Mining Complex Patterns of Argumentative Reasoning in Natural Language Dialogue

Ramon Ruiz-Dolz[1], Zlata Kikteva[2], and John Lawrence[1]
[1]Centre for Argument Technology, University of Dundee, United Kingdom
[2]University of Passau, Germany

Abstract

Argumentation scheme mining is the task of automatically identifying reasoning mechanisms behind argument inferences. These mechanisms provide insights into underlying argument structures and guide the assessment of natural language arguments. Research on argumentation scheme mining, however, has always been limited by the scarcity of large enough publicly available corpora containing scheme annotations. In this paper, we present the first state-of-the-art results for mining argumentation schemes in natural language dialogue. For this purpose, we create QT-SCHEMES, a new corpus of 441 arguments annotated with 24 argumentation schemes. Using this corpus, we leverage the capabilities of LLMs and Transformer-based models, pre-training them on a large corpus containing textbook-like argumentation schemes and validating their applicability in real-world scenarios.

1 Introduction

Argument mining is the task of identifying argument structures in unstructured natural language inputs (Mochales and Moens, 2011; Lippi and Torroni, 2016; Lawrence and Reed, 2020). Most of the argument mining research in recent years has focused on the automatic classification of argument components (i.e., premises and claims) (Levy et al., 2014; Habernal and Gurevych, 2017; Haddadan et al., 2019), the identification of argumentative relations (i.e., supports, attacks, or rephrases) between them (Cocarascu and Toni, 2017; Menini et al., 2018; Ruiz-Dolz et al., 2021; Kikteva et al., 2023), or both tasks at the same time with an end-to-end architecture (Bao et al., 2022; Morio et al., 2022). This is, however, a relatively minimal approach when viewed through the lens of the long-standing field of theoretical argumentation that offers an array of fine-grained argument analysis frameworks (Kienpointner (1992); Grennan (1997); Walton et al. (2008); Wagemans (2016)).

One of the contributions from the theoretical field introduces the concept of argumentation schemes that represent common patterns of human argumentative reasoning (Walton et al., 2008). With more than 60 semi-structured representations of inferential patterns, this approach aims to capture argumentation on a granular level. The schemes are represented in sets of premises and conclusions like in Example (1) which illustrates an *Argument from Waste*. An argument that follows this argumentation pattern requires two premises, with the first one in (1-a) explaining how ceasing current actions would result in wasting previous efforts and the second one in (1-b) making it explicit how such waste would result in an undesirable outcome. The conclusion is to continue with the previous course of action. Other scheme structures represent different patterns of reasoning, such as *Cause to Effect*, *Expert Opinion*, and *Popular Practice* among many others.

(1) a. Premise 1: *If* a *stops trying to realise* A *now, all* a*'s previous efforts to realise* A *will be wasted.*
b. Premise 2: *If all* a*'s previous attempts to realise* A *are wasted, that would be a bad thing.*
c. Conclusion: *Therefore,* a *ought to continue trying to realize* A.

This kind of fine-grained analysis extends beyond simple argument detection and allows us to uncover mechanisms underlying human reasoning. For instance, by applying argumentation schemes, we are able to evaluate argument quality by examining how closely real-life arguments match the scheme structure as well as the strength of individual premises (Kondo et al., 2021). Similarly, they can aid in such tasks as fallacy identification (Ruiz-Dolz and Lawrence, 2023), misinformation detection (Musi et al., 2023), and enthymeme re-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7421–7435
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

construction (Delas et al., 2024). The templatic nature of these classifications also allows for their integration into computational systems. However, so far they have only been used in small-scale experiments (Walton, 2012; Lawrence and Reed, 2016; Green, 2018). The lack of extensive annotated corpora containing argumentation scheme information has limited the development of state-of-the-art NLP algorithms, with the matter of automatic analysis of the reasoning patterns in natural language argumentation remaining largely unaddressed.

A recent corpus, developed by Ruiz-Dolz et al. (2024), of almost 2,000 automatically generated arguments that closely follow 20 different types of argumentation scheme structures, like the one in Example (1), opens new avenues for research. However, while undoubtedly a valuable resource, it does not account for the fact that arguments in natural communication rarely follow structures defined in theoretical work precisely. Instead, speakers heavily rely on anaphora, enthymemes, and a wide variety of rhetorical structures when constructing their arguments.

Example (2) offers a real-life *Argument from Waste*[1]. While on the surface level it differs in form from the description in (1), content-wise it matches the scheme requirements. The first premise in (1-a) that the efforts will be wasted is expressed through the use of the word ‘embed’ in (2-a) which implies that the current success is not stable and progress might be lost if certain actions are not taken. The second premise in (1-b) is left out and can be reconstructed from the use of the word ‘success’ in (2-a) since losing something described in such a manner does not constitute a positive outcome.

(2) a. Premise: *We need to make sure that we embed the successes that we have had.*
b. Conclusion: *There is still work to do.*

In this paper, we aim to address the gap in argument mining research by working on the identification of complex argumentative reasoning structures, i.e., argumentation schemes, in natural language dialogue, in particular, in the BBC’s political debate program ‘Question Time’. Creating a corpus of natural language arguments belonging to a set of 24 argumentation schemes that is large enough to train and fine-tune deep learning models is a highly challenging task due to a large number of potential classes and a very unequal class distribution. To mitigate this, we make use of an already existing corpus of automatically generated textbook-like arguments by Ruiz-Dolz et al. (2024) to implement a state-of-the-art NLP algorithm. We then create our own corpus of natural language argumentation schemes uttered in real debates, QT-SCHEMES, for fine-tuning and validating the model’s performance in real-world scenarios.

