Dialo-AP: A Dependency Parsing Based Argument Parser for Dialogues

Sougata Saha, Souvik Das, Rohini Srihari
State University of New York at Buffalo
Department of Computer Science and Engineering
{sougatas, souvikda, rohini}@buffalo.edu

Abstract

While neural approaches to argument mining (AM) have advanced considerably, most of the recent work has been limited to parsing monologues. With an urgent interest in the use of conversational agents for broader societal applications, there is a need to advance the state-of-the-art in argument parsers for dialogues. This enables progress towards more purposeful conversations involving persuasion, debate and deliberation. This paper discusses Dialo-AP, an end-to-end argument parser that constructs argument graphs from dialogues. We formulate AM as dependency parsing of elementary and argumentative discourse units; the system is trained using extensive pre-training and curriculum learning comprising nine diverse corpora. Dialo-AP is capable of generating argument graphs from dialogues by performing all subtasks of AM. Compared to existing state-of-the-art baselines, Dialo-AP achieves significant improvements across all tasks, which is further validated through rigorous human evaluation.

1 Introduction

Argumentation is the process of reasoning systematically in support of an idea, action, or theory. It is prevalent in our daily communication and conversations, including online conversations. Since argumentation represents an intrinsic human attribute, the ability of artificial agents (bots) to exhibit this skill can be seen as strong evidence for judging such agents as "human-like". While computational models of argumentation have been investigated (Bench-Capon and Dunne, 2007; Rahwan and Simari, 2009; Atkinson et al., 2017), current progress is impeded by the scarcity of large scale corpora exemplifying use of argumentation and reasoning patterns from discourse. Such corpora are necessary if we are to make progress in training argumentative *conversational agents*. In this paper we experiment with computational argumentation mining (AM) (Mochales and Moens, 2011; Lippi and Torroni, 2016; Lawrence and Reed, 2019) for automatically analyzing discourse at a pragmatics level, and parsing argumentation structures from dialogues. Furthermore, although considerable research can be found in the field of AM, most of the current work has focused on parsing monologues (*micro-level* models), while neglecting dialogues (*macro-level* models) (Bentahar et al., 2010; Grasso, 2002). Here, we aim to fill the research gap in dialogical models for AM by proposing Dialo-AP, a novel end-to-end argument parser for dialogues.

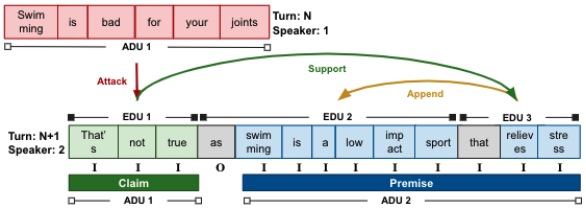


Figure 1: Dependency representation of dialogical arguments, across two turns.

Arguments primarily comprise *claims* and *premises*, with the claim being the central controversial statement of an argument, and the premise provides reasoning by supporting or attacking the claims (Stab and Gurevych, 2014b). End-to-end AM for dialogues generally involves performing text segmentation, component classification, and intra/inter-turn relation detection & classification. Formulating AM as dependency parsing (DP) problem operating at a mixture of elementary discourse unit (EDU) and argumentative discourse unit (ADU) granularity, Dialo-AP is an end-to-end argument parser which takes as input entire conversations, and outputs an argument graph comprised of arguments and relations. Figure 1 illustrates our DP formulation, where speaker 1's utterance consisting of the ADU "Swimming is bad for your joints" is *attacked* by speaker 2's claim "That's not true", which in turn is *supported* by the premise

Proceedings of the 29th International Conference on Computational Linguistics, pages 887–901
October 12–17, 2022.

comprising EDUs “it’s a low impact sport" and “it reduces stress".

Trained on the annotated dialogical *ChangeMyView* (CMV) corpus released by Chakrabarty et al. (2019b), and further utilizing robust pretraining on large scale parallel corpora, followed by fine-tuning on diverse argumentation datasets using *curriculum learning*, Dialo-AP attains significantly higher results compared to internal and external baselines for both in and out-of-domain examples.

2 Related Work

Significant advancements have been made in computational model for AM in recent years. Stab and Gurevych (2014c) implemented a feature engineering based pipelined approach for performing all four sub-tasks of AM, on the Persuasive Essays (PE) corpus (Stab and Gurevych, 2014a), which was further improved by the Integer Linear Programming (ILP) based approach proposed by Persing and Ng (2016). Stab and Gurevych (2017) introduced a larger version of the PE corpus and implemented an ILP constrained pipelined approach for AM. Mirko et al. (2020) improved upon the pipelined approach for AM introduced by Nguyen and Litman (2018), and further implemented a novel graph construction process to create argument graphs. Recently, Bao et al. (2021) proposed a neural transition-based model for component classification and relationship detection, which incrementally builds an argumentation graph by generating a sequence of actions, and can handle both tree and non-tree argumentation structures.

Eger et al. (2017) formulated the tasks of AM as a token level DP, and achieved state-of-the-art performance on the PE dataset, using a neural dependency parser. Inspired by the success of incorporating biaffine classifiers for semantic DP (Dozat and Manning, 2016, 2018), Ye and Teufel (2021) further improved the DP based approach by using biaffine layers, and leveraged pre-trained BERT (Devlin et al., 2018) for richer argument representations. Instead of operating at a word level, Morio et al. (2020) experimented with proposition level AM and used a joint learning framework for jointly performing the tasks of component classification, relation detection and classification. For AM in dialogues, Chakrabarty et al. (2019b) proposed Ampersand (AMP), a computational model for AM in online persuasive discussion forums.

Considerable work has also been done in trying to establish relationships between ADUs and EDUs. Peldszus (2015); Peldszus and Stede (2016); Musi et al. (2018); Hewett et al. (2019) studied the mapping from discourse structure from Rhetorical Structure Theory (RST) to argumentation structures and showed that discourse relations from RST often correlate with argumentative relations.

3 Methods

Formulating AM as dependency parsing, we introduce a multi-task learning (MTL) framework, where unlike existing pipelined approaches, all the sub-tasks are trained together in an end-to-end fashion. Since large scale annotated data for AM from dialogues is scarce, we augment existing monological datasets for our purpose, and leverage pretraining and curriculum learning to learn from the available datasets, before fine tuning on the target CMV corpus.

3.1 Dependency Representation of Arguments

Inspired by the token level dependency representations of arguments in monologues by Eger et al. (2017) and Ye and Teufel (2021), we formulate the following EDU level dependency representation for dialogues (Figure 1), encompassing all the sub-tasks for AM:

Text Segmentation & Component Classification: An argument (ADU) comprises fully or partially overlapping EDUs, which in turn contains labeled argumentative/non-argumentative tokens, using the IO tagging scheme. Identifying such EDUs by predicting the token tags, and further combining consecutive EDUs into ADUs by predicting the existence of relationship constitutes performing the sub-task of *text segmentation*. For example in Figure 1, EDU 2 in turn N+1 partially overlaps with ADU 2, as the token “as” is tagged as “O”, whereas the EDUs 1 and 3 fully overlap with ADU 1 and 2 respectively, which is indicated by all the tokens in the EDUs labelled as “I”. Further, EDU 2 and 3 can be combined using the “Append” relationship to construct ADU 2, after removing the non-argumentative token “as” (marked as O). Each EDU can belong to 1 of 4 classes $\in$ [**M**ajor **C**laim (**MC**), **C**laim (**C**), **P**remise (**P**), **N**on **A**rgument (**NA**)], and predicting the type of a constituent EDU constitutes performing *component classification*.

