# Overview of DialAM-2024: Argument Mining in Natural Language Dialogues

**Ramon Ruiz-Dolz, John Lawrence, Ella Schad and Chris Reed**
Centre for Argument Technology (ARG-tech)
University of Dundee
Dundee DD1 4HN, United Kingdom
`{rruizdolz001, j.lawrence, e.m.schad, c.a.reed}@dundee.ac.uk`

## Abstract

Argumentation is the process by which humans rationally elaborate their thoughts and opinions in written (e.g., essays) or spoken (e.g., debates) contexts. Argument Mining research, however, has been focused on either written argumentation or spoken argumentation but without considering any additional information, e.g., speech acts and intentions. In this paper, we present an overview of DialAM-2024, the first shared task in dialogical argument mining, where argumentative relations and speech illocutions are modelled together in a unified framework. The task was divided into two different sub-tasks: the identification of propositional relations and the identification of illocutionary relations. Six different teams explored different methodologies to leverage both sources of information to reconstruct argument maps containing the locutions uttered in the speeches and the argumentative propositions implicit in them. The best performing team achieved an F1-score of 67.05% in the overall evaluation of the reconstruction of complete argument maps, considering both sub-tasks included in the DialAM-2024 shared task.

## 1 Introduction

Argument Mining (Lawrence and Reed, 2020) investigates the automatic extraction of argument structures from natural language inputs. The nature of argumentation, however, can be very variable depending on its context, presenting significant differences between written and spoken argumentation (Hitchcock, 2009), and between monological and dialogical argumentation (O'Keefe, 1977). Research in argument mining has mainly focused on the extraction of arguments only considering argument annotations such as premises and claims (Stab et al., 2018; Reimers et al., 2019) or attacks and supports between propositions (Hou and Jochim, 2017; Ruiz-Dolz et al., 2021; Saadat-Yazdi et al., 2023), without bringing into consideration additional relevant information that could be extracted from the speeches uttered in the dialogues. This is mostly due to the lack of presence of a framework annotating dialogue information in addition to argument structures in argument mining work.

Inference Anchoring Theory (IAT) was proposed as an annotation framework for dialogue argumentation where not only the structure of arguments is captured, but also the speech acts and speaker intent is also annotated to support and contextualise argumentation in dialogues (Budzynska and Reed, 2011; Janier et al., 2014). Therefore, when approaching argument mining in dialogues, IAT represents an ideal framework to expand the standard annotations typically used in argument mining research (i.e., attacks and supports between premises and claims) integrating dialogical information to the argument mining process. Although several corpora and resources annotated with IAT such as US2016 (Visser et al., 2019), QT30 (Hautli-Janisz et al., 2022), RIP (Schad et al., 2024), or FORECAST (Górska et al., 2024) have been released in the last years[1], there is a lack of systematic research in dialogical argument mining integrating these speech features into argument mining systems.

DialAM-2024 represents the first shared task in dialogical argument mining bringing together argument and speech annotations in an attempt to systematically explore the potential benefits of combining both when developing argument mining systems to be used in transcribed argumentative dialogues. The DialAM-2024 shared task received submissions from six different teams exploring a broad set of approaches to integrate and combine argument and dialogue features for argument mining. In this paper, we describe the DialAM-2024 shared task, summarise the most important aspects

[1] All publicly available at the AIFdb: `http://www.aifdb.org/search`

*Proceedings of the 11th Workshop on Argument Mining*, pages 83–92
August 15, 2024 ©2024 Association for Computational Linguistics

of the submitted systems, and provide an in-depth analysis of the final results of the shared task. Furthermore, we perform a qualitative analysis of the output of the best performing system, pointing out the open challenges that will need to be addressed in future work.

## 2 DialAM-2024

### 2.1 Task

The DialAM-2024[2] shared task explores, for the first time, argument mining in dialogues where information from both arguments and dialogues is modelled together. For this purpose, we use IAT, a domain independent annotation framework designed for capturing argument structures anchored to locutions via illocutions. DialAM-2024, therefore, consists of two sub-tasks: the identification of propositional (argumentative) relations, and the identification of illocutionary (speech act) relations. The data used to develop and evaluate the systems submitted to the DialAM-2024 task includes annotations for both sub-tasks together, providing a direct connection between the dialogue speeches and the annotated argumentative structures. With this shared task, it is our goal to motivate the research on the relations between dialogical information and argumentative structures jointly. This way, it is our goal to take a step forward from previous sequence modelling-based approaches, only considering pairs of sentences or argumentative discourse units (ADUs) to automatically identify argument structures, where much of the relevant information to argumentation remains implicit behind the natural language.

This way, the two DialAM-2024 sub-tasks are defined as follows:

A. **Identification of Propositional Relations**. In the first task, the goal is to detect argumentative relations existing between the argumentative propositions directly extracted from the locutions uttered in the argumentative dialogues. Such relations are: Inference, Conflict, and Rephrase.

B. **Identification of Illocutionary Relations**. In the second task, the goal is to detect illocutionary relations existing between locutions uttered in the dialogue and the argumentative propositions associated with them including: Asserting, Agreeing, Arguing, Disagreeing, Challenging, Restating, Pure Questioning, Rethorical Questioning, and Assertive Questioning.