Using this approach, we investigate various pre-processing strategies and methods to leverage the automatically generated arguments. This, in turn, improves the performance of models deployed on real dialogues with natural language arguments. We find that the model trained on textbook-like examples of argumentation schemes exhibits poor performance when evaluated on real-life data, however, pre-processing of the training corpus and fine-tuning the model on a small set of real-life arguments improves it.

To the best of our knowledge, this is the first work integrating the concept of argumentation schemes into a state-of-the-art argument mining model at this scale and validating it on natural language dialogue argumentation data. The contributions of this paper are therefore three-fold: (1) We present QT-SCHEMES, a dataset of 441 natural language arguments annotated with 24 different argumentation schemes, making it the largest corpus in size and class dimensionality regarding Walton’s schemes. (2) We conduct a set of thorough experiments revealing the most effective way to implement an argument mining model for the identification of argumentation schemes in natural language dialogue with a limited amount of available data. (3) We evaluate the capabilities of Large Language Models (LLMs) to process and identify complex structures of argumentative reasoning.

2 Related Work

The idea of including argumentation schemes into the argument mining process is something that has been discussed and strongly motivated in the past (Walton, 2012). This idea was further developed in works by both Feng and Hirst (2011) and Lawrence and Reed (2016) in which small datasets annotated with up to five different argumentation schemes were used to train machine learning models for the argumentation scheme mining task. Handcrafted features and a set of scheme-related keywords were

[1] Example taken from our corpus (node set ID 23797 from QT-2September2021, access via `http://ova3.arg.tech/`)

used to create argument representations, leading to promising results in a small-scale experimental setup. Green (2018) presented an alternative approach that leverages the logical structures of argumentation schemes to mine scientific discourse. This work, however, did not incorporate features belonging to the natural language of the arguments or any lexical features. Pushing forward natural language argumentation scheme resource availability, Visser et al. (2019) annotated fact, value, and policy arguments belonging to 21 different argumentation schemes in a small-scale corpus consisting of 488 arguments. The authors, however, focused on Wagemans (2016) periodic table of arguments, which exhibits important structural differences from Walton's argumentation schemes.

The use of argumentation scheme structures for argument quality assessment was also explored by Kondo et al. (2021) who pointed out the usefulness of the schemes not only for mining arguments but also for exploring additional aspects of computational argumentation. Finally, in the most recent work, there have been attempts to overcome the issue of limited availability of large enough corpora for argumentation mining. Both Saha and Srihari (2023) and Ruiz-Dolz et al. (2024) use generative NLP models to automatically generate large corpora containing 69,428 and 3,810 natural language arguments respectively. Saha and Srihari (2023) generate arguments belonging to six groups vaguely based on the original structures defined by Walton, while Ruiz-Dolz et al. (2024) provide a compilation of natural language arguments following the complete structural definitions of twenty argumentation schemes, allowing to analyse a broader range of reasoning patterns in natural language. Resources of this nature allow further investigation into state-of-the-art NLP algorithms for mining argumentation schemes in natural language communication.

3 Data

3.1 NLAS

Given the scarcity of large enough resources to implement state-of-the-art NLP algorithms for mining argumentation schemes in human argumentative discourse, we use the Natural Language Argumentation Scheme (NLAS) corpus (Ruiz-Dolz et al., 2024) as the starting point for this paper. The NLAS is the largest publicly available corpus of natural language argumentation schemes consisting of 20 different Walton schemes instantiated into 50 varied topics such as animal testing, climate change, and freedom of speech with two possible stances for each of them, i.e., in favour and against the topic.

The complete corpus (referred to as NLAS-COMP in this paper) contains a total of 1,893 English language independent arguments, evenly distributed with almost 100 arguments per argumentation scheme. These arguments were automatically created using generative LLMs and validated by expert annotators. However, the generation strategy defined in Ruiz-Dolz et al. (2024) largely depends on the complete semi-formal argument structures proposed by Walton et al. (2008), resulting in a very homogeneous corpus of textbook-like arguments, differing substantially from the argument structures found in natural language dialogue, where speakers frequently rely on enthymemes (Breitholtz, 2020) – deliberate omissions of premises and conclusions, a common feature of natural language argumentation.

In an attempt to automatically approximate arguments that are found in natural communication, we pre-process the NLAS-COMP corpus to create a new version for pre-training our models (referred to as NLAS-PROC in this paper). We process the corpus by splitting the propositions (i.e., premises and conclusion) that make up original NLAS arguments into the individual elements to produce subsets of all possible proposition combinations of the original structures. For instance, for the Example (1), we produce the {(1-a)}, {(1-b)}, {(1-c)}, {(1-a), (1-b)}, {(1-a), (1-c)}, and {(1-b), (1-c)} subsets. While this strategy might seem reductive at first, it is effective with a corpus like NLAS where due to their strictly templatic nature, some elements of the arguments would be considered redundant in natural communication. Consider this conclusion from an argument from *Popular Opinion*: "There is a reason to be skeptical of the benefits of intermittent fasting, as it often involves skipping breakfast, which goes against conventional wisdom". Here, the general acceptance premise that breakfast is widely considered the most important meal of the day is implicit while the presumption premise that "skipping breakfast goes against conventional wisdom" is embedded within the conclusion itself. In this way, this conclusion alone functions as an argument.