Intra/Inter-turn Relation Detection & Classification: Within a speaker’s turn, ADUs are related

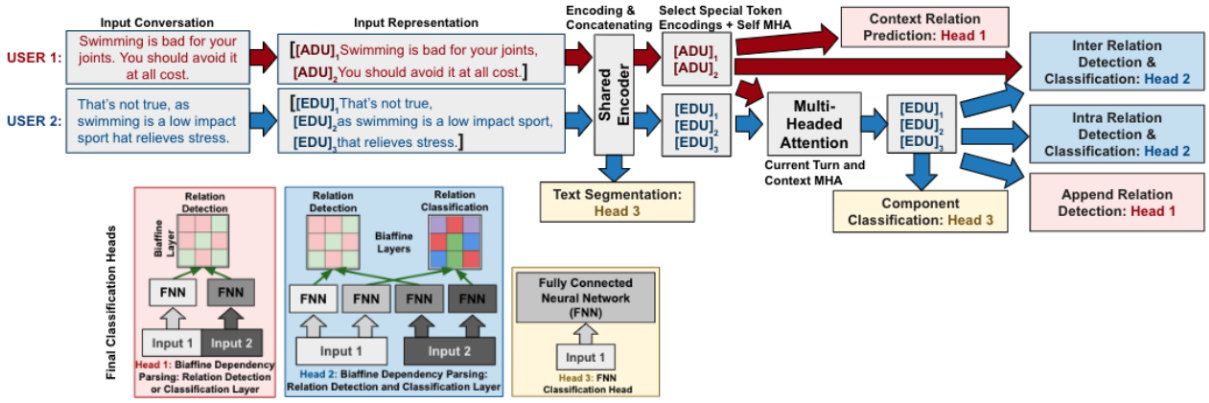


Figure 2: End-to-end Model Architecture.

using "**Sup**port" (**Sup**) or "**Att**ack" (**Att**) relationships which originate from the last EDU of the parent and terminate in the last EDU of the child ADU. Predicting the existence of such relationship between EDUs and further labeling it comprises the sub-tasks *intra-turn relation detection & classification*. Across turns inter-speaker support and attack relationships are established by relating and labeling (using Sup or Att) the last EDU of the source ADU (child) from the current turn, with the target ADU (parent) from the context. *inter-turn relation detection & classification* encompasses determining and labeling such relationships.

Thus, in each turn, ADUs constitute EDUs which are related using directed "Append" edges between consecutive EDUs. Support and attack relationship between ADU pairs are established by associating a labeled directed edge between the last EDUs from the origin and target ADU. The arguments in each turn are further parented by the contextual ADUs by associating a labeled directed edge originating from the contextual ADU, and terminating in the last EDU of the supporting/attacking ADU in the current turn.

3.2 Model Architecture

Dialo-AP is trained in a multi-task setting, where all the sub-tasks share a common encoded representation followed by task-specific layers. Figure 2 illustrates our architecture in detail[1].

Input Representation The model inputs EDU segments for the current turn, which are delimited by a special *[EDU]* token, which not only signifies the start of an EDU span, but is also responsible for encoding and representing its meaning. Turns with length greater than 300 tokens are split into shorter sequences of maximum 300 tokens, while ensuring that an EDU does not span multiple splits. The model inputs a list of ADU spans from prior turns as context, which unlike the current turn is not segmented to an EDU level, and always starts with the start of sequence (*sos*) token.

Encoding We use a shared transformer encoder to independently encode the current turn tokens $\mathrm{S}^{\mathrm{i}}_{\mathrm{curr}}$ and the context $\mathrm{S}^{\mathrm{i}}_{\mathrm{ctx}}$. Lengthy turns which are split into shorter sequences are sequentially encoded and concatenated into a single representation $\mathrm{E}_{\mathrm{curr}}$ (Eqn. 1). In order to preserve the temporal aspect of the text across splits, the position ids of the tokens in each split are cumulatively incremented after every encoding step. The context tokens are also encoded using the same encoder, which yields the context representation $\mathrm{E}_{\mathrm{ctx}}$ (Eqn. 2).

Post encoding, the final context representation $\mathrm{E}^{\mathrm{SOS}}_{\mathrm{ctx}}$ is obtained by selecting and concatenating the sos token encodings of the context ADUs, followed by a multi-headed self-attention layer mha with dropout $drop$ (Eqn. 3, 4). The final current turn representation $\mathrm{E}^{\mathrm{EDU}}_{\mathrm{curr}}$ is constructed by selecting the encodings of the *[EDU]* tokens, followed by a multi-headed self-attention layer and a multi-headed cross-attention between the current turn *[EDU]* token encodings and $\mathrm{E}^{\mathrm{SOS}}_{\mathrm{ctx}}$ (Eqn. 5, 6, 7).

$$\mathrm{E}_{\mathrm{curr}}=\mathrm{concat}(\mathrm{enc}(\mathrm{S}^{\mathrm{i}}_{\mathrm{curr}})|^{\mathrm{n_{splits}}}_{\mathrm{i=1}}) \quad (1)$$

$$\mathrm{E}_{\mathrm{ctx}}=\mathrm{enc}(\mathrm{S}^{\mathrm{i}}_{\mathrm{ctx}})|^{\mathrm{n_{ctx}}}_{\mathrm{i=1}};\ \mathrm{get}(\mathrm{X},\mathrm{idx})=\mathrm{X}[\mathrm{idx},:] \quad (2)$$

$$\mathrm{E}^{\mathrm{SOS}}_{\mathrm{ctx}}=\mathrm{concat}(\mathrm{get}(\mathrm{E}_{\mathrm{ctx}},\mathrm{idx}_{\mathrm{SOS}})) \quad (3)$$

$$\mathrm{E}^{\mathrm{SOS}}_{\mathrm{ctx}}=\mathrm{E}^{\mathrm{SOS}}_{\mathrm{ctx}}+\mathrm{drop}(\mathrm{mha}(\mathrm{E}^{\mathrm{SOS}}_{\mathrm{ctx}},\mathrm{E}^{\mathrm{SOS}}_{\mathrm{ctx}})) \quad (4)$$

[1]Code and data: https://github.com/sougata-ub/dialo-ap

$$\mathrm{E}^{\mathrm{EDU}}_{\mathrm{curr}}=\mathrm{get}(\mathrm{E}_{\mathrm{curr}}, \mathrm{idx}_{\mathrm{EDU}}) \quad (5)$$

$$\mathrm{E}^{\mathrm{EDU}}_{\mathrm{curr}}=\mathrm{E}^{\mathrm{EDU}}_{\mathrm{curr}}+\mathrm{drop}(\mathrm{mha}(\mathrm{E}^{\mathrm{EDU}}_{\mathrm{curr}}, \mathrm{E}^{\mathrm{EDU}}_{\mathrm{curr}})) \quad (6)$$

$$\mathrm{E}^{\mathrm{EDU}}_{\mathrm{curr}}=\mathrm{E}^{\mathrm{EDU}}_{\mathrm{curr}}+\mathrm{drop}(\mathrm{mha}(\mathrm{E}^{\mathrm{EDU}}_{\mathrm{curr}}, \mathrm{E}^{\mathrm{SOS}}_{\mathrm{ctx}})) \quad (7)$$

$$\mathrm{Biaf}(x, y)=x^T \mathrm{U} y + \mathrm{W}(x \oplus y) + b \quad (8)$$

Task Specific Layers We incorporate task-specific layers to perform the final prediction for each sub-task. Illustrated in Figure 2, we use single-layered feed-forward neural networks (Head 3) as the final layer for both text segmentation and component classification, with the input for text segmentation being the concatenated current turn representation $\mathrm{E}_{\mathrm{curr}}$, and $\mathrm{E}^{\mathrm{EDU}}_{\mathrm{curr}}$ for component classification.