The final goal of the DialAM-2024 shared task is, therefore, to reconstruct graph-structured argument maps, containing locutions and argument propositions previously identified and segmented from argumentative dialogues.

### 2.2 Evaluation

We measured the macro-averaged Precision, Recall and F1-score to evaluate the performance of the submitted systems. The evaluation of the DialAM-2024 shared task was performed at two different levels: *focused* and *general*. In the *focused* evaluation setup, we only considered the related propositions/locutions in the gold standard files, ignoring all the possible combinations of non related propositions/locutions. To complement it, we also considered a *general* evluation setup, where the whole argument map was included in the evaluation. This way, a high performance in the *general* setup but low in the *focused* setup represents a pessimistic approach that leaves more nodes without any relation than it should be. For an extreme case of this first situation, see the majority baseline described below. Conversely, a high performance in the *focused* setup but low in the *general* setup represents an optimistic approach establishing more relations between propositions/locutions than actually exist.

Furthermore, the evaluation was conducted independently for the two sub-tasks included in DialAM-2024, and globally combining the scores of the two independent evaluations. We named as ARI (from argument relation identification) the evaluation of the performance on Task A: Identification of Propositional Relations, and as ILO (from illocutionary relation identification) the evaluation of the performance of the submitted systems on Task B: Identification of Illocutionary Relations. Finally, we refer to the final results combining both sub-tasks and considering the complete argument maps as the Global evaluation.

### 2.3 Baselines

We included two different baselines as a reference for the submitted systems to the DialAM-2024 shared task: a majority baseline that always assigns the majority class (no relation) to all the possible pairs of sequences, and a pre-trained RoBERTa-

[2] The website for this shared task can be found at the following link: dialam.arg.tech

large model for sequence pair classification based on (Ruiz-Dolz et al., 2021).

- MAJORITY-BL: Given that most of the possible combinations of propositions/locutions are not related at all, no relations are assigned between nodes in the argument maps for all ARI, ILO, and Global evaluations.

- ROBERTA-BL: The system consists of two RoBERTa-large (Liu et al., 2019) model architectures fine-tuned independently for both Tasks A and B, approaching the problem as a sequence pair classification problem. No interaction between argumentative and dialogue information is considered in this baseline.

## 3 Data

### 3.1 The QT30 Corpus

QT30 (Hautli-Janisz et al., 2022) is the largest individual corpus of analysed dialogical argumentation at 280,000 words, made up of thirty episodes of one of the most viewed political talk show in the UK, "Question Time"; it features topical debates where the audience members ask questions or request justifications from the panel members: people who are political or societal figures [3]. The audience members will typically be from the area in which they host that week's show, thereby also determining the kind of questions that are asked. For instance, if the show was hosted in Scotland, then there may be questions about Independence or relations between England and Scotland. These thirty episodes were broadcast in 2020 and 2021, covering national scandals and controversial debates such as Brexit, how the government handled COVID19, the subject of vaccination, as well as topics such as PartyGate (politicians in power partying during lockdowns). Question Time (QT) is moderated by a neutral third party who takes questions from the audience and prompts panel members.

The purpose of QT30 was to identify the argumentative structure within these politically relevant debates by annotating the dialogical and propositional structure, as well as identifying the relations used (support, rephrase, or attack) and the illocutionary force of contributions. The authors report inter-annotator agreement (IAA) of 0.56, using CASS (Duthie et al., 2016).

The analysis within the QT30 paper reveals interesting facets of argumentation within broadcast debate, e.g., how the use of conflicts and supports differ between roles. We use this dataset for two reasons: its size and the depth of annotation captured. The size provides us with more data with which to train models; as for depth of annotation, IAT was specifically developed to capture argumentative dialogues and, in a task where we ask participants to identify argumentative relations and illocutionary forces while incorporating additional dialogical information, is ideally suited to provide us the necessary annotation.

### 3.2 Annotation

IAT provides a theoretical scaffold to handle dialogue and argument structures, and the relations between them. It is used in order to represent, and to gain insight into, the arguments people make in complex dialogues. For IAT diagramming we use OVA+, an online tool developed for the analysis of arguments (Janier et al., 2014). The IAT framework and its OVA tool have been used for more than 2.5 million words of analysed argumentation.[4]

The smallest units of the IAT analysis are argumentative discourse units (ADUs), typically directly analysed as locutions. Locutions are in the text boxes on the right of the graph structure and are known as L-nodes. 'Edges' (incoming and outgoing) is the term used to describe the relations, illocutionary forces, and Default Transitions (TAs) anchored in the nodes. Propositions are on the left-hand side and are reconstructed locutions, where linguistic features like anaphora, pronouns, and deixis are resolved. IAT has three types of relations: (i) relations between locutions in a dialogue, called transitions; (ii) relations between content (propositional content of locutions); and (iii) illocutionary connections that link locutions with their content. Locutions have speakers and typically also have timestamps. The text of locutions is not reconstructed or changed in any way from the source data in contrast to the propositional content of those locutions. Locutions and propositions are connected via illocutionary connections. The guidelines used for annotation are available publicly[5].