This step allows us to create a more heterogeneous corpus of arguments better approximating

	Ad Hominem	Based on Cases	Defeasible Rule-based	Discovery	Popular Acceptance	Position to Know	Practical Reasoning	Chained Arguments with Rules and Cases	Total
NLAS-PROC	1,455	3,269	2,771	4,231	2,162	4,555	3,498	1,530	**23,471**
QT-SCHEMES	34	52	19	135	15	53	132	1	**441**

Table 1: Statistics of the NLAS-PROC and QT-SCHEMES corpora (grouped per argumentation scheme family).

those found in natural language dialogue. With this, we obtain a total of 23,471 incomplete arguments ranging between 1,000 and 2,000 per scheme thus also increasing the corpus size. Corpus statistics of both NLAS-COMP and NLAS-PROC in full is available in Appendix A; Table 1 reports distribution of arguments in NLAS-PROC grouped per argumentation scheme family (an experimental design choice explained in Section 4.1).

3.2 QT-SCHEMES

For validating proposed models with naturally occurring arguments in dialogue, we use the QT30 corpus (Hautli-Janisz et al., 2022). The corpus consists of 30 episodes of the BBC political debate show ‘Question Time’ (QT) manually annotated with Inference Anchoring Theory (IAT) (Budzynska et al., 2014, 2016), a framework that captures how arguments evolve in dialogical settings.

We extract arguments from five QT episodes found in the corpus: QT-18March2021, QT-10June2021, QT-19August2021, QT-2September2021, and QT-14October2021. When selecting the episodes we make sure that the topics discussed during the debates do not coincide with the topics used to generate NLAS arguments. Some of the topics covered in the episodes included discussions on the aftermath of the war in Afghanistan, the future of the hospitality industry in the wake of the pandemic, and sexual harassment against women. We extract a total of 891 arguments from the five debates and annotate them with the argumentation schemes based on Walton et al. (2008). We use 24 argumentation schemes which for the most part overlap with the schemes in NLAS corpus except for four of them that are only found in QT data. After the annotation, we end up with a QT-SCHEMES corpus of 441 arguments annotated with one of the argumentation schemes. Statistics of the corpus with argumentation schemes grouped into families is provided in Table 1; full statistics is available in Appendix A.

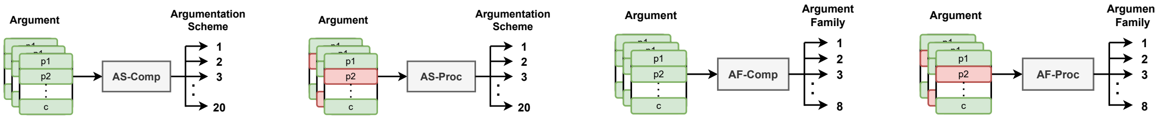
Argumentation scheme identification is a complex task that leverages a large number of classes and requires a certain level of familiarity with argumentation theory. The task, therefore, warrants expert annotations; in this case, two of the authors of the paper annotated the data. The annotation process was guided by the scheme descriptions offered by Walton et al. (2008) (presented in full in Appendix B) and produced annotations of arguments like in Example (2) with one of the schemes. The inter-annotator agreement (IAA) for the annotation process was validated on 12% of the data resulting in Cohen’s κ of 0.39. This fair agreement reflects the inherent complexity of capturing implicit reasoning in natural language, and our subsequent grouping of schemes into families and fine-tuning on the dialogue corpus help mitigate the impact of annotation noise on model performance. Direct comparison to earlier work is challenging due to variability in domain and task configurations. However, the reported agreement is in line with that expected for a complex multi-class argument mining task (see Visser et al. (2020) for discussion on argumentation scheme annotation challenges), considering that even for the simpler task of argument relation identification (support or attack) agreement varies widely, from Krippendorff’s α of 0.43 (Mestre et al., 2021) to 0.81 (Stab and Gurevych, 2014).

4 Methodology

4.1 Classification Task Designs

For all of our models, we consider two classification designs. In the first design (referred to as AS), we use the argumentation schemes as they are found in NLAS and QT30-SCHEMES corpora involving more than 20 classes of arguments.

In our second design (AF) we explore a class dimensionality reduction by grouping argumentation schemes of a similar nature into eight different argumentation scheme families. We follow Walton and Macagno (2015) for the mapping between the argumentation schemes and corresponding families. For instance, *Arguments from Analogy*, *Example*, and *Precedent* belong to the family of *Defeasible Rule-based Arguments*, while *Arguments from Best Explanation*, *Ignorance*, *Random Sample To Popu-*



(a) AS-COMP model. (b) AS-PROC model. (c) AF-COMP model. (d) AF-PROC model.

Figure 1: Models proposed for detecting argumentation schemes in natural language dialogue. The green colour indicates that the argument component (i.e., premise or conclusion) was included as part of the input. The red colour indicates the components that were excluded during the pre-processing.

lation, and *Sign* are part of the *Discovery Argument* family. The aim of this approach is not only to reduce the class dimensionality, but also to improve the learning process by putting together arguments that share similar linguistic features and separating those that do not. In natural dialogue speakers often rely on implicit reasoning and omit certain premises. By grouping similar schemes into families, we reduce the sensitivity to these structural differences, allowing the model to capture more robust, generalised linguistic features. For the complete distribution and mapping of the argumentation schemes to families see Appendix A.

4.2 Classification Model Configurations

We leverage the two versions of NLAS (NLAS-COMP and NLAS-PROC) for pre-training our classification models. Combination of the two task designs and the two versions of NLAS for pre-training results in four models: AS-COMP, AS-PROC, AF-COMP, and AF-PROC, where AS and AF correspond to argumentation scheme and family classification design respectively, while COMP and PROC stand for complete and processed versions of NLAS used for pre-training. These approaches are visualised in Figure 1.

