-class classification approaches.

2 Related Work

This section provides an overview of the diverse applications and methodological approaches to argument schemes within computational argumentation. We explore the classification and analysis of these schemes, their incorporation into datasets, and the challenges in their annotation and automated generation. The concept of argument schemes suggests that arguments can be organized based on diverse characteristics, reflecting commonly used patterns of argumentative reasoning (Macagno and Walton, 2015). This idea has ancient roots, tracing back to the works of Aristotle, as discussed in (Macagno et al., 2017). One of the most debated issues in this context is how such schemes should be appropriately classified, leading to the development of multiple approaches. The dialectical approach, highlighted by van Eemeren and Grootendorst (2003), focuses on the abstract representation of arguments within debates, while Wagemans (2016) organizes arguments into three main distinctions, culminating in the periodic table of arguments. Other approaches, such as those by Kienpointner (1992) and Grennan (1997), aimed to identify common argumentative features. In this tradition, Walton schemes are empirically developed in a bottom-up manner (Walton, 1996; Walton et al., 2008), involving the selection and analysis of arguments from varied domains. This method has led to the documentation of over 60 primary schemes and more than 100 sub-schemes (Walton et al., 2008). An initial approach to grouping schemes together was

ID	Which option applies to the argument?
ID-17	A: Conclusion is about a course of action B: Conclusion is not specifically action-oriented
ID-32	A: Argument explicitly mentions values B: Argument is not specifically value-based
ID-47	A: Argument relies on a causal relation B: Argument does not specifically rely on causality

Table 2: Dichotomous questions from the ASK decision tree.

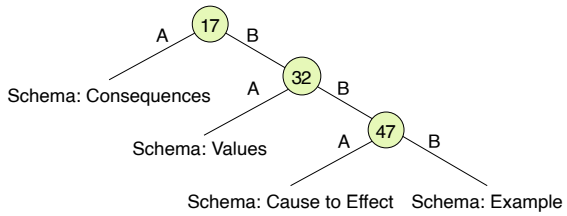

Figure 1: Trimmed ASK decision tree for classifying four argumentation schemes. Classification starts at the root node (ID-17), where the user selects the option that best matches the argument under analysis. Each response directs the user to the next relevant question, guiding them through the tree until the correct scheme is identified. The options for each node are listed in Table 2. The ID number assigned to each node corresponds to the original node IDs in the ASK decision tree from Visser et al. (2021). To correctly identify a scheme as 'Cause to Effect' (see Table 1 for examples), one must answer node ID-17 with 'B', node ID-32 with 'B', and node ID-47 with 'A'.

first proposed by Walton et al. (2008) and later refined by Walton and Macagno (2015).

Various datasets have been created to support research on Walton schemes. The Araucaria dataset, for instance, includes arguments from various media and institutional sources (Katzav et al., 2004; Reed, 2006; Moens et al., 2007). The dataset by Visser et al. captures the dynamics of the 2016 presidential debates (Visser et al., 2021), while the ReCAP dataset focuses on German education policy (Dumani et al., 2021). Further, datasets like those curated by Macagno (2022) and Bezou-Vrakatseli et al. (2024) expand the scope to argumentative tweets and ethical debates, respectively. A significant challenge in this field is achieving high levels of annotator agreement. Studies such as those by Lindahl et al. (2019) have revealed inconsistencies in annotation, underscoring the need for clearer guidelines. The use of decision trees for scheme annotation has been shown to significantly improve annotator agreement (Visser et al., 2021; Macagno, 2015, 2022).

Walton schemes are used to analyze various areas, such as newspapers (Lindahl et al., 2019), elections (Hansen and Walton, 2013) or student work (Duschl, 2007). They are also used to analyze paralogisms in student work (Rapanta and Walton, 2016a). Automated generation of arguments aligned with specific schemes has been explored in works like those by Saha and Srihari (2023) and the NLAS-multi corpus (Ruiz-Dolz et al., 2024a), showcasing the potential for synthetic argument generation. Several methods have been developed to classify schemes in texts (Bezou-Vrakatseli et al., 2024). Feng and Hirst (2011) analyze the five most common arguments from the Araucaria dataset to construct decision trees based on argumentative structural and linguistic features. The approach of Moens et al. (2007) leverages the same dataset to detect arguments using a multinomial naive Bayes classifier and a maximum entropy model. Song et al. (2014) develop protocols for annotating Walton schemes and their associated critical questions. Furthermore, Bezou-Vrakatseli et al. (2024) utilizes a range of BERT-based classifiers for automated scheme classification, while Lawrence and Reed (2016) leverages argumentation schemes to identify argumentative structures. Similarly, Green (2018) utilizes logic programs and schemes to mine arguments in biomedical research articles, building on earlier work (Green, 2015). Walton schemes are utilized across various domains to analyze content from newspapers, election campaigns, and educational settings, highlighting their adaptability and relevance in real-world applications (Lindahl et al., 2019; Hansen and Walton, 2013; Duschl, 2007; Rapanta and Walton, 2016a).