As an example of IAT annotation, Figure 1 shows the typical structure of a QT episode: the

[3] QT30 as a corpus is publicly accessible at http://corpora.aifdb.org/qt30

[4] The OVA tool is made available at the following address: ova.arg.tech

[5] Annotation guidelines: https://www.arg.tech/index.php/annotation-guidelines/

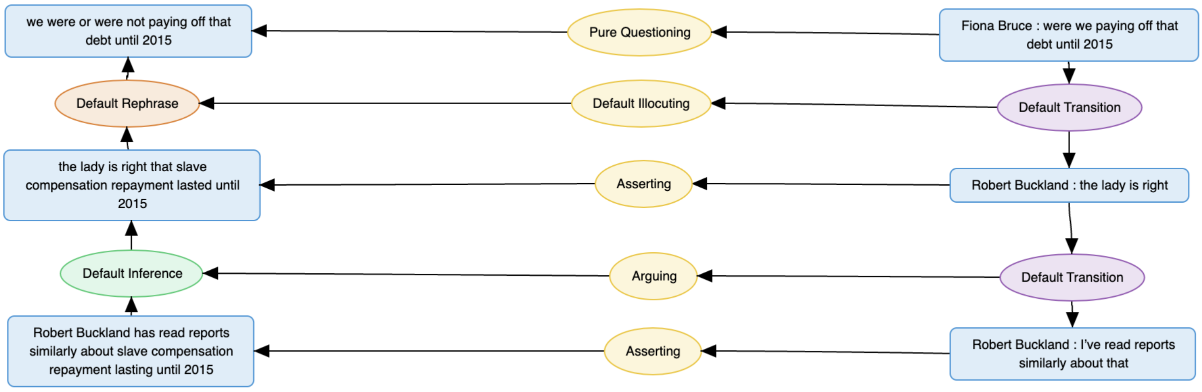
we were or were not paying off that debt until 2015
Pure Questioning
Fiona Bruce : were we paying off that debt until 2015
Default Rephrase
Default Illocuting
Default Transition
the lady is right that slave compensation repayment lasted until 2015
Asserting
Robert Buckland : the lady is right
Default Inference
Arguing
Default Transition
Robert Buckland has read reports similarly about slave compensation repayment lasting until 2015
Asserting
Robert Buckland : I've read reports similarly about that

Figure 1: Example of data annotated using IAT: rectangular blue boxes indicating propositions and locutions, yellow ovals for illocutionary connections, purple ovals for discourse transitions, the green oval for inference relation, the orange oval either showing a rephrase relation or signalling the answer to a question, as is the case here.

moderator puts forward a question and a panel member replies. The Default Rephrase node is used to denote a rephrase relation, unless anchored by a "Pure Questioning" illocutionary force, as is the case here. Robert Buckland's intention of creating an argument is captured through the Default Inference. As each proposition should be understandable by itself, the second and third proposition have been reconstructed; the second with what "the lady" was right about and the third with who "I" and "that" refers to.

### 3.3 Training Phase

For the training phase, the participants were given the QT30 data in JSON format. We provided additional information sheets to participants on the DialAM website[6][7] about the style of annotation, as well as how to read the JSON-based format. The QT30 corpus contains 10,818 propositional relations and 32,303 illocutionary relations in 1,478 JSON files.

### 3.4 Evaluation Phase

For the evaluation phase, we chose eleven JSON files containing argument maps that included some challenging argument structures. This was for the purpose of evaluating how participants systems dealt with these complex cases. An example of the complex argumentation available in the data and used for evaluation would be linked, convergent, and divergent arguments.

## 4 Submissions

Fifteen different teams registered for the DialAM-2024 shared task, of which six submitted their system outputs during the evaluation phase. All the submissions addressed the two sub-tasks of the shared task: (A) identification of propositional relations, and (B) identification of illocutionary relations. The submitted systems include a broad set of different language modelling approaches and model architectures. A summary of the submitted systems together with the references to their system description papers where the approaches are described in-depth is provided below.

### 4.1 System Descritpions

**dfki-mlst** (Binder et al., 2024) approaches the shared task as an n-ary classification problem. Their proposed method relies on three main steps: normalise the nodeset, encode the relations for both propositions and illocutions, and train a classification model. The authors submitted a fine-tuned DeBERTa-v3 model (He et al., 2021) as their classification model for the evaluation phase. In addition to DeBERTa-v3, the authors carried out an extensive comparison of different model architectures for the base classifier after the end of the evaluation phase, pointing out that DeBERTa-v1 performed better on test (although DeBERTa-v3 got better results during validation), and that combining the text data included in L and I nodes helps to improve the performance of the submitted system in both *focused* and *general* evaluation setups.

**KnowComp** (Wu et al., 2024) proposes a three-stage sequential inference pipeline to approach the shared task based on prompt-based fine-tuning.