Biaffine classifiers (Eqn. 8) are generalizations of linear classifiers, which include multiplicative interactions between two vectors. Since relation detection and classification require performing inference over argument pairs, we implement biaffine dependency parsing (Head 2 and 3 in Figure 2) for both sub-tasks. For intra-relation prediction, the current turn EDU encodings $\mathrm{E}^{\mathrm{EDU}}_{\mathrm{curr}}$ are split into two parts using FNNs–a parent $\mathrm{H}^{\mathrm{i\_parent}}_{\mathrm{intra}}$ and a dependent child $\mathrm{H}^{\mathrm{i\_child}}_{\mathrm{intra}}$ representation, which in turn are passed through a biaffine classifier for detecting or labelling relationships between the EDUs (Eqn. 9, 10). For inter-relation prediction, the parent and child representations $\mathrm{H}^{\mathrm{i\_parent}}_{\mathrm{inter}}$ and $\mathrm{H}^{\mathrm{i\_child}}_{\mathrm{inter}}$ for the biaffine classifier are obtained by passing the context encoding $\mathrm{E}^{\mathrm{SOS}}_{\mathrm{ctx}}$ and current turn EDU encodings $\mathrm{E}^{\mathrm{EDU}}_{\mathrm{curr}}$ through FNNs respectively.

$$\mathrm{H}^{\mathrm{i\_j}}_{\mathrm{k}}=\mathrm{FNN}(x)|x \in (\mathrm{E}^{\mathrm{EDU}}_{\mathrm{curr}}, \mathrm{E}^{\mathrm{SOS}}_{\mathrm{ctx}}), \quad (9)$$
$$\mathrm{i} \in (\mathrm{detect}, \mathrm{label}), \mathrm{j} \in (\mathrm{parent}, \mathrm{child}),$$
$$\mathrm{k} \in (\mathrm{inter}, \mathrm{intra})$$

$$\mathrm{sc}^{\mathrm{i}}_{\mathrm{j}}=\mathrm{Biaf}(\mathrm{H}^{\mathrm{i\_parent}}_{\mathrm{k}}, \mathrm{H}^{\mathrm{i\_child}}_{\mathrm{k}})| \quad (10)$$
$$\mathrm{i} \in (\mathrm{detect}, \mathrm{label}), \mathrm{k} \in (\mathrm{inter}, \mathrm{intra})$$

The sub-tasks of append relation detection and the additional context relationship prediction are performed in a similar way to intra-relationship detection and labeling respectively, where $\mathrm{E}^{\mathrm{EDU}}_{\mathrm{curr}}$ is used for append relation detection, and $\mathrm{E}^{\mathrm{SOS}}_{\mathrm{ctx}}$ for labeling relationships between the context ADUs.

$$\mathcal{L}_{total}=\sum \lambda_x \mathcal{L}^x | x \in (\mathrm{subtasks}) \quad (11)$$

All the sub-tasks are jointly trained end-to-end by minimizing the aggregated interpolated loss $\mathcal{L}_{total}$ (Eqn. 11), where text segmentation, component classification, and inter/intra/contextual relationship labelling are trained by minimizing the cross entropy loss, whereas inter/intra/append relationship detection is trained by minimizing the binary cross entropy loss.

3.3 Pre-training

Since the size of the CMV corpus is small for modern deep learning approaches, we pre-train our parser for most sub-tasks, on large scale noisy labelled corpora.

Component & Intra-Turn Relation Prediction We use the **IMHO corpus** (Chakrabarty et al., 2019a) for pre-training the parser on the sub-tasks of *component classification*, *append relation detection*, and *intra/inter-turn relation detection*. The IMHO corpus comprises 5.5 million opinionated claims from Reddit, which are self-labeled by their authors using the internet acronyms IMO/IMHO (in my (humble) opinion). For example "IMO, Lakers are in big trouble next couple years. Their players are out of contract". We tokenize each example into sentences, and label a sentence as claim only if it contains the acronyms IMO/IMHO, and further associated with a noisy premise by choosing either the preceding or succeeding non-claim sentence, depending on which has a higher levenshtein distance based similarity with the claim tokens. Argument components are further segmented into EDUs, which we detail in Appendix A.2.1. The training targets constitute claim and premise labels, two binary relation matrices for predicting presence of argumentative and "Append" relations between EDUs, and a label matrix for predicting the "support/attack" relationship type.

Inter-Turn Relation Prediction We use the **args.me** (Ajjour et al., 2019), and **QR** corpus (Chakrabarty et al., 2019b) for pre-training the parser on the *inter-turn relation detection and classification* sub-tasks. The args.me corpus comprises 387,606 macro-level arguments crawled from diverse debate portals and already identifies source and target arguments along with pro/con stance labels, which we further convert to support/attack inter-turn relationships. The QR dataset comprises 97,636 pairs of original post and replies from the CMV sub Reddit, where the respondent used Reddit's "quote" feature to reply, signifying an attack relationship on the quoted section from the original post. We combine both the macro-level datasets consisting of source argument, target argument and the inter-argument relationship, and further generate 10,000 random argument pairs with "no re-

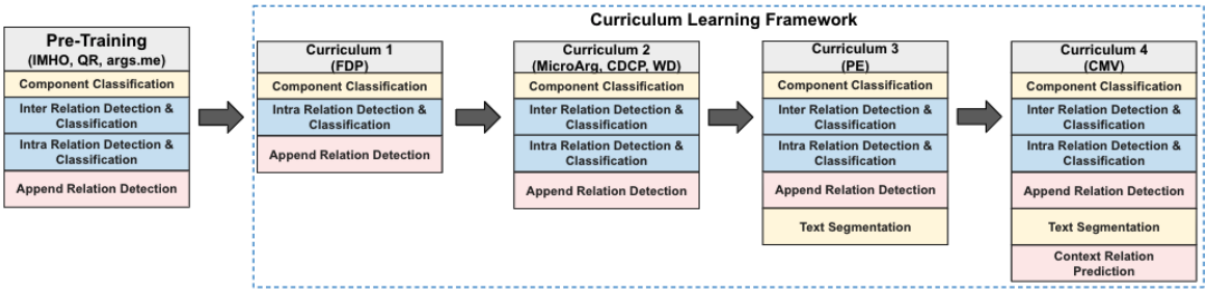


Figure 3: Curriculum Learning Framework.

lationship" labels. For constructing the training relationship label matrices, similar to the IMHO corpus processing we tokenize the source argument into EDUs. However, as discussed in sub-section 3.1, distinct from intra-turn relationship prediction the targets here are ADUs from prior turn, which we identify by using a version of Dialo-AP trained only on the processed IMHO corpus for extracting claims/premises from the context.

3.4 Curriculum Learning (CL) Framework

Computational AM being a relatively new field, suffers from the lack of large scale annotated data, specially for dialogues. Most of the available datasets pertain to distinct and diverse aspects of AM for monologues (Habernal and Gurevych, 2017). Although monologues are distinct from dialogues, parsing both the forms of discourse entails performing comparable tasks like text segmentation, component classification, and intra-turn relationship prediction, which are more local in nature. Further, with a few adaptations, monologues can be augmented to mimic dialogues, and engender noisy training data for inter-turn relationship prediction. Inspired by these observations we design a curriculum learning framework, where we leverage diverse corpora spanning both monologues and dialogues pertaining for incrementally training our parser on all the AM sub-tasks. We define four curriculum spanning six distinct datasets, with the final curriculum comprising fine-tuning on the target CMV dataset. Figure 3 illustrates our curriculum learning framework. Further, all the datasets are pre-processed to conform with our defined dependency representation, which we detail in Appendix A.2.1 and A.2.2.

Curriculum 1 (C1): Component & Intra-Relation Prediction The first curriculum comprises training the component classification and intra-relation prediction sub-tasks, where we leverage the Feedback Prize Dataset[2] (**FDP**), which consists of 15,000 argumentative essays written by U.S students in grades 6-12, and were annotated by expert raters for elements commonly found in argumentative writing.