3 Multi-Class and Means-End Approaches

Humans often struggle with annotating argumentation schemes, partly because many schemes rely on implicit assumptions (enthymemes), and some schemes such as consequences require multiple steps of reasoning (Macagno and Walton, 2015). To simplify the annotation process for schemes, Visser et al. (2021) developed a binary decision tree that systematically guides annotators through the annotation task. Rather than classifying the argument directly, annotators make a series of choices between two characteristics of the argument at each step. For instance, one choice might involve determining whether the conclusion of the argument is about a course of action. Each decision narrows

down the classification path by determining the next characteristic to be evaluated. An illustration of this classification procedure is provided in Figure 1.

Instead of relying on a human annotator, a language model can follow the steps outlined in the decision tree—a method we refer to as the means-end approach for argument classification. In contrast, traditional multi-class classification presents the model with an argument—optionally enriched with contextual information—and requires it to select the appropriate argumentation scheme from a predefined set. The means-end approach, by contrast, decomposes the classification task into a sequence of smaller, more manageable decisions. At each step, the model identifies a specific characteristic of the argument, which then determines the subsequent step in the classification sequence. This procedure is guided by external expert knowledge structures, such as decision trees. The approach is inspired by the *means-end analysis* problem-solving technique, in which an agent incrementally selects and applies actions to achieve a goal, based on an information gain heuristic (Newell and Simon, 1995). Note that this approach is different from merely breaking a problem into smaller steps; it also encodes a specific sequence for how the classification should be performed. The correctness of this sequence is ensured by the expert knowledge employed.

A key advantage of the employed step-wise decomposition is enhanced explainability: unlike the multi-class approach, which often operates as a black box, the means-end method allows for detailed analysis of each individual decision. This not only helps identify sources of classification errors but also makes it easier to refine specific steps within the classification process. Moreover, this approach is not limited to the domain of argumentation and can be applied to other complex classification tasks that benefit from codified expert knowledge.

4 Experiments and Evaluation

Our experiments are designed to address two primary objectives. First, we evaluate whether the means-end approach—guided by the ASK decision tree—offers improved results over the traditional multi-class classification approach for argument scheme classification. Second, we conduct a detailed analysis of the means-end approach to determine which decision points are most effective and where the classification process is most susceptible to errors, all while providing a high level of explainability.[2]

4.1 Dataset and Decision Tree

For argument scheme classification, we utilize two datasets: EthiX (Bezou-Vrakatseli et al., 2024) and the US2016G1tvWALTON dataset (referred to as USTV) (Visser et al., 2021). The EthiX dataset consists of 686 arguments extracted from ethical debates on Kialo[3], spanning 22 topics and covering eight distinct argumentation schemes. The USTV dataset includes 505 arguments in total, spanning 38 argumentation schemes. Its content is sourced from the first head-to-head debate of the 2016 U.S. general election and was transformed into the Argument Interchange Format (Chesñevar et al., 2006). A key advantage of these two datasets is that the human annotators applied the same ASK decision tree logic from Visser et al. (2021) that we utilize for the means-end classification approach. To facilitate the classification process and ensure a sufficient amount of training data, we focus on four schemes that are included in both the EthiX and USTV datasets. For the means-end approach, we simplify the original ASK decision tree by retaining only the three nodes necessary to differentiate between the four considered schemes. This refinement removes questions related to not-considered schemes while preserving the consistency and integrity of the remaining ones. Although the original annotators had to answer a greater number of questions, those included in the reduced tree are answered identically to the original process, allowing for a meaningful comparison between human and machine judgment. Figure 1 presents the modified decision tree, used in our means-end experiments. A summary of the refined dataset, along with key statistics, is provided in Table 3. We split the datasets in a 70/10/20 ratio for training, validation and testing, respectively. Minor adjustments were made to ensure that each scheme was represented in every split. For the EthiX dataset specifically, we ensured that every combination of scheme and topic appeared in each split. Additionally, we made sure that the test set for each dataset contained at least 11 distinct arguments for each scheme.