In all these four approaches we model the following conditional probability:

$$\hat{s} = \arg\max_{s \in S} P(s | a_1^N) \tag{1}$$

Where S represents the complete set of 20 argumentation schemes in the AS-COMP and AS-PROC models and the set of eight different argumentation families in the AF-COMP and AF-PROC models. The argument a of length N (i.e., $1 \ldots N$), where N represents the number of argument components, can be instantiated as the complete natural language scheme in the AS-COMP and AF-COMP models or as a subset of its propositions in AS-PROC and AF-PROC models. This way, for a given argument a of length N in the AS-COMP and AF-COMP models, the length of the same argument will be equal or smaller in the AS-PROC and AF-PROC models.

As an extension of the previous four models pre-trained on NLAS only, we consider the four dialogue versions of them by leveraging the newly annotated QT-SCHEMES corpus containing argumentation schemes in natural language dialogue like in the Example (2). Taking the AS-COMP, AS-PROC, AF-COMP, and AF-PROC models as our starting point, we further fine-tune them to model the dialogue argumentation schemes. We refer to the dialogue versions of our models as AS-COMP-DIAL, AS-PROC-DIAL, AF-COMP-DIAL, and AF-PROC-DIAL respectively. Given the nature of the language modelled in the dialogue versions of the original models where a pair of propositions connected by inference is always used, we make the previously defined conditional probability more specific:

$$\hat{s} = \arg\max_{s \in S} P(s | p_1^I, c_1^J) \tag{2}$$

Instead of having a unique argument sequence a, the dialogue models explicitly incorporate two propositions, the premise p and the conclusion c, of variable lengths I and J respectively. By modelling this more specific conditional probability, we obtain results that are more consistent with existing models of argumentative dialogue such as IAT where inference nodes (i.e., argument relations) link two propositional nodes. In the case of linked arguments (i.e., arguments with multiple supporting premises) all of the premises are concatenated as part of the input premise p.

4.3 LLM Prompting Strategies

In order to allow for the comparison between the results of different models, we try to keep our LLM prompting strategies consistent with the classification model configurations. With this in mind, we prompt the LLMs for the tasks involving argumentation schemes (AS) and families (AF), following the zero-shot (ZS) and few-shot (FS) strategies. In the zero-shot setting, the prompt includes descriptions of the argumentation schemes as offered by Walton et al. (2008). For the few-shot strategy, we run two different experiments: one including examples from the textbook-like NLAS corpus, and another one with examples from the QT-SCHEMES dialogue corpus. The combination of these configurations results in six prompting strategies: three in the argumentation scheme task design in zero-shot (AS-ZS), few-shot with NLAS examples (AS-FS), and few-shot with QT-SCHEMES examples (AS-FS-DIAL) prompting settings, as well as corresponding strategies in the argumentation family design (AF-ZS, AF-FS, AF-FS-DIAL). We prompt the models to generate both a label (i.e., argumentation scheme or family) as well as a short explanation supporting this decision[2].

5 Experiments

5.1 Experimental Setup

The models described in Section 4.2 are implemented with a pre-trained RoBERTa-large (Liu et al., 2019) architecture chosen based on its stronger performance compared to other models reported in previous argument mining studies (Ruiz-Dolz et al., 2021). The models are further pre-trained on NLAS and fine-tuned on QT-SCHEMES for 20 epochs with a learning rate of 1e-5 and a weight decay of 0.01. We use a batch size of 42 for pre-training and 112 for fine-tuning. Additionally, we conduct experiments using two different state-of-the-art LLMs, Qwen 2.5 and Llama (3.1 and 3.3 versions), across three model sizes – 7, 8, and 70 billion parameters. For the evaluation of the predicted labels, we use macro-averaged precision, recall, and F1 scores.

We use the NLAS dataset as our pre-training corpus for RoBERTa models and the QT-SCHEMES as our fine-tuning and evaluation corpus. This decision is based on two main considerations: first, the size of the NLAS allows us to train Transformer models for the task of argumentation scheme mining. Second, while the theoretical character of the textbook-like arguments in natural language enables us to model general lexical features of the argumentation schemes, the dialogical nature of the QT-SCHEMES corpus allows us to enhance our models with the lexical features of argumentation observed in natural communication. With this approach, we are able to validate the real impact of the theory-based NLAS when identifying argumentation schemes in the context of arguments uttered in an unstructured debate. Examples from both corpora are used in the corresponding few-shot prompting strategies.

We split the NLAS corpus into train and development according to a 90-10 proportion. The QT-SCHEMES corpus is divided into the fine-tuning and test sets by episode, thus preventing any potential data leakage. QT-10June2021, QT-19August2021, QT-2September2021, and QT-14October2021 are used in fine-tuning and QT-18March2021 is used for evaluation. We select QT-18March2021 as our test file given its completeness as it contains most of the annotated argumentation schemes in the QT-SCHEMES corpus. In order to present consistent and comparable results, all our experiments (i.e., all RoBERTa configurations and LLMs in each prompt setting) are evaluated using the same test file.

Due to some discrepancy in scheme distribution between NLAS and QT-SCHEMES (see Appendix A for details), the four schemes that were absent from NLAS were only seen by the models at the fine-tuning or prompting stages and used in the evaluation. Additionally, due to the episode-wise split of QT-SCHEMES, one scheme was not available for fine-tuning and two were not included in the test set. The code used in our experiments as well as the prompts and the data can be publicly accessed at https://github.com/raruidol/ACL25-ArgumentationSchemeMining. The best-performing fine-tuned model can be downloaded from the Huggingface repository[3].