Curriculum 2 (C2): Component & Intra/Inter-Turn Prediction We train the resulting model from curriculum 1 on the component classification, intra and inter-turn relation prediction sub-tasks by combining training data from three existing corpora: (i) the argumentative microtext corpora (**MicroArg**) (Peldszus, 2015) featuring 112 short argumentative monologues, which were annotated with argumentation structures, following the scheme proposed in Peldszus and Stede (2013); (ii) Consumer Debt Collection Practices (**CDCP**) corpora (Park and Cardie, 2018), comprising 4,931 elementary unit and 1,221 support relation annotations; (iii) Web Discourse (**WD**) corpora (Habernal and Gurevych, 2017), comprising 340 documents annotated with the extended Toulmin model (Toulmin, 2003).

Curriculum 3 (C3): Text Segmentation, Component & Intra/Inter-Turn Relation Prediction Using the Persuasive Essays (Stab and Gurevych, 2017) (**PE**) corpus, we train the resultant model from curriculum 2 on the text segmentation, component classification, intra and inter-relation prediction tasks. PE comprises 402 randomly selected essays from an online forum, which are annotated with argumentation structures. Treating each paragraph as a turn, we convert the dataset to dialogues by considering the major claims as conversation context, and re-labeling the existing "for/against" relationship between a claim and major claim as inter-turn "support/attack" relationship.

Curriculum 4 (C4): Target Dataset Fine-tuning Finally, we fine-tune the resultant model from curriculum 3 on the Change My View (**CMV**) dataset

[2]https://www.kaggle.com/competitions/feedback-prize-2021/overview

(Chakrabarty et al., 2019b), for all sub-tasks. Consisting of 112 discussions, the CMV dataset expands the existing data collected by Hidey et al. (2017) by annotating both inter-turn and intra-turn relations, along with additional argument components. Further, in order to facilitate learning meaningful representations, we introduce an additional task during inter-turn relation prediction, where we train the model to predict relationships that exist between the contextual ADUs.

4 Experiments and Results

We use the CMV data (C4 in Section 3.4) for our experiments, and repeat each experiment five times. In each run, 10% of the data is randomly set aside for testing, and we report the average and maximum results across all runs. More details pertaining to model configuration and setup are shared in Appendix A.1. We train a baseline variant **Baseline-C4** by fine-tuning only on the CMV data (without pre-training and curriculum learning), and use as our internal baseline for model comparison. We further compare our implementation against the following strong external baselines, and report results.

AMP-BERT: Ampersand's (Chakrabarty et al., 2019b) (AMP) BERT baseline using only the pre-trained model without additional fine-tuning.

AMP-Pre-Train: AMP's fine-tuned models *IMHO Context Fine-Tuned BERT* for component classification and intra-turn relation prediction, and the *QR Context Fine-Tuned BERT* for inter-turn relation prediction.

AMP-Best: AMP's best models–*IMHO Context Fine-Tuned BERT* for component classification, *IMHO Context Fine-Tuned BERT+ RST Ensemble* for intra-turn relation prediction, and *IMHO Context Fine-Tuned BERT+ RST Features + Extractive Summarizer* for inter-turn relation prediction. Note that in order to facilitate uniform comparison across experiments, we disregard the variants that incorporate additional rule-based post processing.

AMP-Pre-Train-Re: Since the models reported in AMP are not available for public use, in order to perform qualitative analysis we re-create their fine-tuned models that incorporate pre-training (AMP-Pre-Train), for all common sub-tasks.

4.1 Quantitative Results

Component Classification Table 1 shares our results for component classification, where we report and compare F1 score against external and internal baselines. We observe that although external baselines perform better for identifying non arguments, our implementation (C1+C4) which is trained on curriculum 1 followed by fine-tuning on the target dataset, significantly outperforms all baselines for claim and premise classification, which is more beneficial for constructing argument graphs. We reason that since curriculum 1 constitutes learning only component classification and intra-relation prediction using the fairly large FDP dataset, it is better able to classify components due to lower cognitive load.

We also observe that training the model on all curricula (CL+C4) yields good results, which is not further improved by pre-training. We attribute this to the noisy nature of the pre-training data for component classification.

Inter/Intra Relation Detection Table 1 shares our results for both inter and intra-turn relation detection, where we compare F1, precision, and recall scores across models. In comparison to our internal baseline, for inter-relation detection we observe that training using a curriculum learning framework yields better results, specially for curriculum 2, which constitutes training the relation prediction sub-tasks using the MicroArg, CDCP and WD datasets. Furthermore, we observe higher F1 scores with pre-training, which is further increased by incorporating curriculum learning, yielding the best overall results for inter-relation detection. For intra-turn relation detection we obtain best overall F1 results when incorporating all curricula in our curriculum learning framework. Although pre-training does not seem to be further enhanced the intra-turn relation F1 score, it helps achieve a higher precision model, which can be useful depending on the intended use case of the parser.

For both inter and inter relation detection, we observe that in comparison to AMP based external baselines, our models yield higher precision, lower recall, and higher F1 scores. We also observe that in contrast to our models, which balances precision and recall, all AMP variants generally have disproportionately higher recall compared to their precision. We attribute it to the fact that AMP formulates relationship detection as a binary prediction task between sentence pairs, and constructs all possible permutations of possible sentence pairs from text, which inadvertently spans all arguments, thus increasing recall. On the contrary, our biaffine

Model	Component Classification			Inter-Turn Relation Detection			Intra-Turn Relation Detection		
	Non-Arg	Claim	Premise	Precision	Recall	F1	Precision	Recall	F1
AMP-BERT	71.3	62.0	72.2	8.8	76.0	15.8	12.0	67.0	20.3
AMP-Pre-Train	-	-	-	11.0	75.3	19.1	14.3	69.0	23.7
AMP-Best	75.7	67.1	72.5	16.0	79.4	26.8	16.7	73.0	27.2
AMP-Pre-Train-Re	82.7	63.6	60.9	8.0	52.5	14.0	11.7	77.0	20.4
Baseline-C4 *	70.1 (77.5)	63.7 (71.9)	74.3 (80.4)	23.9 (40.4)	31.2 (39.6)	26.4 (37.7)	17.2 (23.1)	19.3 (29.6)	16.5 (21.5)
C1+C4	72.9 (77.3)	**68.5** (74.7)	**75.5** (82.6)	43.8 (72.9)	33.2 (42.5)	35.2 (42.1)	23.4 (31.8)	28.0 (44.0)	23.0 (29.4)
C2+C4	67.0 (73.4)	59.2 (67.9)	72.1 (76.1)	46.4 (56.9)	30.3 (38.8)	36.4 (44.0)	22.1 (40.3)	13.6 (27.8)	12.9 (24.1)
C3+C4	66.5 (71.8)	63.2 (65.8)	72.4 (76.0)	35.4 (49.6)	29.4 (38.8)	31.2 (40.9)	20.3 (36.9)	25.2 (31.5)	20.7 (26.0)
CL+C4	**73.2** (79.9)	68.4 (73.6)	75.1 (80.7)	33.0 (46.2)	**35.1** (46.3)	33.8 (44.1)	28.0 (37.0)	**34.3** (50.7)	**29.2** (37.9)
Pre-Train+C4	67.4 (77.7)	65.2 (71.8)	73.7 (80.3)	**63.9** (91.1)	27.4 (38.1)	38.4 (53.7)	12.9 (19.1)	16.8 (27.3)	14.1 (20.6)
Pre-Train+CL+C4	70.0 (76.1)	67.0 (71.5)	75.3 (80.1)	55.3 (70.3)	31.3 (43.3)	**39.7** (49.8)	**30.7** (45.7)	26.4 (31.8)	27.2 (37.5)

Table 1: Average and (maximum results) for Component Classification & Intra/Inter-Turn Relationship Detection. For each metric best results w.r.t internal baseline (*) is highlighted in bold, and overall best result underlined.