[2]All our code is available at: https://github.com/webis-de/Argmining-25

[3]https://www.kialo.com/

Argument Schemes			Datasets					
			EthiX		USTV		Σ	
Name	Walton-ID	DT-Depth	#	%	#	%	#	%
Example	6	3	120	24.0	81	44.0	201	29.4
Values	19	2	118	23.6	15	8.2	133	19.5
Cause to Effect	28	3	87	17.4	48	26.1	135	19.8
Consequences	33	1	174	34.9	40	21.7	214	31.3
Σ			499		184		683	

Table 3: Overview of the four schemes and their frequency in the EthiX (Bezou-Vrakatseli et al., 2024) and USTV (Visser et al., 2021) datasets. The second column, labeled 'Walton-ID', shows the canonical scheme numbers as defined by Walton et al. (2008). The 'DT-depth' column (Decision Tree Depth) indicates the number of decisions required in the trimmed ASK decision tree to correctly identify each scheme (see Figure 1).

	Multi-Class				Means-End			
	EthiX		USTV		EthiX		USTV	
	PR	SV	PR	SV	PR	SV	PR	SV
Macro F_1	0.63	0.72	0.44	0.44	0.44	0.68	0.33	0.38
Micro F_1	0.65	0.72	0.48	0.50	0.45	0.68	0.35	0.45

Table 4: Macro and Micro F_1 scores are reported for multi-class and means-end approaches using two classification methods: few-shot prompting (PR) with GPT-4o-mini and a supervised training approach (SV) with BERT. Results are presented for the EthiX and USTV datasets.

4.2 Experiments Overview

For each classification approach, we employ two distinct methods. The first is prompting, which leverages a large language model—specifically, GPT-4o-mini (2024-07-18) (OpenAI, 2023). Prompting enables us to provide the model with the same natural language instructions used by human annotators, making it particularly suitable for executing decision trees designed for human reasoning. The second method is supervised learning, in which we fine-tune a conventional BERT-based classifier (Devlin et al., 2019) on the training data. These two methods also incorporate different model architectures. GPT-4o-mini processes text unidirectionally, from left to right, predicting each token based solely on the preceding tokens. In contrast, BERT's bidirectional architecture allows it to consider both the preceding and following context around every token simultaneously, enabling a holistic understanding of the text.

In the multi-class approach using prompting, the model receives an argument along with a list of argumentation schemes and is tasked with selecting the most appropriate scheme. The prompt also includes definitions of all the schemes (for an example, see Table 1), adapted from Bezou-Vrakatseli et al. (2024) and based on the original formulations in Walton et al. (2008). Similarly, the means-end approach combined with prompting provides the model with an argument paired with a characterization derived from the ASK decision tree, where the model's task is to determine which characterization best applies to the argument. In all prompting-based methods, we employ a few-shot learning strategy by including example instances. To minimize randomness and encourage precise, controlled outputs, we set the temperature to 0.2 and the top-p value to 0.1 across all tasks. In the multi-class approach using supervised learning, a single classifier is trained to differentiate among the four argumentation schemes. In contrast, the supervised learning means-end approach trains a separate binary classifier for each of the three nodes in the decision tree. As a result, nodes deeper in the tree receive fewer training examples, since each node—except the root—handles only a subset of the full set of arguments. Table 4 presents the macro and micro F_1 scores for both the multi-class and means-end classification approaches. The detailed results for the multi-class classification methods are shown in Table 5. Table 6 reports the scores for the means-end approach, along with the accuracy of the corresponding decision tree nodes. To ensure consistent evaluation across argument schemes, classification approaches, and datasets, we sample 10 arguments per scheme from each dataset. The same set of arguments is used across all experiments to compute the reported scores.