5.2 Results

The results of our experiments including both the proposed models and the prompted LLMs are presented in Table 2. The first major finding is that the pre-trained models on the 20 different argumentation scheme classes from the NLAS corpus are far from being useful when deployed in a natural language context. The ROBERTA-AS-COMP and ROBERTA-AS-PROC models perform poorly on this task, highlighting the significant differences between textbook-like natural language arguments

[2] Complete examples of the prompts are included in the supplementary material.

[3] ROBERTA-AF-PROC-DIAL: https://huggingface.co/raruidol/RoBERTa-AF-Proc-Dial

and the arguments used by humans in real debates. While a slight improvement can be observed in a setting with the second model compared to the first one indicating that creating combinations of the incomplete subsets of the argumentation schemes is beneficial, none of the models truly succeed in this task. However, we do observe a drastic improvement in the performance of the RoBERTa-AF-Comp and RoBERTa-AF-Proc models (31.7 and 49.7 F1 respectively) which utilise argumentation scheme family groupings, indicating that the features learned in this setting maintain their relevance when considering the same families presented in the natural language dialogues included in the QT-Schemes corpus. Interestingly, LLMs exhibit better performance in the 20+ argumentation scheme classification setting than RoBERTa models. In particular, Llama3.3(70b)-AS-FS achieves an F1-score of 29.4, which makes it the best-performing model for this task configuration. However, despite the improved score, this model would still not perform sufficiently well for the automatic identification of argumentation schemes in natural dialogues.

Model	Precision	Recall	F1-score
RoBERTa-AS-Comp	0.8	4.5	1.0
RoBERTa-AS-Proc	3.4	6.2	3.1
RoBERTa-AS-Comp-Dial	7.4	9.4	8.0
RoBERTa-AS-Proc-Dial	8.2	10.7	9.0
Qwen2.5(7b)-AS-ZS	4.1	14.3	5.7
Llama3.1(8b)-AS-ZS	9.9	8.9	6.6
Llama3.3(70b)-AS-ZS	18.9	24.4	18.7
Qwen2.5(7b)-AS-FS	3.7	11.0	5.5
Llama3.1(8b)-AS-FS	4.5	12.2	5.4
Llama3.3(70b)-AS-FS	**31.2**	**45.4**	**29.4**
Qwen2.5(7b)-AS-FS-Dial	7.4	16.2	7.8
Llama3.1(8b)-AS-FS-Dial	18.6	18.9	14.4
Llama3.3(70b)-AS-FS-Dial	22.1	27.9	22.3
RoBERTa-AF-Comp	45.5	38.1	31.7
RoBERTa-AF-Proc	57.7	56.3	49.7
RoBERTa-AF-Comp-Dial	**65.1**	47.7	49.3
RoBERTa-AF-Proc-Dial	62.1	**66.9**	**62.3**
Qwen2.5(7b)-AF-ZS	12.3	26.1	13.5
Llama3.1(8b)-AF-ZS	13.1	21.3	13.9
Llama3.3(70b)-AF-ZS	44.4	44.2	34.7
Qwen2.5(7b)-AF-FS	8.3	20.8	11.0
Llama3.1(8b)-AF-FS	10.1	17.2	11.9
Llama3.3(70b)-AF-FS	18.7	34.4	23.8
Qwen2.5(7b)-AF-FS-Dial	27.9	16.9	14.7
Llama3.1(8b)-AF-FS-Dial	38.7	28.4	28.1
Llama3.3(70b)-AF-FS-Dial	34.7	36.8	31.8

Table 2: Results of the evaluation of our argumentation scheme identification models on the natural language dialogue QT-Schemes test corpus. The first half of the table contains the argumentation scheme (AS) classification experiments. The second half contains the scheme family (AF) classification experiments.

Although there is an improvement in the dialogue versions of our pre-trained models, RoBERTa-AS-Comp-Dial and RoBERTa-AS-Proc-Dial still exhibit poor performance. This is most probably due to the trade-off between the size of the fine-tuning part of the QT-Schemes corpus and the large class dimensionality. Our experiments with LLMs reveal that they exhibit limited improvement when provided with dialogue examples. Interestingly, the Llama 3.3 model prompted with schemes extracted from dialogues (Llama3.3(70b)-AS-FS-Dial) performed slightly worse than the one prompted with textbook-like arguments (Llama3.3(70b)-AS-FS) which suggests that LLMs can not generalise well enough when dealing with the complexity and implicitness of reasoning in natural dialogue scenarios.

Finally, regarding our experimental results when considering argumentation scheme families, we observe significant differences. The best performance is achieved with the RoBERTa-AF-Proc-Dial model with an F1-score of 62.3, higher by more than 10 points when compared to our second best-performing model, RoBERTa-AF-Proc. Moreover, the improvement of the RoBERTa-AF-Proc-Dial over RoBERTa-AF-Comp-Dial further highlights the benefits of our pre-processing strategy for the NLAS training corpus. Such notable performance improvement, however, can not be observed in our experiments with LLMs, which underperformed in the task, presenting similar results to those achieved in the argumentation scheme classification setup despite the significant reduction in class dimensionality. Here, the highest score is exhibited in the zero-shot setting by LLAMA3.3(70B)-AF-ZS instead. This reinforces our claim that LLMs struggle to generalise after a certain point, making them difficult to use in a real-world setting.

These findings lead us to two main conclusions. First, the automatically generated textbook-like natural language argumentation schemes are indeed helpful for developing argumentation scheme mining systems after some theory-based pre-processing. Second, fine-tuning on natural language dialogue data is a necessary step to effectively deploying such systems in real-world dialogue scenarios.