[6] Data format details: http://dialam.arg.tech/res/files/data-format.pdf

[7] Annotation details: http://dialam.arg.tech/res/files/annotation-details.pdf

The first stage is aimed at identifying direct illocutionary relations between L and I nodes. The second stage identifies argument relations between I nodes. Finally, the third stage detects indirect illocutionary relations between TA nodes and I nodes. In all the three stages, the text information included in L and I nodes is combined with a specifically curated prompt as the input to the developed models. Team KnowComp ran their experiments considering DeBERTa-base, DeBERTa-large, RoBERTa, and ALBERT (Lan et al., 2019). The best results were observed with the DeBERTa-based model architecture, which was the one selected for the final submission.

**Pokemon** (Zheng et al., 2024) proposes a two-stage pipeline to identify both argument relations with a two-step model filtering relations and classifying them into one of the three classes, and illocutionary relations with an eleven-class classification model covering all the possible YA nodes. The authors experiment with three different model architectures to implement the proposed pipeline, DeBERTa, RoBERTa, and LLaMa (Touvron et al., 2023). Based on the reported experimentation, their final submission consists of a DeBERTa-base combined with a RoBERTa-MNLI for the first stage (two-step) model, and a DeBERTa-large for the second stage model, meaning that RoBERTa-MNLI worked better for argument relation classification and DeBERTa-large for illocutionary relation identification.

**Pungene** (Chaixanien et al., 2024) approach the shared task with a pipeline system consisting of three main parts: the first part focuses on identifying illocutionary forces between locutions and propositions, the second part detects argumentative relations between propositions, and the third part identifies YA nodes between transitions and S nodes. This way, the proposed system gradually reconstructs the argument map by adding relation nodes. For the first part, the proposed system calculates BERTscore between pairs of locutions and propositions to establish the pairs, and then classifies the pair into one of the illocutionary relation classes using a fine-tuned BERT-base model. The second part of the proposed system connects argument propositions and detects the relation type between them by fine-tuning a BERT model for multi-class classification. Finally, the third part establishes the connection between transitions and S nodes by considering the natural language context of the nodes involved in the transition and the argumentative relation.

**Turiya** (Saha and Srihari, 2024) investigates two methods for argumentative dialogue analysis. First, by training a classification model using RoBERTa embeddings and two biaffine classifiers (Dozat and Manning, 2016). The first biaffine classifier is in charge of determining the relationship between argument propositions, between locutions and propositions, and between transitions and propositions. The second biaffine classifier is then trained to identify the remaining set of relations, the ones existing between transitions and propositions. The second explored method consists on leveraging the capabilities of generative LLMs to identify all the relations by prompting the language model with all the information extracted from the argument maps to generate an output pointing out all the potential relations between the nodes included in the map. From the reported results, it is possible to observe how in the general evaluation the LLMs perform better, but for the focused evluation combining both methods provides better results.

The sixth team, **misaka**, did not submit a system description paper.

## 5 Results

In order to provide an insightful analysis of the performance of the submitted systems to the DialAM-2024 shared task, we have divided the evaluation into three parts. First, the evaluation of the submitted systems when identifying propositional relations. Second, the evaluation of the submitted systems when identifying illocutionary relations. Finally, a global evaluation of the submitted systems when reconstructing argument maps looking at both, argument and discourse structures together. Furthermore, each evaluation is also done considering two different setups: by considering exclusively the related pairs of nodes in the evaluation maps (i.e., *focused*), and by considering the complete map including non-related nodes (i.e., *general*).

### 5.1 Propositional Relation Evaluation

The final results of the propositional relation evaluation, also known in the argument mining community as argument relation identification (ARI), have been described in Table 1.

Regarding the performance of the submitted systems on the specific aspect of identifying propositional relations, we observed that, in the *focused* setup POKEMON team was the best, while in the

| Model | Rank | Precision | Recall | F1-score |
|---|---|---|---|---|
| Pokemon | 1st | **46.26** | **32.43** | **35.89** |
| dfki-mlst | 2nd | 43.87 | 24.82 | 30.40 |
| RoBERTa-BL | 3rd | 37.10 | 18.42 | 22.80 |
| Pungene | 4th | 30.18 | 17.59 | 20.51 |
| KnowComp | 5th | 23.47 | 5.85 | 9.06 |
| misaka | 5th | 23.47 | 5.85 | 9.06 |
| Turiya | 7th | 18.95 | 4.21 | 6.65 |
| Majority-BL | 8th | 0 | 0 | 0 |
| dfki-mlst | 1st | **61.96** | **53.30** | **55.33** |
| Pungene | 2nd | 49.21 | 46.32 | 46.22 |
| KnowComp | 3rd | 32.43 | 33.79 | 32.75 |
| misaka | 3rd | 32.43 | 33.79 | 32.75 |
| Turiya | 5th | 30.81 | 31.52 | 30.75 |
| Pokemon | 6th | 32.00 | 46.56 | 30.64 |
| Majority-BL | 7th | 28.79 | 30.28 | 29.52 |
| RoBERTa-BL | 8th | 28.59 | 34.69 | 26.46 |

Table 1: Results of the ARI evaluation. First half reports the *focused* evaluation setup and second half the *general* setup.