5 Discussion

Classifying arguments remains a particularly challenging task, as reflected in our results. First, we

	Multi-Class											
	EthiX						USTV					
	Prompting			Supervised			Prompting			Supervised		
Scheme	Pre.	Rec.	F_1	Pre.	Rec.	F_1	Pre.	Rec.	F_1	Pre.	Rec.	F_1
Example	1.00	0.40	0.57	0.75	0.60	0.67	0.33	0.10	0.15	0.67	0.60	0.63
Values	0.64	0.70	0.67	0.73	0.80	0.76	0.73	0.80	0.76	0.0	0.0	0.0
Cause to Effect	0.62	0.50	0.56	0.78	0.70	0.74	0.40	0.40	0.40	0.36	0.80	0.50
Consequences	0.59	1.00	0.74	0.67	0.80	0.73	0.38	0.60	0.46	0.67	0.60	0.63

Table 5: Overview of multi-class classification results for Precision ('Pre.'), Recall ('Rec.'), and F_1 on the EthiX and USTV datasets. 'Prompting' refers to the few-shot approach using the GPT-4o-mini model, while 'Supervised' denotes the fine-tuned BERT-based classifier.

		Means-End											
		Prompting						Supervised					
		DT-Nodes			Scheme classification			DT-Nodes			Scheme classification		
		ID-17	ID-32	ID-47				ID-17	ID-32	ID-47			
Dataset	**Scheme**	Acc.	Acc.	Acc.	Pre.	Rec.	F_1	Acc	Acc	Acc	Pre.	Rec.	F_1
EthiX	Example	0.80	0.80	0.70	0.43	0.60	0.50	0.90	0.90	0.60	0.55	0.60	0.57
	Values	0.80	0.60		0.44	0.40	0.42	0.90	0.70		0.70	0.70	0.70
	Cause to Effect	1.00	0.80	0.60	0.50	0.50	0.50	1.00	0.90	0.70	0.67	0.60	0.63
	Consequences	0.30			0.43	0.30	0.35	0.80			0.80	0.80	0.80
USTV	Example	0.60	0.60	0.60	0.22	0.20	0.21	1.00	1.00	0.90	0.36	0.90	0.51
	Values	0.40	0.50		0.33	0.20	0.25	1.00	0.0		0.0	0.0	0.0
	Cause to Effect	0.90	0.80	0.50	0.50	0.30	0.38	1.00	1.00	0.60	0.50	0.60	0.55
	Consequences	0.70			0.37	0.70	0.48	0.30			1.00	0.30	0.46

Table 6: Overview of two classification method for the Means-End approach. 'Prompting' refers to the few-shot prompting method using the GPT-4o-mini model, while 'Supervised' denotes the fine-tuned BERT-based classifier. 'DT-Nodes' represents the nodes in the ASK decision tree that an argument must pass through to be correctly classified. The node IDs correspond to those listed in Table 2. The Accuracy ('Acc.') columns indicate the proportion of the 10 arguments per scheme that were correctly identified at the respective decision nodes. Accuracy is computed by tracing each argument's correct path through the decision tree and recording the decision at each node. The Precision ('Pre.'), Recall ('Rec.'), and F_1 columns represent overall classification performance, with each argument's scheme determined by following the decision tree logic. The evaluation is conducted on the EthiX and USTV datasets.

observe that the supervised learning method consistently outperforms LLM prompting. The limitations of large language models in classification tasks stem from the nature of their pretraining, which often does not sufficiently prepare them for domain-specific or fine-grained distinctions without additional adaptation. It is unlikely that an LLM has encountered highly specialized tasks—such as argument scheme classification using a means-end approach—during its training, which limits its effectiveness in this context. In contrast, the supervised learning approach benefits from explicit fine-tuning on the relevant argument schemes and datasets, resulting in substantially improved scores. Classification scores on the EthiX dataset are consistently higher than those on the USTV dataset, regardless of the approach or methods used. This disparity can be attributed to the nature of the USTV arguments, which are especially difficult to interpret without a clear understanding of the specific speech context in which they were made. In particular, the notably weak scores of the supervised method on the 'Values' scheme in the USTV dataset can be attributed to the extremely limited number of training examples available for that category. In contrast, the prompting method achieves better results for this scheme, leveraging the extensive pre-training of large language mod-

els. However, due to the high complexity of the arguments in the datasets and the small sample size used for the comparison, these results should be interpreted with caution.

As shown in Table 4, a comparison of the multi-class and means-end approaches indicates that, despite comparable overall F_1 scores, the multi-class approach achieves marginally higher results. Nonetheless, the scores for the means-end approach remain solid, especially given the reduced amount of training data available for nodes deeper in the classification tree. Examining the scores for individual schemes reveals varying results. For the Ethix and 'Consequences' schemes, the supervised means-end approach achieves the highest F_1 score among all compared configurations (Table 6). In the same configuration, the 'Example' scheme produces the lowest F_1 score. A similar variation in scheme scores is observed in the multi-class approach (Table 5). This suggests that some argument schemes (e.g., 'Consequences') are easier to classify than others. A key challenge in classification arises from the nature of the arguments themselves: they are often highly enthymematic, containing implicit or omitted components. In contrast, arguments associated with certain schemes may be more explicit, leading to higher classification scores.