5.3 Error Analysis

In order to better understand the results, we perform error analysis on the predictions of RoBERTa-AF-Proc and RoBERTa-AF-Proc-Dial (best-performing configurations with RoBERTa) as well as LLAMA3.3(70B)-AS-FS and LLAMA3.3(70B)-AF-ZS (best-performing models among LLMs) [4].

We find that RoBERTa-AF-Proc is better at identifying *Position to Know* arguments, which frequently follow the corresponding argumentation scheme closely by providing the speaker's relevant background and making the fact that they are asserting a claim regarding a certain entity within their 'expertise' explicit. This, however, also leads to an increased number of false positive *Position to Know* predictions by the model as it is more affected by the presence of named entities. This is further supported by the fact that another argumentation scheme family that results in a high number of false positives in this setting is the *Ad Hominem* one that presupposes an attack on an entity. Here, a general negative sentiment of the argument (rather than a targeted one) and the presence of named entities might serve as a false predictor to the model.

As for the results in RoBERTa-AF-Proc-Dial setting, it predicts more correct labels for *Discovery* and *Practical Reasoning* argumentation families, the most prevalent ones in our corpus. However, most of the false positives were also assigned to the *Discovery* family. Walton's scheme for one of the most frequent arguments of the family, *Best Explanation*, requires one premise of "a finding or a given set of facts" and two premises clarifying how the first premise is most satisfactory. The fact that the first premise can take a variety of different forms in natural communication and the remaining two premises are often implied potentially explains the overprediction of this category.

When it comes to the generative models, the overall performance is significantly lower than that of RoBERTa, however, error analysis still offers a few insightful observations. LLAMA3.3(70B)-AF-ZS exhibits better performance with both *Ad Hominem*, seemingly demonstrating a stronger capability to detect premises with targeted negative sentiment than RoBERTa models, and *Popular Acceptance*, potentially, due to the fact that LLM embeddings better encode generally accepted premises that are often implicit in natural communication. However, the majority of *Discovery* arguments were misclassified as *Practical Reasoning*. In this case, varied linguistic surface of *Best Explanation* arguments in the *Discovery* family hinders LLMs ability to predict this class correctly. It is difficult to make generalisations concerning LLAMA3.3(70B)-AS-FS results due to the high-class dimensionality of this configuration, however, we observe that in a few instances the model did not follow the instructions when assigning the appropriate label.

We also examine the explanations for the correct predictions and find that they seem reasonable on the surface level with the model relying on patterns in scheme structures as well as direct or indirect references to the arguments themselves. However, some of them fall apart under closer scrutiny. For instance, several explanations refer to the claims or premises in a generalised way that fits the argumentation scheme or family but does not provide a direct reference to the appropriate elements of the argument that is being analysed, while in other cases, the explanations quote the wrong parts of the argument to support correct 'reasoning'.

6 Conclusion

In this paper, we present the first large-scale experiments in argumentation scheme mining in natural language dialogue. For this, we create QT-Schemes, a corpus consisting of more than 400 natural language arguments annotated with 24 different argumentation schemes. Using this corpus for fine-tuning and evaluation, we leverage the capabilities of the state-of-the-art NLP models pretrained with textbook-like automatically generated arguments, and evaluate the capabilities of LLMs to identify complex patterns of argumentative reasoning.

We are not surprised to find that the high dimensionality of the task (with 20+ possible argumentation scheme classes) poses a considerable challenge to the models' abilities to distinguish between different types of schemes and generalise from automatically generated to natural language arguments. However, while LLM performance is underwhelming, our pre-processing strategy of RoBERTa's pretraining data and fine-tuning the model on natural language dialogue data show promising results when validating in real-world scenarios, specifically when grouping the data into larger argumen-

[4] Confusion matrices are available in the supplementary materials.

tation scheme families. With this paper, we report new state-of-the-art results in the under-researched area of mining argumentation schemes, motivating further research of more complex aspects of argument mining.

Acknowledgments

This work is partially funded by: the 'AI for Citizen Intelligence Coaching against Disinformation (TITAN)' project, funded by the EU Horizon 2020 research and innovation programme under grant agreement 101070658, and by UK Research and innovation under the UK governments Horizon funding guarantee grant numbers 10040483 and 10055990; the 'Artificial Intelligence for Institutionalised, Multimodal, Gamified, Mass Democratic Deliberations' project, funded by the EU Horizon Europe Framework Programme (HORIZON) under grant agreement 101178806; the 'CLARUS' project, funded by the EU Horizon Europe Framework Programme (HORIZON) under grant agreement 101121182; and by VolkswagenStiftung under grant Az. 98544 'Deliberation Laboratory'.

Limitations

The main limitation of this paper comes with the size of the annotated QT-schemes corpus. Annotating argumentation schemes is a challenging and expensive task due to the large number of potential classes and their unequal distribution in natural communication. We were able to able to annotate in a reasonable time 891 arguments of which only 441 belonged to one argumentation scheme type. Furthermore, the large class dimensionality of the problem addressed in this paper can affect the stability of our results, with small variations in the test sample having a significant impact on the performance scores. This is also one of the reasons why we added the scheme family grouping, which allowed us to provide more robust results. As future work, it would be interesting to enlarge the size of the dialogue corpus and explore if our findings remain consistent and improve considering a wider set of domains and topics. That is, however, out of the scope of this paper. Furthermore, other alternative corpora containing annotated argumentation schemes are smaller and contain narrower sets of scheme types, making it difficult to compare the performance of our systems in these datasets.

References

Jianzhu Bao, Yuhang He, Yang Sun, Bin Liang, Jiachen Du, Bing Qin, Min Yang, and Ruifeng Xu. 2022. A generative model for end-to-end argument mining with reconstructed positional encoding and constrained pointer mechanism. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10437–10449.