Model	Inter-Turn Rel. Classify		Intra-Turn Rel. Classify		Text Segmentation		Append
	Support	Attack	Support	Attack	Non-Arg	Arg	
Baseline-C4 *	74.9 (82.3)	66.0 (78.9)	97.8 (99.4)	53.0 (85.7)	77.3 (80.1)	89.1 (90.5)	18.5 (30.5)
C1+C4	76.9 (84.9)	71.3 (81.0)	98.3 (99.2)	50.3 (61.5)	78.1 (81.8)	89.3 (91.4)	59.2 (61.0)
C2+C4	75.1 (84.1)	69.5 (79.5)	98.7 (99.3)	56.3 (75.0)	74.0 (79.1)	87.9 (90.0)	15.7 (25.8)
C3+C4	67.6 (80.3)	60.5 (74.1)	98.2 (99.2)	48.9 (80.0)	75.7 (77.9)	87.9 (89.2)	36.3 (45.4)
CL+C4	**78.7** (85.2)	**78.2** (81.3)	98.9 (99.3)	58.0 (66.7)	**78.5** (83.9)	**89.5** (91.4)	62.5 (65.6)
Pre-Train+C4	75.6 (80.5)	77.3 (82.2)	**99.2** (99.7)	**60.3** (92.3)	75.5 (83.2)	87.9 (90.5)	61.5 (65.5)
Pre-Train+CL+C4	77.2 (86.3)	76.6 (81.5)	98.5 (99.3)	51.4 (80.0)	78.5 (82.4)	89.1 (90.1)	**81.2** (83.3)

Table 2: Average and (maximum) F1 scores for Inter/Intra-Turn Relationship Classification, Text Segmentation and Append relationship prediction. In each column, best result w.r.t baseline (*) is highlighted in bold.

Model	Component		Inter-Turn		Intra-Turn	
	TP-C	TP-A	TP-C	TP-A	TP-C	TP-A
AMP-Re	71.1	75.6	54.3	46.7	67.6	72.7
Dialo-AP	**82.2**	**80.4**	**87.5**	**90.9**	**71.3**	**73.2**

Table 3: Comparison of Human Evaluation Results between AMP-Pre-Train-Re (AMP-Re) and Pre-Train+CL+C4 (Dialo-AP)

dependency parsing based formulation operates at an entire turn level, and facilitates information exchange across EDUs, thus resulting in a balanced score.

Inter/Intra Relation Labeling, Text Segmentation and Append Detection We report our results for inter/intra-turn relationship label prediction and append relationship detection in Table 2. We only perform comparison amongst our implemented variants and inter baselines, due to lack of external baselines for these tasks. For inter-turn relation classification, we observe that incorporating both pre-training and curriculum learning yields better results compared to baseline. Further, training on all curricula yields best F1 score for predicting both support and attack relationship. Similarly for intra-turn relation classification, we observer both pre-training and curriculum learning yields superior results compared to baseline. However, compared to curriculum learning, incorporating pre-training yields better results. We also observe that compared to inter-turn, all model variants are perform intra-turn *support* relationship classification better and *attack* relationship classification worse, compared to inter-turn. Further, for each model, the difference in *support* and *attack* classification F1 scores for intra-turn is higher compared to inter-turn, signifying. We attribute this to the fact that occurrence of *support* relationships are more prevalent within a turn compared to *attack* relationships, which is the converse for inter-turn relationships (Table 4, Appendix A.3).

Although the task of text segmentation is more dependent on linguistic features, we observe best results (Table 2) when training using curriculum learning, proving the efficacy of training using diverse curriculum, in a multi-task learning framework. Also, for detecting *append* relationship between EDUs (Table 2), we observe significantly better results when incorporating pre-training along with curriculum learning, compared to other means.

4.2 Qualitative Results

We further perform human evaluations to ascertain Dialo-AP's usefulness in real world scenarios, where the topic of the discussion might be unre-

stricted. For our purpose, we collect discussion threads from the *ChangeMyView* subreddit on the controversial and out-of-domain topics of *abortion*, *gun violence*, *minimum wage* and *death penalty*, and perform human evaluation on the component classification, inter-turn and intra-turn relation detection subtasks, using a subset of 100 discussions (Table 5, Appendix A.3).

Since our motivation is to create a parser that can identify salient arguments with high precision, we introduced and compared two new metrics: (i) **TP-C:** Mean **T**rue **P**ositive rate at a **C**onversation level, signifying for a conversation, the number of model predictions that are correct on an average. (ii) **TP-A:** Mean **T**rue **P**ositive rate at an over**a**ll level, signifying on an average, the number of model predictions that are correct. We parse each discussion using Dialo-AP variant incorporating pre-training and curriculum learning (Pre-Train+CL+C4), and the recreated version of AMP that leverages pre-training (AMP-Pre-Train-Re), and use Amazon Mechanical Turk (AMT) Human Intelligence Task (HIT) to collect human evaluation on the parsed outputs. In each HIT we provide the entire discussion thread, followed by either the identified arguments with their predicted claim/premise labels, or inter/intra argument pairs predicted by the parsers, and ask the evaluators to mark (by ticking a checkbox) if they think the prediction is correct. Appendix A.4 details the human evaluation task and the AMT collection framework. We compute inter-annotator agreement using 2 evaluators, and observer a Cohen's Kappa score of 0.15, 0.22 and 0.16 for component classification, intra and inter-turn relationship detection respectively, signifying fair amount of agreement (Table 6, Appendix A.4).

Table 3 shares the results from the human evaluation. We observe that our formulation yields significantly better results for all three subtasks, with inter-turn relation detection reporting highest gains compared to the competing model. We attribute this to our robust pre-training and curriculum learning framework, which trains the parser on existing and augmented dialogical data, for identifying inter-turn argumentative relationships. In comparison to itself, leveraging pre-training on the monological IMHO and the noisy QR corpus, AMP performs best on component classification followed by intra-turn and inter-turn relation detection subtasks, whereas Dialo-AP performs best on inter-turn relation detection, followed by component classification and intra-turn relation detection subtasks. Thus, signifying Dialo-AP's better applicability for mining arguments from dialogues.

5 Discussion

Our aim with Dialo-AP was to devise an end-to-end argument parser that can not only enable discourse analysis, but also aid in argument generation by engendering *argument graphs* comprising salient (support-attack) chains of arguments from dialogues. In Figure 4 we illustrate an argument graph generated by Dialo-AP (on the right) on a randomly sampled CMV discussion on *death penalty*, and further compare it against the output by recreated AMP (on the left). Firstly, we observe that although both the parsers yield similar number of components, operating at a combination of EDU and token level, Dialo-AP better identifies and labels argumentative spans. For instance, AMP incorrectly labels the non-arguments *P5* and *P6* as premises. Further, operating at a sentence level, AMP classifies the component *C2* as a single claim, whereas Dialo-AP is correctly able to segment it into 2 components *P1* and *C2*.

Secondly, we observe that since AMP formulates relation prediction as a binary classification problem between argument pairs, it predicts copious relations between components which need not hold, thus hurting it's usefulness for constructing argument graphs. For example, none of the relationships predicted between user 2's argument components hold. On the contrary, not only are the relationships identified by Dialo-AP more meaningful, it also labels the relationship type, making it more useful for constructing argument graphs.

Although Dialo-AP yields better results, it comes with its own set of predicaments. As illustrated on the right, it is unable to relate and utilize all identified argumentative components in the argument graph. For instance, although Dialo-AP identifies the component *P4*, it is unable to establish relationship, and leaves them out of the graph. Further, we observe that both the parsers lack epistemic reasoning capabilities, and could possibly benefit from the use of external knowledge graphs and knowledge bases, which we point as the next possible research direction for argument parsing.