One of the key advantages of the means-end approach is its explainability, as illustrated in Table 6. Here, differences appear notably at the root node ID-17. For most schemes—except 'Consequences'—the prompting method classifies this node correctly. However, since this node is intended to distinguish 'Consequences' from other schemes, it is not an appropriate choice at this point. In contrast, the supervised learning method shows better accuracy in detecting 'Consequences' arguments. We also observe that node ID-47 consistently struggles to differentiate between the 'Example' and 'Cause to Effect' schemes across both prompting and supervised learning methods in both datasets. This kind of insight underscores a key advantage of the means-end approach: when specific decision points in the tree underperform, human experts can intervene to refine the relevant nodes, thereby enhancing the overall system (Visser et al., 2021). Additionally, the means-end approach offers flexibility by allowing adaptation to the granularity of the classification task. If the objective is to classify broader categories of argument schemes rather than individual ones, the decision tree can be truncated at a desired depth—for example, by omitting node ID-47. In doing so, the classification process can be adjusted dynamically without requiring further training.

6 Conclusion

There are several compelling reasons why automated classification of Walton schemes is valuable. First, an automated classifier enables large-scale analysis of argumentation patterns across diverse domains, such as legal reasoning, online debates, and news articles. Second, once a scheme is classified, it becomes possible to identify corresponding critical questions as provided by Walton et al. (2008), facilitating the detection of errors in argumentation. These critical questions can also serve as commonplace arguments (Bilu et al., 2019). Third, schemes support enthymeme reconstruction, the training of argumentation skills and critical thinking (Figueras and Agerri, 2024), and the enhancement of existing debate systems (Rapanta and Walton, 2016b). Reliable scheme identification poses a significant challenge for human annotators due to the high cognitive load involved (Bezou-Vrakatseli et al., 2024). Additionally, while multi-class classification proves more effective for scheme detection, the means-end approach delivers comparable results with only a slight decrease in scores. To this end, the means-end approach offers significant advantages by providing valuable insights into the classification process, highlighting potential sources of error, and clearly identifying which specific argument characteristic are inconsistently recognized. Our findings confirm that automatically detecting argument schemes continues to be a challenging task. Additionally, our results show that the supervised training approach leveraging BERT surpasses the prompting method in performance across both multi-class and means-end approaches.

For future work, several directions are promising. One avenue is to further fine-tune the decision tree nodes, particularly those deeper in the tree that have fewer training examples. In this context, supplementary datasets—including synthetically generated arguments—may prove valuable. Another promising direction is the exploration of alternative datasets that feature more formal argumentation (Saha and Srihari, 2023; Ruiz-Dolz et al., 2024a). Hybrid methods for argument scheme classification deserve further investigation. For example,

the vast knowledge contained in large language models might be used to create contextual information that enhances a fine-tuned classifier based on the means-end framework—thus effectively merging the advantages of prompting and supervised learning techniques. Additionally, alternative decision tree structures—such as those proposed by Macagno (2015) and Macagno (2022)—or other classification frameworks could be employed to further improve classification scores within the means-end approach.

7 Limitations

The effectiveness of the means-end approach depends on the quality of the underlying decision tree. For the approach to be practical, each node's task must be clearly defined, precise, and easily interpretable. This enables annotators or language models to make accurate decisions without relying on extensive prior knowledge. However, when tasks are overly complex or ambiguous, the overall effectiveness of the approach declines. As a result, designing an effective decision tree poses a significant challenge, even for experts.

Ideally, a well-optimized decision tree would position nodes that classify frequently used schemes closer to the root, reducing the expected external path length. However, the ASK decision tree is imbalanced. For example, identifying an argument as the frequently used 'Example' scheme (see Table 3) requires correctly answering three successive decisions. The more decisions that must be made, the higher the risk of misclassification. This structural imbalance is also evident in the original ASK tree presented by Visser et al. (2021).