Ellen Breitholtz. 2020. *Enthymemes and Topoi in Dialogue: the use of common sense reasoning in conversation*. Brill.

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. Towards argument mining from dialogue. In *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014)*, pages 185–196. IOS Press.

Katarzyna Budzynska, Mathilde Janier, Chris Reed, and Patrick Saint Dizier. 2016. Theoretical foundations for illocutionary structure parsing. *Argument & Computation*, 7(1):91–108.

Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1374–1379.

Zvonimir Delas, Brian Plüss, and Ramon Ruiz-Dolz. 2024. An argumentation scheme-based framework for automatic reconstruction of natural language enthymemes. In *Computational Models of Argument*, pages 61–72. IOS Press.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 987–996.

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

Wayne Grennan. 1997. *Informal Logic: Issues and Techniques*. McGill-Queen's Press-MQUP.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational linguistics*, 43(1):125–179.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. QT30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages

3291–3300, Marseille, France. European Language Resources Association.

Manfred Kienpointner. 1992. *Alltagslogik: struktur und funktion von argumentationsmustern*. Frommann-Holzboog.

Zlata Kikteva, Alexander Trautsch, Patrick Katzer, Mirko Oest, Steffen Herbold, and Annette Hautli-Janisz. 2023. On the impact of reconstruction and context for argument prediction in natural debate. In *Proceedings of the 10th Workshop on Argument Mining*, pages 100–106, Singapore. Association for Computational Linguistics.

Takahiro Kondo, Koki Washio, Katsuhiko Hayashi, and Yusuke Miyao. 2021. Bayesian argumentation-scheme networks: A probabilistic model of argument validity facilitated by argumentation schemes. In *Proceedings of the 8th Workshop on Argument Mining*, pages 112–124.

John Lawrence and Chris Reed. 2016. Argument mining using argumentation scheme structures. In *COMMA*, pages 379–390.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Never retreat, never retract: Argumentation analysis for political speeches. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial intelligence and law*, 19:1–22.

Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, and Kohsuke Yanai. 2022. End-to-end argument mining with cross-corpora multi-task learning. *Transactions of the Association for Computational Linguistics*, 10:639–658.

Elena Musi, Elinor Carmi, Chris Reed, Simeon Yates, and Kay O'Halloran. 2023. Developing misinformation immunity: How to reason-check fallacious news in a human–computer interaction environment. *Social Media+ Society*, 9(1):20563051221150407.

Ramon Ruiz-Dolz, Jose Alemany, Stella M Heras Barberá, and Ana García-Fornes. 2021. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36(6):62–70.

Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *Proceedings of the 10th Workshop on Argument Mining*. Association for Computational Linguistics.

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

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)*, pages 8373–8388.

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.

Jacky Visser, John Lawrence, Chris Reed, Jean Wagemans, and Douglas Walton. 2020. Annotating argument schemes. In *Argumentation through languages and cultures*, pages 101–139. Springer.

Jacky Visser, John Lawrence, Jean Wagemans, and Chris Reed. 2019. An annotated corpus of argument schemes in us election debates. In *Proceedings of the 9th Conference of the International Society for the Study of Argumentation (ISSA), 3-6 July 2018*, pages 1101–1111.

Jean Wagemans. 2016. Constructing a periodic table of arguments. In *Argumentation, objectivity, and bias: Proceedings of the 11th international conference of the Ontario Society for the Study of Argumentation (OSSA), Windsor, ON: OSSA*, pages 1–12.

Douglas Walton. 2012. Argument mining by applying argumentation schemes. *Studies in Logic*, 4(1):2011.

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

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

A Argumentation Schemes Distribution and Families Mapping

Argumentation Family	Argumentation Scheme	NLAS		QT-SCHEMES	
		COMP	PROC	Total	Ft/Te
Ad Hominem Arguments	Allegation of Bias	0	0	1	0/1
	Direct Ad Hominem	100	573	16	13/3
	Inconsistent Commitment	89	882	17	15/2
Arguments Based on Cases	Cause to Effect	99	1,146	41	35/6
	Established Rule	95	1,008	3	1/2
	Verbal Classification	99	1,115	8	4/4
Defeasible Rule-based Arguments	Analogy	100	1,165	8	7/1
	Example	97	550	5	4/1
	Precedent	94	1,056	6	4/2
Discovery Arguments	Best Explanation	100	2,112	111	86/25
	Ignorance	93	1,122	5	3/2
	Random Sample to Population	0	0	2	1/1
	Sign	100	997	17	11/6
Popular Acceptance Arguments	Popular Opinion	99	1,096	10	5/5
	Popular Practice	94	1,066	5	4/1
Position to Know Arguments	Expert Opinion	100	1,195	16	15/1
	Position to Know	100	1,182	28	16/12
	Witness Testimony	100	2,178	9	3/6
Practical Reasoning Arguments	Consequences	0	0	34	34/0
	Practical Reasoning	0	0	63	48/15
	Sunk Costs	93	1,098	8	7/1
	Threat	88	1,520	18	17/1
	Waste	86	880	9	8/1
Chained Arguments with Rules and Cases	Slippery Slope	76	1,530	1	1/0
Total	-	1,902	23,471	441	331/100

Table 3: Distribution of the argumentation schemes and mapping onto argumentation scheme families according to Walton and Macagno (2015). For NLAS we include the number of the arguments in the original version of the corpus (COMP) and in the version that was pre-processed for this paper (PROC). The numbers for QT-SCHEMES include the total number of arguments, fine-tuning (Ft), and test (Te) splits. Note, that the split is done on the basis of the corpus structure (one episode with the most representative distribution of classes used for evaluating all models and the rest is used for fine-tuning (only RoBERTa models).