6 Conclusion

In this paper, we present Dialo-AP, a state-of-the-art end-to-end dependency parsing based argument

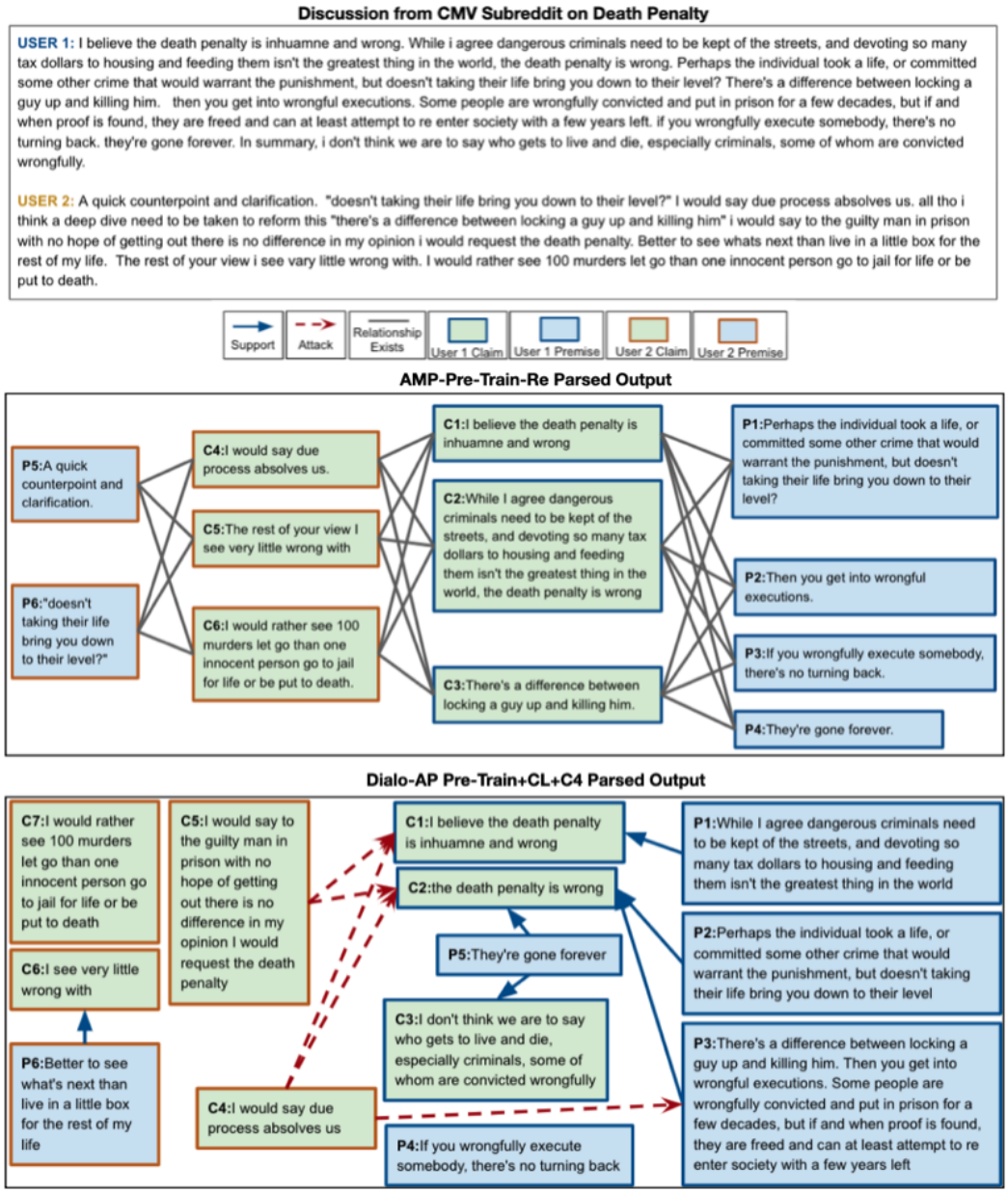


Figure 4: Comparison of parsed CMV post.

parser for parsing arguments from dialogues. Formulating AM as dependency parsing of EDUs and ADUs, and trained in a multi-task setting over diverse curriculum, Dialo-AP is capable of engendering argument graphs from dialogues, by performing all sub-tasks of AM. Dialo-AP's efficacy is exhibited by its superior experimental and human evaluation results, in comparison to strong internal and external baselines. We further discuss Dialo-AP's limitations, and point towards possible next research steps.

Acknowledgements

We thank the anonymous reviewers for providing valuable feedback on our manuscript. This work is partly supported by NSF grant number IIS2214070. The content in this paper is solely the responsibility of the authors and does not necessarily represent the official views of the funding entity.

References

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data Acquisition for Argument Search: The args.me corpus. In *42nd German Conference on*

Artificial Intelligence (KI 2019), pages 48–59, Berlin Heidelberg New York. Springer.

Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Simari, Matthias Thimm, and Serena Villata. 2017. Towards artificial argumentation. *AI magazine*, 38(3):25–36.

Jianzhu Bao, Chuang Fan, Jipeng Wu, Yixue Dang, Jiachen Du, and Ruifeng Xu. 2021. A neural transition-based model for argumentation mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6354–6364, Online. Association for Computational Linguistics.

Trevor JM Bench-Capon and Paul E Dunne. 2007. Argumentation in artificial intelligence. *Artificial intelligence*, 171(10-15):619–641.

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259.

Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. 2019a. IMHO fine-tuning improves claim detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 558–563, Minneapolis, Minnesota. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019b. AMPERSAND: Argument mining for PERSuAsive oNline discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Timothy Dozat and Christopher D. Manning. 2016. Deep biaffine attention for neural dependency parsing.

Timothy Dozat and Christopher D. Manning. 2018. Simpler but more accurate semantic dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490, Melbourne, Australia. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Floriana Grasso. 2002. Towards a framework for rhetorical argumentation. In *EDILOG 02: Proceedings of the 6th workshop on the semantics and pragmatics of dialogue*, pages 53–60.

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

Freya Hewett, Roshan Prakash Rane, Nina Harlacher, and Manfred Stede. 2019. The utility of discourse parsing features for predicting argumentation structure. In *Proceedings of the 6th Workshop on Argument Mining*, pages 98–103, Florence, Italy. Association for Computational Linguistics.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

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

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2).

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*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization.

LENZ Mirko, Premtim Sahitaj, Sean Kallenberg, Christopher Coors, Lorik Dumani, Ralf Schenkel, and Ralph Bergmann. 2020. Towards an argument mining pipeline transforming texts to argument graphs. *Computational Models of Argument: Proceedings of COMMA 2020*, 326:263.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, Yuta Koreeda, and Kohsuke Yanai. 2020. Towards better non-tree argument mining: Proposition-level biaffine parsing with task-specific parameterization. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3259–3266, Online. Association for Computational Linguistics.

Elena Musi, Manfred Stede, Leonard Kriese, Smaranda Muresan, and Andrea Rocci. 2018. A multi-layer annotated corpus of argumentative text: From argument schemes to discourse relations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Huy V. Nguyen and Diane J. Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In *AAAI*.

Joonsuk Park and Claire Cardie. 2018. A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Andreas Peldszus. 2015. An annotated corpus of argumentative microtexts.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):1–31.

Andreas Peldszus and Manfred Stede. 2016. Rhetorical structure and argumentation structure in monologue text. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 103–112, Berlin, Germany. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394, San Diego, California. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Iyad Rahwan and Guillermo R Simari. 2009. *Argumentation in artificial intelligence*, volume 47. Springer.

Christian Stab and Iryna Gurevych. 2014a. 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.

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014c. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967, Brussels, Belgium. Association for Computational Linguistics.

Yuxiao Ye and Simone Teufel. 2021. End-to-end argument mining as biaffine dependency parsing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 669–678, Online. Association for Computational Linguistics.