Both datasets largely consist of enthymemes, containing implicit premises or conclusions. In the EthiX dataset, arguments are drawn from Kialo debates; however, the specific context—such as whether an argument supports or attacks another—is not explicitly provided. In the USTV dataset, arguments originate from a televised debate, where many points rely on prior context and earlier topics that are not directly present within the arguments themselves. In such cases, contextual understanding and enthymeme reconstruction are essential for accurate classification by both human annotators and language models. This absence of explicit context makes the classification task particularly challenging. A markedly improved outcome is observed when classifying arguments that strictly follow the semi-formal Walton scheme definitions, as demonstrated by Ruiz-Dolz et al. (2024b), with near perfect F_1 scores. Lastly, it should also be considered that, due to the limited available data, only 10 arguments could be tested per scheme and dataset, which restricts the generalizability of the results.

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BMBF) through the project "DIALOKIA: Überprüfung von LLM-generierter Argumentation mittels dialektischem Sprachmodell" (01IS24084A-B).

References

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. 2019. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*, pages 54–64. Association for Computational Linguistics.

Elfia Bezou-Vrakatseli, Oana Cocarascu, and Sanjay Modgil. 2024. Ethix: A dataset for argument scheme classification in ethical debates. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, pages 3628–3635. IOS Press.

Yonatan Bilu, Ariel Gera, Daniel Hershcovich, Benjamin Sznajder, Dan Lahav, Guy Moshkowich, Anael Malet, Assaf Gavron, and Noam Slonim. 2019. Argument invention from first principles. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1013–1026. Association for Computational Linguistics.

Carlos I. Chesñevar, Jarred McGinnis, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo R. Simari, Matthew South, Gerard Vreeswijk, and Steven Willmott. 2006. Towards an argument interchange format. *Knowl. Eng. Rev.*, 21(4):293–316.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*,

pages 4171–4186. Association for Computational Linguistics.

Lorik Dumani, Manuel Biertz, Alex Witry, Anna-Katharina Ludwig, Mirko Lenz, Stefan Ollinger, Ralph Bergmann, and Ralf Schenkel. 2021. The recap corpus: A corpus of complex argument graphs on german education politics. In *15th IEEE International Conference on Semantic Computing, ICSC 2021, Laguna Hills, CA, USA, January 27-29, 2021*, pages 248–255. IEEE.

Richard A. Duschl. 2007. Quality argumentation and epistemic criteria. In Sibel Erduran and María P. Jiménez-Aleixandre, editors, *Argumentation in Science Education*, volume 35, pages 159–175. Springer Netherlands, Dordrecht.

Vanessa W. Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 987–996. The Association for Computer Linguistics.

Blanca C. Figueras and Rodrigo Agerri. 2024. Critical questions generation: Motivation and challenges. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 105–116, Miami, FL, USA. Association for Computational Linguistics.

Nancy L. Green. 2015. Annotating evidence-based argumentation in biomedical text. In *2015 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2015, Washington, DC, USA, November 9-12, 2015*, pages 922–929. IEEE Computer Society.

Nancy L. Green. 2018. Towards mining scientific discourse using argumentation schemes. *Argument Comput.*, 9(2):121–135.

Wayne Grennan. 1997. *Informal Logic: Issues and Techniques*. McGill-Queen's University Press, Montreal; Buffalo.

Hans V. Hansen and Douglas N. Walton. 2013. Argument kinds and argument roles in the ontario provincial election, 2011. *Journal of Argumentation in Context*, 2(2):226–258.

Joel Katzav, Chris Reed, and Glenn Rowe. 2004. Argument research corpus. In *Proceedings of Practical Applications in Language and Computers (PALC 2003): 4th Biennial International Conference on Practical Applications in Language Corpora, 4–6 April 2003, Łódź*. Peter Lang.

Manfred Kienpointner. 1992. *Alltagslogik: Struktur und Funktion von Argumentationsmustern*. Number 126 in Problemata. Frommann-Holzboog, Stuttgart-Bad Cannstatt.

John Lawrence and Chris Reed. 2016. Argument mining using argumentation scheme structures. In *Computational Models of Argument - Proceedings of COMMA 2016, Potsdam, Germany, 12-16 September, 2016*, volume 287 of *Frontiers in Artificial Intelligence and Applications*, pages 379–390. IOS Press.

Dieu-Thu Le, Cam-Tu Nguyen, and Kim A. Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 121–130. Association for Computational Linguistics.

Anna Lindahl, Lars Borin, and Jacobo Rouces. 2019. Towards assessing argumentation annotation - A first step. In *Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL 2019, Florence, Italy, August 1, 2019*, pages 177–186. Association for Computational Linguistics.