B Scheme descriptions (based on Walton et al. (2008))

Allegation of Bias

Major Premise: If x is biased, then x is less likely to have taken the evidence on both sides into account in arriving at conclusion A.
Minor Premise: Arguer a is biased.
Conclusion: Arguer a is less likely to have taken the evidence on both sides into account in arriving at conclusion A.

Direct Ad Hominem

Premise: a is a person of bad character.
Conclusion: Therefore, a's argument α should not be accepted.

Inconsistent Commitment

Initial Commitment Premise: a has claimed or indicated that he is committed to proposition A (generally, or by virtue of what he has said in the past).
Opposed Commitment Premise: Other evidence in this particular case shows that a is not really committed to A.
Conclusion: a's commitments are inconsistent.

Cause to Effect

Major Premise: Generally, if A occurs, then B will (might) occur.
Minor Premise: In this case, A occurs (might occur).
Conclusion: Therefore, in this case, B will (might) occur.

Established Rule

Major Premise: If carrying out types of actions including A is the established rule for x, then (unless the case is an exception), x must carry out A.
Minor Premise: Carrying out types of actions including A is the established rule for a.
Conclusion: Therefore, a must carry out A.

Verbal Classification

Individual Premise: a has property F.
Classification Premise: For all x, if x has property F, then x can be classified as having property G.
Conclusion: a has property G.

Analogy

Similarity Premise: Generally, case C_1 is similar to case C_2.
Base Premise: A is true (false) in case C_1.
Conclusion: A is true (false) in case C_2.

Example

Premise: In this particular case, the individual a has property F and also property G.
Conclusion: Therefore, generally, if x has property F, then it also has property G.

Precedent

Major Premise: Generally, according to the established rule, if x has property F, then x also has property G.
Minor Premise: In this legitimate case, a has F but does not have G.
Conclusion: Therefore, an exception to the rule must be recognized, and the rule appropriately modified or qualified.

Best Explanation

Premise: F is a finding or given set of facts.
Premise: E is a satisfactory explanation of F.
Premise: No alternative explanation E_1, . . . E_n given so far is as satisfactory as E.
Conclusion: Therefore, E is plausible, as a hypothesis

Ignorance

Major Premise: If A were true, then A would be known to be true.
Minor Premise: It is not the case that A is known to be true.
Conclusion: Therefore, A is not true.

Random Sample to Population

Sign

Specific Premise: *A* (a finding) is true in this situation.
General Premise: *B* is generally indicated as true when its sign, *A*, is true.
Conclusion: *B* is true in this situation.

Popular Opinion

General Acceptance Premise: *A* is generally accepted as true.
Presumption Premise: If *A* is generally accepted as true, that gives a reason in favor of *A*.
Conclusion: There is a reason in favor of *A*.

Popular Practice

Major Premise: *A* is a popular practice among those who are familiar with what is acceptable or not in regard to *A*.
Minor Premise: If *A* is a popular practice among those familiar with what is acceptable or not with regard to *A*, that gives a reason to think that *A* is acceptable.
Conclusion: Therefore, *A* is acceptable in this case.

Expert Opinion

Major Premise: Source *E* is an expert in subject domain *S* containing proposition *A*.
Minor Premise: *E* asserts that proposition *A* is true (false).
Conclusion: *A* is true (false).

Position to Know

Major Premise: Source *a* is in position to know about things in a certain subject domain *S* containing proposition *A*.
Minor Premise: *a* asserts that *A* is true (false).
Conclusion: *A* is true (false).

Witness Testimony

Position to Know Premise: Witness *W* is in a position to know whether *A* is true or not.
Truth Telling Premise: Witness *W* is telling the truth (as *W* knows it).
Statement Premise: Witness W states that *A* is true (false).
Conclusion: *A* may be plausibly taken to be true (false).

Consequences

Premise: If *A* is brought about, good (bad) consequences will plausibly occur.
Conclusion: Therefore, *A* should (not) be brought about.

Practical Reasoning

Goal Premise: Bringing about S_n is my goal.
Means Premise: In order to bring about S_n, I need to bring about S_i.
Conclusion: Therefore, I need to bring about S_i.

Sunk Costs

t1: Time of the proponent's commitment to a certain action (pre-commitment)
t2: Time of proponent's confrontation with the decision whether carry out the pre-commitment or not.
Premise 1: There is a choice at *t2* between *A* and *not-A*.
Premise 2: At *t2* I am precommitted to *A* because of what I did or committed myself to at *t1*.
Conclusion: Therefore, I should choose *A*.

Threat

Premise 1: If you bring about *A*, some cited bad consequences, *B*, will follow.
Premise 2: I am in position to bring about *B*.
Premise 3: I hereby assert that in fact I will see to it that *B* occurs if you bring about *A*.
Conclusion: Therefore, you had better not bring about *A*.

Waste

Premise 1: If *a* stops trying to realize *A* now, all *a*'s previous efforts to realize *A* will be wasted.
Premise 2: If all *a*'s previous attempts to realize *A* are wasted, that would be a bad thing.
Conclusion: Therefore, *a* ought to continue trying to realize *A*.

Slippery Slope

First Step Premise: A_0 is up for consideration as a proposal that seems initially like something that should be brought about.
Recursive Premise: Bringing up A_0 would plausibly lead (in the given circumstances, as far as we know) to A_1, which would in turn plausibly lead to A_2,
and so forth, through the sequence A_2, . . . A_n.
Bad Outcome Premise: A_n is a horrible (disastrous, bad) outcome.
Conclusion: A_0 should not be brought about.