| Model | Rank | Precision | Recall | F1-score |
|---|---|---|---|---|
| RoBERTa-BL | 1st | **73.10** | **72.55** | **72.09** |
| Pungene | 2nd | 71.18 | 69.23 | 69.95 |
| dfki-mlst | 3rd | 69.12 | 66.25 | 66.10 |
| Pokemon | 4th | 54.15 | 49.87 | 51.39 |
| KnowComp | 5th | 48.44 | 41.27 | 44.33 |
| misaka | 5th | 48.44 | 41.27 | 44.33 |
| Turiya | 7th | 43.81 | 26.09 | 30.41 |
| Majority-BL | 8th | 0 | 0 | 0 |
| Pungene | 1st | 81.99 | **80.79** | **81.17** |
| KnowComp | 2nd | **82.35** | 76.26 | 78.90 |
| misaka | 2nd | **82.35** | 76.26 | 78.90 |
| dfki-mlst | 4th | 81.08 | 79.25 | 78.78 |
| Pokemon | 5th | 56.41 | 64.57 | 59.36 |
| Turiya | 6th | 51.37 | 57.05 | 53.31 |
| RoBERTa-BL | 7th | 39.11 | 62.07 | 45.75 |
| Majority-BL | 8th | 34.71 | 35.90 | 35.29 |

Table 2: Results of the ILO evaluation. First half reports the *focused* evaluation setup and second half the *general* setup.

*general* setup dfki-mlst outperformed the others. From the ARI results, it is also possible to observe that systems performed much better in the *general* setup than in the *focused* setup, meaning that most of them estimated that more argument propositions are not related than related, which was not the case. Furthermore, this part of the shared task was also the most challenging one, achieving significantly lower performance scores than in the illocutionary relation identification task.

### 5.2 Illocutionary Relation Evaluation

The final results of the illocutionary relation evaluation (ILO) are summarised in Table 2.

It is interesting to observe how, in the *focused* setup, the RoBERTa-large baseline performed the best, but in the *general* setup was one of the worst systems. This is mostly due to the fact that this baseline does not correctly model the non-related pairs of sequences. Due to this, and the high class imbalance where assertions represent the majority of illocutionary relations, the model obtains good results when only looking at the set of related nodes but performs poorly when considering the complete argument maps, being not the best option for illocutionary relation identification. The best submission in this sub-task was Pungene, providing consistent strong results in both *focused* and *general* evaluation setups, followed by dfki-mlst. KnowComp and misaka performed well in the *general* setup, but their performance significantly dropped in the *focused* evaluation, contrary to the baseline. This means that these systems modelled better the non-related locution-proposition pairs, but missed a lot of the existing illocutionary relations.

### 5.3 Global Results

| Model | Rank | Precision | Recall | F1-score |
|---|---|---|---|---|
| dfki-mlst | 1st | **56.50** | **45.53** | **48.25** |
| RoBERTa-BL | 2nd | 55.1 | 45.49 | 47.45 |
| Pungene | 3rd | 50.68 | 43.41 | 45.23 |
| Pokemon | 4th | 50.20 | 41.15 | 43.64 |
| KnowComp | 5th | 35.95 | 23.56 | 26.70 |
| misaka | 5th | 35.95 | 23.56 | 26.70 |
| Turiya | 7th | 31.38 | 15.15 | 18.53 |
| Majority-BL | 8th | 0 | 0 | 0 |
| dfki-mlst | 1st | **71.52** | **66.28** | **67.05** |
| Pungene | 2nd | 65.60 | 63.55 | 63.70 |
| KnowComp | 3rd | 57.39 | 55.03 | 55.82 |
| misaka | 3rd | 57.39 | 55.03 | 55.82 |
| Pokemon | 5th | 44.20 | 55.57 | 45.00 |
| Turiya | 6th | 41.09 | 44.29 | 42.03 |
| RoBERTa-BL | 7th | 33.85 | 48.38 | 36.10 |
| Majority-BL | 8th | 31.75 | 33.09 | 32.40 |

Table 3: Results of the Global evaluation. First half reports the *focused* evaluation setup and second half the *general* setup.

The global results of the DialAM-2024 shared task were calculated by aggregating the performance of the systems in tasks A and B. The final results can be observed in Table 3.

The best overall system was the one submitted by team dfki-mlst, with 48.25 and 67.05 F1-scores