Fabrizio Macagno. 2015. A means-end classification of argumentation schemes. In Frans H. van Eemeren and Bart Garssen, editors, *Reflections on Theoretical Issues in Argumentation Theory*, pages 183–201. Springer International Publishing.

Fabrizio Macagno. 2022. Argumentation profiles and the manipulation of common ground. the arguments of populist leaders on twitter. *Journal of Pragmatics*, 191:67–82.

Fabrizio Macagno, Douglas Walton, and Chris Reed. 2017. Argumentation schemes. history, classifications, and computational applications. *IfCoLog Journal of Logics and Their Applications*, 8(4):2493–2556.

Fabrizio Macagno and Douglas N. Walton. 2015. Classifying the patterns of natural arguments. *Philosophy & Rhetoric*, 48(1):26–53.

Marie-Francine Moens, Erik Boiy, Raquel M. Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *The Eleventh International Conference on Artificial Intelligence and Law, Proceedings of the Conference, June 4-8, 2007, Stanford Law School, Stanford, California, USA*, pages 225–230. ACM.

Allen Newell and Herbert A. Simon. 1995. *GPS, a Program that Simulates Human Thought*, pages 415–428. American Association for Artificial Intelligence, USA.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Geetanjali Rakshit, Kevin K. Bowden, Lena Reed, Amita Misra, and Marilyn A. Walker. 2017. Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents - 8th International Workshop on Spoken Dialog Systems, IWSDS 2017, Farmington, PA, USA, 6-9 June 2017, revised selected papers*, volume 510 of *Lecture Notes in Electrical Engineering*, pages 45–52. Springer.

Chrysi Rapanta and Douglas N. Walton. 2016a. Identifying paralogisms in two ethnically different contexts at university level / identificación de paralogismos en dos contextos universitarios diferenciados étnicamente. *Journal for the Study of Education and Development*, 39(1):119–149.

Chrysi Rapanta and Douglas N. Walton. 2016b. The use of argument maps as an assessment tool in higher education. *International Journal of Educational Research*, 79:211–221.

Chris Reed. 2006. Preliminary results from an argument corpus. In *Linguistics in the Twenty First Century*, pages 185–195. Cambridge Scholar Press, Newcastle, UK.

Ramon Ruiz-Dolz, Joaquín Taverner, John Lawrence, and Chris Reed. 2024a. Nlas-multi: A multilingual corpus of automatically generated natural language argumentation schemes. *Data in Brief*, 57:111087.

Ramon Ruiz-Dolz, Joaquín Taverner, John Lawrence, and Chris Reed. 2024b. Nlas-multi: A multilingual corpus of automatically generated natural language argumentation schemes. *CoRR*, arXiv:2402.14458.

Sougata Saha and Rohini K. Srihari. 2023. Argu: A controllable factual argument generator. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8373–8388. Association for Computational Linguistics.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, and 34 others. 2021. An autonomous debating system. *Nat.*, 591(7850):379–384.

Yi Song, Michael Heilman, Beata B. Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argument Mining, hosted by the 52nd Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2014, June 26, 2014, Baltimore, Maryland, USA*, pages 69–78. The Association for Computer Linguistics.

Frans H. van Eemeren and Rob Grootendorst. 2003. *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. Cambridge University Press, Cambridge.

Bart Verheij. 2003. Dialectical argumentation with argumentation schemes: An approach to legal logic. *Artificial Intelligence and Law*, 11(2/3):167–195.

Jacky Visser, John Lawrence, Chris Reed, Jean H. M. Wagemans, and Douglas N. Walton. 2021. Annotating argument schemes. *Argumentation*, 35(1):101–139.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3753–3765. Association for Computational Linguistics.

Jean H. M. Wagemans. 2016. Constructing a periodic table of arguments. In P. Bondy and L. Benacquista, editors, *Argumentation, Objectivity, and Bias: Proceedings of the 11th International Conference of the Ontario Society for the Study of Argumentation (OSSA)*, pages 1–12, 18–21. OSSA.

Douglas N. Walton. 1996. *Argumentation Schemes for Presumptive Reasoning*. L. Erlbaum Associates, Mahwah, N.J.

Douglas N. Walton and Fabrizio Macagno. 2015. A classification system for argumentation schemes. *Argument & Computation*, 6(3):219–245.

Douglas N. Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.