A Appendix

A.1 Experiment Setup

We use Roberta (base) (Liu et al., 2019) as the base encoder, and increase its embedding layer to accommodate the special *[EDU]* token. The size of positional embedding layer is increased to 2500. Two layers comprising four attention heads are used for MHA, where the MHA result in each layer is sum pooled with the residual output while applying dropout with 0.1 probability to the MHA result. The hidden size of the FNNs in the biaffine layer is set to 600. An interpolation factor of 0.4 each is used for aggregating the inter and intra-turn relation prediction losses, 0.1 for component classification, and 0.01 for the additional contextual relationship loss prediction. The remaining factor of 0.09 is split equally among text segmentation, append relation prediction, and inter/intra-turn relation labeling. Further, a weight of 3.0 is applied to positive examples during computing binary cross entropy loss for inter and intra-turn relation prediction. All models are trained with a learning rate of 1e-5 for 15 epochs and optimised using AdamW (Loshchilov and Hutter, 2017), with early stopping if the validation loss doesn't reduce for 2 epochs. We repeat each experiment five times and report

the average and maximum results across all runs. During inference, a threshold of 0.5 is used for relationship detection, which is lowered to 0.2, for parsing the out-of-domain samples for human evaluation.

A.2 Pre-processing Steps for Dependency Representation

A.2.1 EDU Segmentation

Depending on the volume of data that needs to be processed, we incorporate the following two segmentation strategies for segmenting text into EDUs: (i) **Neural Segmentation:** For low volume data we use the Bi-LSTM-CRF based discourse segmenter by Wang et al. (2018). (ii) **Rule Based Segmentation:** For larger volumes of data, we use a rule based discourse segmenter, where we segment on encountering the following punctuation :".", "?", "!", ",", ";", and further use a pre-defined set of 113 commonly used discourse markers for finer segmentation (example: "however", "in conclusion", "besides", etc.).

We use the rule based segmentation for preprocessing the large scale IMHO, args.me, FDP, and QR corpus, whereas the neural segmentation is used for segmenting the PE, WD and CMV corpus. Further, we always resort to neural segmentation during inference. Due to it's relatively shorter argument length, for CDCP corpus we treat each proposition as an EDU, and for MicroArg each sentence is considered to be an EDU.

A.2.2 Dataset Specific Pre-processing

IMHO Corpus: We enforce a minimum length of 10 and a maximum length of 300 tokens for each segment. Further, we remove examples whose tokens are split into more than 25 segments. Further, while associating claims with noisy premise, by default the succeeding sentence is chosen as premise if its similarity score is within a margin of 10% compared to the preceding sentence.

arge.me & QR corpus: For args.me, we consider a "pro" stance as "support", and "con" as "attack". For QR, all relationships are considered to be "attack".

FDP Corpus: We remap "Position"→"Major Claim", ("Claim", "Counterclaim", "Rebuttal", "Concluding Statement")→"Claim", and "Evidence"→"Premise". We use textual entailment to associate a claim with a premise. For each claim we construct a set of four candidate premises: two preceding and two following. Using AllenAI's (Gardner et al., 2017) ELMO (Peters et al., 2018) based Textual Entailment we select the premise with the highest entailment score above a threshold of 0.7, as the most likely connected premise, else select the premise immediately following the claim.

MicroArg Corpus: We remap "NA"→"No Relation", ("sup", "exa", "add", "pro")→"Support", and ("reb", "opp")→"Attack". Further, since MicroArg constitutes monologues, we consider the prompt as the previous turn, and convert the monologues to dialogues.

CDCP Corpus: Here we make an assumption and only mark propositions as claim if they have any associated evidence. Else, it's tagged as premise.

WD & PE Corpus: For the WD corpus, we remap "Backing", "Rebuttal" and "'Refutation" to premise. For the PE dataset, we derive dialogues from each paragraph by treating the "Major Claims" (or the essay prompt if major claim is not present) as prior conversation context.

CMV: We remap ("support", "agreement", "partial_agreement", "understand")→"Support", and ("rebuttal_attack", "partial_attack", "rebuttal","undercutter_attack", "partial_disagreement", "disagreement", "undercutter", "attack")→"Attack".

A.3 Additional Stats

Relation	Support	Attack
Intra	96.1	3.9
Inter	45.0	55.0

Table 4: Percentage distribution of Support and Attack Relationships for Inter and Intra-Turn Relations.

Topic	Search Keywords	Count Discussions
abortion	abortion, foeticide	53
gun control	gun control, own gun, second amendment, gun violence, ban gun	19
death penalty	death penalty, capital punishment	18
minimum wage	minimum wage	10

Table 5: Topic distribution of out-of-domain examples collected from CMV.

A.4 Amazon Mechanical Turk Annotations

We leveraged Amazon Mechanical Turk (AMT) in order to collect human evaluations on the model

Task	% Agreement	Cohen's Kappa	Krippendorff's α	N Agreements	N Disagreements	N Cases	N Decisions
Inter-Turn Relation Detection	57.1	0.16	0.13	12	9	21	42
Intra-Turn Relation Detection	57.1	0.22	0.1	12	9	21	42
Component Classification	68.0	0.15	0.14	17	8	25	50

Table 6: Inter-Annotator Agreement of Human Evaluations

generated parsed outputs. We set up human intelligence task (HIT) in the AMT platform, with two evaluators per example and each task worth $0.01. The evaluators were provided with clear instructions on what to annotate and how to annotate the examples, along with a few worked out examples, which are illustrated as screenshots in Figures 5 and 7. The tasks comprised of reading a conversation context, and determining if the presented claim/premise labels are true for component classification (Figure 6), or if the presented argument pairs are valid for inter and intra-turn relation prediction (Figure 8).

In order to ensure quality of annotations, a random portion of the examples presented to each annotator would not be related to the provided conversation, and would have to be marked as "Not in Conversation". Any annotations that failed the quality check were discarded. Further, we discarded annotations which were quickly submitted (less than 2 minutes of work time), and also removed samples where the evaluators missed unchecking the checkboxes, resulting in ambiguity.

This task requires you to read a conversation between multiple participants, and then determine if the claim, premise and non argument labels associated with each of 5 extracted phrase from the conversation are correct.

Background

A brief background on identifying arguments, claims and premises from text. Please read through this brief background in order to understand how to perform the annotation task.

An argument consists of several statements. In its simplest form, it includes a claim that is supported by at least one premise.
Claim: The claim represents a controversial statement which the author tries to persuade the reader of. It is usually a proposition or assumption and should not be accepted by the reader without additional support. This characteristic distinguishes arguments from explanations where the conclusion is a true statement that is not arguable (e.g. an event that happened in the past).
Premise: The premise, underpins the plausibility of the claim, and is usually added for persuading the reader of the claim. Considering the simplest form of an argument, a premise can be seen as a justification for the claim, whereas more complex argumentation structures can also include premises that aim at refuting a claim.
NonArgument: Any other phrases in the conversation which does not belong to a claim or premise are termed as non arguments.

Below we list 3 examples identifying claims and premises from text:

1. *[An advanced gun background check should become routine in all gun sales]$_1$ because [it will prevent gun rampages]$_2$.*
   In this example there are two argument components, where the phrase **1** is the claim, and phrase **2** is the premise, which provides iustification for the first one.
2. *First of all, [people cannot predict their own future or know what will happen tomorrow]$_1$. [The world is full of disasters such as wars, pollution, famine, drought, starvation, natural disasters and diseases]$_2$. So [it is just a big mistake to have children]$_3$.*
   In this example there are three argument components, where the phrase **3** is the claim, and the phrases **1** and **2** are the premises, which provides justification for the claim.
3. *Furthermore, [it is a very heavy psychological and physical burden to have children]$_1$. [A mother carries her baby in her womb for 9 months and 10 days and then the baby torments her during and after the birth]$_2$. [There is no peace, no silence or no sleep at home]$_3$. On the other hand, [the father has to work hard and earn more money]$_4$ because [the baby comes with his expenses]$_5$.*
   In this example there are five argument components, where the phrase **1** is the claim, and the phrases **2**, **3** and **4** are the premises, which provides justification for the claim. Phrase **5** is also a premise, and it provides justification for the premise in phrase **4**.