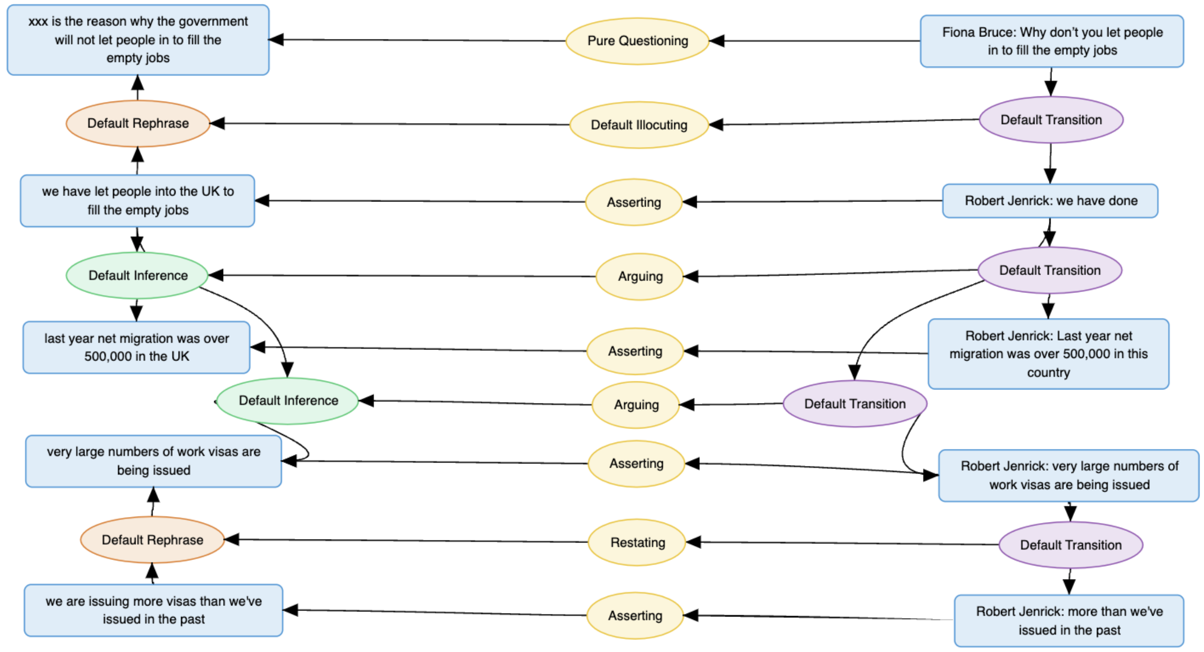
xxx is the reason why the government will not let people in to fill the empty jobs
Pure Questioning
Fiona Bruce: Why don't you let people in to fill the empty jobs
Default Rephrase
Default Illocuting
Default Transition
we have let people into the UK to fill the empty jobs
Asserting
Robert Jenrick: we have done
Default Inference
Arguing
Default Transition
last year net migration was over 500,000 in the UK
Asserting
Robert Jenrick: Last year net migration was over 500,000 in this country
Default Inference
Arguing
Default Transition
very large numbers of work visas are being issued
Asserting
Robert Jenrick: very large numbers of work visas are being issued
Default Rephrase
Restating
Default Transition
we are issuing more visas than we've issued in the past
Asserting
Robert Jenrick: more than we've issued in the past

Figure 2: The gold map number 6 that correctly shows a divergent argument, with "we have let people into the UK to fill the empty jobs" as the premise to two conclusions.

in the *focused* and *general* setups respectively. This represents an improvement of 0.8% against the RoBERTa-large baseline and 3.02% against the best competitor in the *focused* evaluation, and an improvement of 3.35% against the best competitor, PUNGENE, in the general evaluation setup. Again, the RoBERTa baseline performed overall well in the focused setup, but was one of the worst systems in the *general* evaluation, only better than the majority baseline. This means that the systems submitted by the other teams, although they did not beat the RoBERTa-large baseline in the *focused* setup, will be better options for argument mining in dialogues reconstructing argument maps due to their significantly better results in the *general* evaluation.

## 6 Qualitative Analysis

To expand the findings observed in the analysis of the results based on the performance scores achieved by the participants, we carried out a qualitative analysis looking at the content of the submitted argument maps leading us to interesting observations. For that purpose, we compared the maps generated by the submitted systems with the eleven gold standard maps included in our test set focusing on specific aspects that influenced the performance of the systems including conflict relations or more complex argument structures such as convergent (i.e., Figure 3), divergent (i.e., Figure 2), and linked arguments (i.e., Figure 3).

In general, we observed that the submitted argument mining systems had problems recognising conflict relations, failing to identify most of them, and assigning conflicts between non-conflicting propositions. We also observed, in line with the previously reported results that, teams DFKI-MLST and PUNGENE were the ones that produced the most similar outputs compared to the gold standard maps. It was also interesting to observe how, the maps produced by team POKEMON's system contained a significantly larger amount of relations compared to the rest. This is the reason of their higher scores in the *focused* evaluation with a significant drop of performance in the *general* setup.

Although both systems had some problems detecting more complex argument structures, we observed that DFKI-MLST did a better job than PUNGENE on these ones, identifying more convergent, divergent, and linked arguments. In the case of convergent and divergent arguments, the directionality of the relations is fundamental, making the identification of inference relations more challenging. For example, in the test map number 3, DFKI-MLST identified correctly the inference relations, but failed to correctly place the one making the argument divergent. On the other hand, PUNGENE had a rephrase instead of an inference, and the

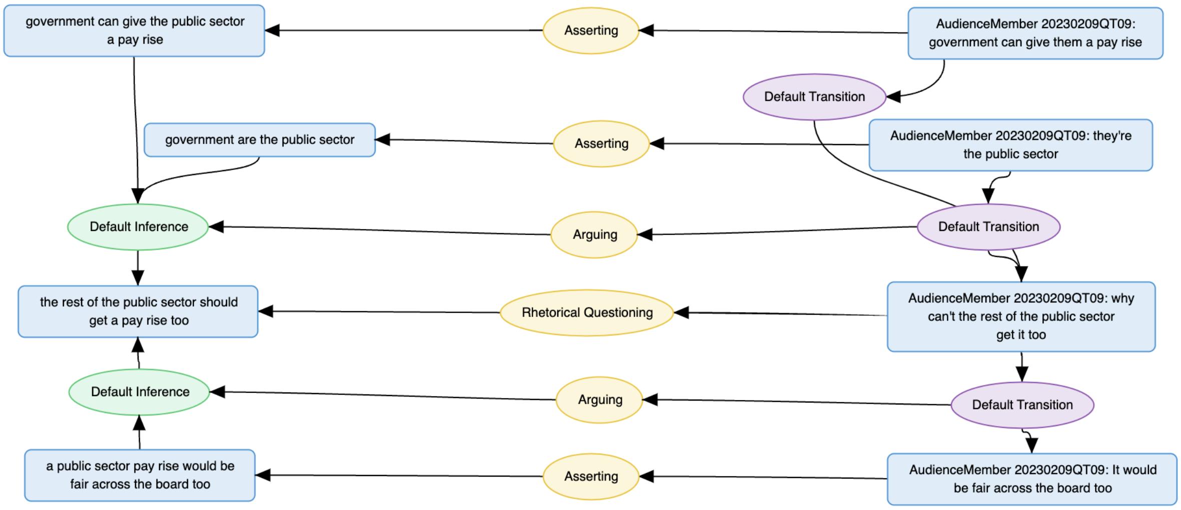
government can give the public sector a pay rise
Asserting
AudienceMember 20230209QT09: government can give them a pay rise
Default Transition
government are the public sector
Asserting
AudienceMember 20230209QT09: they're the public sector
Default Inference
Arguing
Default Transition
the rest of the public sector should get a pay rise too
Rhetorical Questioning
AudienceMember 20230209QT09: why can't the rest of the public sector get it too
Default Inference
Arguing
Default Transition
a public sector pay rise would be fair across the board too
Asserting
AudienceMember 20230209QT09: It would be fair across the board too

Figure 3: The gold map number 5 that correctly shows a convergent argument, with "the rest of the public sector should get a pay rise too" as the conclusion to two premises, where one of the supporting premises a linked argument consisting of two more premises.

assigned directionality made the argument convergent instead of divergent, a similar error was observed in the test map number 4. In the test map number 5, DFKI-MLST captures correctly the structure of the convergent argument, but fails to identify the two linked arguments. When looking at test map number 6, we observed again that both DFKI-MLST and PUNGENE had problems modelling the correct direction of the inference relations, identifying a convergent argument instead of the divergent one existing in the map, as seen within Figures 2 and 4.

With respect to linked arguments, team DFKI-MLST identified them in the test maps 7, 8, and 10, but represented them as convergent arguments instead of linked. A linked argument is represented by a unique inference relation including multiple premises rather than multiple inference relations between the premises and the claim, which would make it a convergent argument. We found, probably due to the implementation of the submitted systems, that this specific case was never considered as an output. In the test map number 10, we included a long linked argument, which consisted of six premises linked together in a unique inference relation towards the claim of the argument. Although the linked relation was not correctly represented (it was modelled as a convergent argument instead), it is interesting that DFKI-MLST correctly identified the six premises supporting the claim in this very particular case.

Finally, we also observed that in the evaluation set, none of the submitted systems was able to capture reported speech connecting locution nodes with illocutionary relations.

## 7 Conclusion

This paper presents DialAM-2024: the first shared task in dialogical argument mining. From the final results, we have been able to observe how the submitted systems that performed better in the DialAM-2024 shared task either addressed both tasks at the same time (modelling argumentative and dialogical features altogether), or first focused on task B and then task A, showing that considering speech acts and dialogical structures helped to improve the performance in the overall reconstruction of argument maps. Furthermore, from our qualitative analysis of the best submissions, we observed that there is still room for improvement in this area, specifically regarding the complex argument structures of convergent, divergent, and linked arguments, where not only the type of relation (i.e., inference) but also its directionality is of utmost important. It was also interesting to observe how, although the illocutionary relations were modelled with a reasonable success, specific cases such as reported speech represented a challenge for the systems submitted to the task.

Therefore, with the DialAM-2024 shared task, it is possible to observe the complexity of argument mining from a new dimension, pointing future work towards a more complete modelling of argumentation, including illocutionary forces and complex

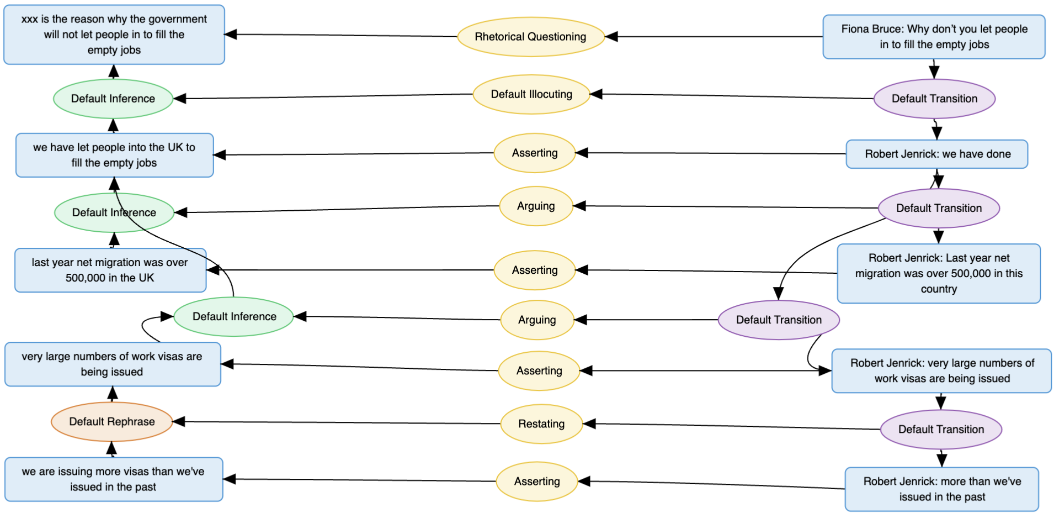
xxx is the reason why the government will not let people in to fill the empty jobs
Rhetorical Questioning
Fiona Bruce: Why don't you let people in to fill the empty jobs
Default Inference
Default Illocuting
Default Transition
we have let people into the UK to fill the empty jobs
Asserting
Robert Jenrick: we have done
Default Inference
Arguing
Default Transition
last year net migration was over 500,000 in the UK
Asserting
Robert Jenrick: Last year net migration was over 500,000 in this country
Default Inference
Arguing
Default Transition
very large numbers of work visas are being issued
Asserting
Robert Jenrick: very large numbers of work visas are being issued
Default Rephrase
Restating
Default Transition
we are issuing more visas than we've issued in the past
Asserting
Robert Jenrick: more than we've issued in the past

Figure 4: Pungene's model incorrectly identifying a convergent argument with "we have let people into the UK to fill the empty jobs" as the conclusion.

argument structures.

## Acknowledgements

This research is supported in part by: Volkswagen Stiftung under grant Az. 98 543; 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 Swiss National Science Foundation under grant 10001FM_200857; and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract 2022-22072200004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

## References

Arne Binder, Tatiana Anikina, Leonhard Henning, and Simon Ostermann. 2024. Dfki-mlst at dialam-2024 shared task: System description. In *Proceedings of the 11th Workshop on Argument Mining*, Thailand. Association for Computational Linguistics.

Katarzyna Budzynska and Chris Reed. 2011. Whence inference. *University of Dundee Technical Report*.

Sirawut Chaixanien, Eugene Choi, Shaden Shaar, and Claire Cardie. 2024. Pungene at dialam-2024: Identification of propositional and illocutionary relations. In *Proceedings of the 11th Workshop on Argument Mining*, Thailand. Association for Computational Linguistics.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

R. Duthie, J. Lawrence, K Budzynska, and C. Reed. 2016. The cass technique for evaluating the performance of argument mining. In *Proceedings of the Third Workshop on Argumentation Mining*, Berlin. Association for Computational Linguistics.

Kamila Górska, John Lawrence, and Chris Reed. 2024. Forecast2023: A forecast and reasoning corpus of argumentation structures. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7395–7405.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

David Hitchcock. 2009. The culture of spoken arguments. In *Proceedings of the 8th OSSA Conference*.

Yufang Hou and Charles Jochim. 2017. Argument relation classification using a joint inference model.

In *Proceedings of the 4th Workshop on Argument Mining*, pages 60–66.

M. Janier, J. Lawrence, and C Reed. 2014. OVA+: An argument analysis interface. In *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014)*, pages 463–464, Pitlochry. IOS Press.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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

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*.

Daniel J O'Keefe. 1977. Two concepts of argument. *The Journal of the American Forensic Association*, 13(3):121–128.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578.

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.

Ameer Saadat-Yazdi, Jeff Z Pan, and Nadin Kökciyan. 2023. Uncovering implicit inferences for improved relational argument mining. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2484–2495.

Sougata Saha and Rohini Srihari. 2024. Turiya at dialam-2024: Inference anchoring theory based llm parsers. In *Proceedings of the 11th Workshop on Argument Mining*, Thailand. Association for Computational Linguistics.

Ella Schad, Jacky Visser, and Chris Reed. 2024. The rip corpus of collaborative hypothesis-making. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16047–16057.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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.

Yuetong Wu, Yukai Zhou, Baixuan Xu, Weiqi Wang, and Yangqiu Song. 2024. Knowcomp at dialam-2024: Fine-tuning pre-trained language models for dialogical argument mining with inference anchoring theory. In *Proceedings of the 11th Workshop on Argument Mining*, Thailand. Association for Computational Linguistics.

Zihao Zheng, Zhaowei Wang, Qing Zong, and Yangqiu Song. 2024. Knowcomp pokemon team at dialam-2024: A two-stage pipeline for detecting relations in dialogue argument mining. In *Proceedings of the 11th Workshop on Argument Mining*, Thailand. Association for Computational Linguistics.