Task Instructions: **Read the below conversation between multiple participants, and mark whether the claim, premise and non argument label associated with each of 5 extracted phrase from the conversation are correct. Note that:**

1. For each extracted phrase there are 3 options:
   1. Correct: Indicating that the associated label (claim, premise non argument) with the extracted phrase from the conversation is correct.
   2. Inorrect: Indicating that the associated label (claim, premise non argument) with the extracted phrase from the conversation is NOT correct.
   3. Not In Conversation: Out of the 5 presented phrases, there is at least 1 phrase which is not from the current converation. Such a phrase should be marked as "Not In Conversation".
2. There can be multiple claims and premises in the conversation.
3. For each phrase, only 1 checkbox should be ticked. Initially all the checkboxes are ticked, and you will have to uncheck the irrelevant boxes.

Figure 5: Instructions provided for evaluating component classification in AMT.

Conversation:

USER 0: I feel like men in long term committed relationships should have a say in their s/o having a abortion. ok first i hate the argument "my body my choice". but in a way i kind of would see it in one night stands, rape, incest, or even just a fwb or fb. but if in a committed ltr the man who is in the relationship, if it's 100% his should have a say if she gets a abortion. i'm not talking about 100 different what if's but solely if in a relationship with her. he should have a say in the decision making. do i think she should be judged or punished? hell naw. but it would be common courtesy to say "hey i'm pregnant what are your thoughts?" instead of "hey i was pregnant and decided to have a abortion"
USER 1: um, the man doesn't carry any of the direct health affects of being pregnant. they don't directly deal with the indirect affects of pregnancy. thus, it seems odd that they can make a choice that can negativity affect the health of another person. the woman, because it is her body, should have the final say.

Phrases to annotate:

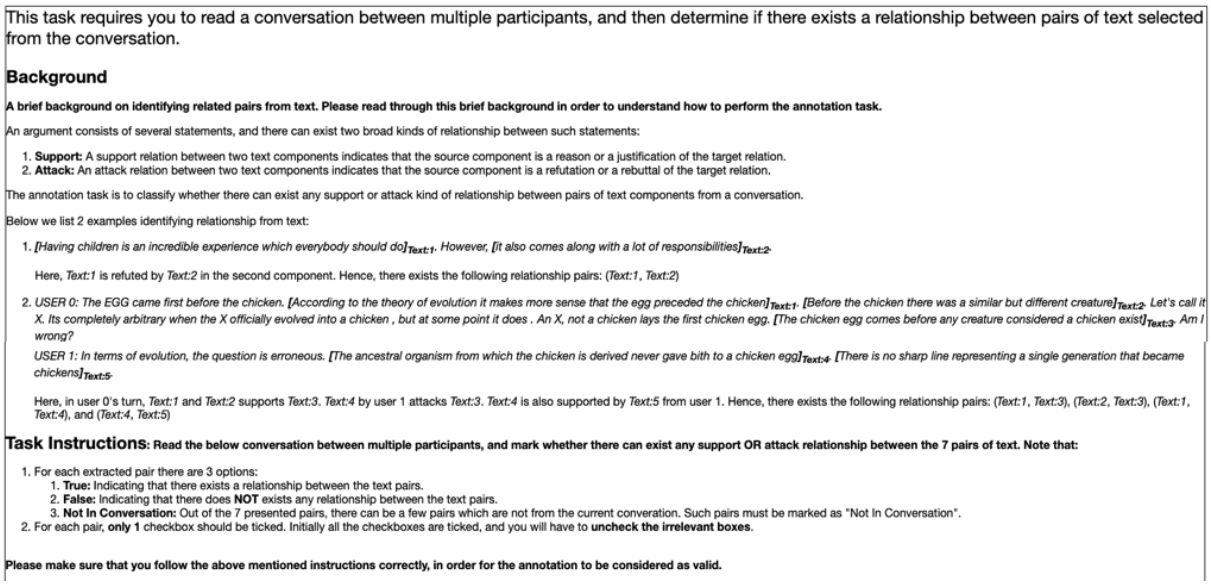
- um, the man doesn't carry any of the direct health affects of being pregnant. : **premise**

  ☑ Correct ☑ Incorrect ☑ Not In Conversation
- they don't directly deal with the indirect affects of pregnancy. : **premise**

  ☑ Correct ☑ Incorrect ☑ Not In Conversation
- the woman, because it is her body, should have the final say. : **premise**

  ☑ Correct ☑ Incorrect ☑ Not In Conversation

Figure 6: Component classification sample from AMT.



This task requires you to read a conversation between multiple participants, and then determine if there exists a relationship between pairs of text selected from the conversation.

Background

A brief background on identifying related pairs from text. Please read through this brief background in order to understand how to perform the annotation task.

An argument consists of several statements, and there can exist two broad kinds of relationship between such statements:

1. **Support:** A support relation between two text components indicates that the source component is a reason or a justification of the target relation.
2. **Attack:** An attack relation between two text components indicates that the source component is a refutation or a rebuttal of the target relation.

The annotation task is to classify whether there can exist any support or attack kind of relationship between pairs of text components from a conversation.

Below we list 2 examples identifying relationship from text:

1. *[Having children is an incredible experience which everybody should do]$_{Text:1}$. However, [it also comes along with a lot of responsibilities]$_{Text:2}$.*

   Here, *Text:1* is refuted by *Text:2* in the second component. Hence, there exists the following relationship pairs: (*Text:1*, *Text:2*)

2. *USER 0: The EGG came first before the chicken. [According to the theory of evolution it makes more sense that the egg preceded the chicken]$_{Text:1}$. [Before the chicken there was a similar but different creature]$_{Text:2}$. Let's call it X. Its completely arbitrary when the X officially evolved into a chicken , but at some point it does . An X, not a chicken lays the first chicken egg. [The chicken egg comes before any creature considered a chicken exist]$_{Text:3}$. Am I wrong?*

   USER 1: In terms of evolution, the question is erroneous. [The ancestral organism from which the chicken is derived never gave bith to a chicken egg]$_{Text:4}$. [There is no sharp line representing a single generation that became chickens]$_{Text:5}$.

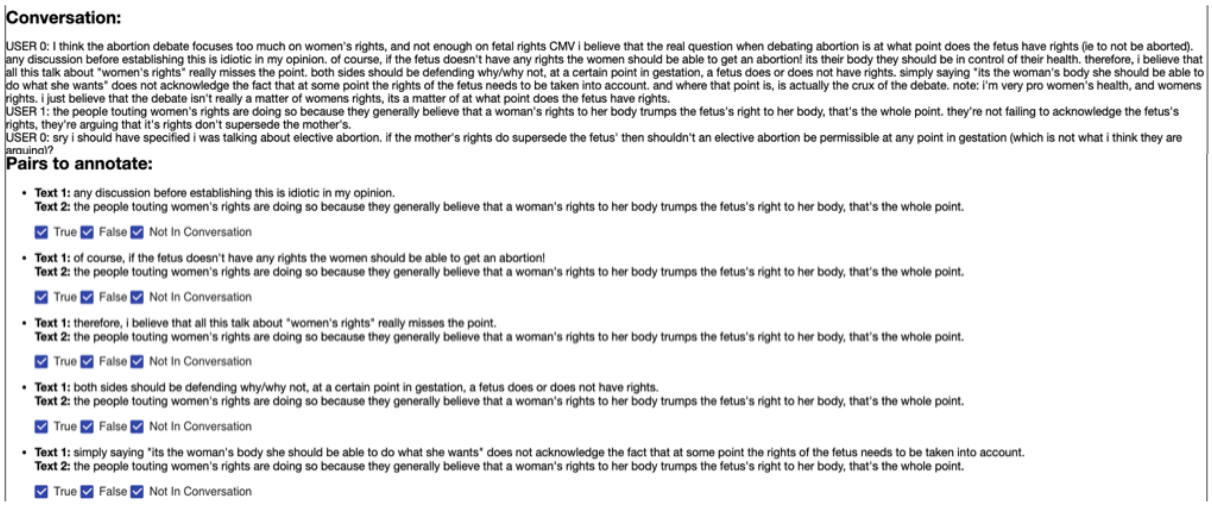
   Here, in user 0's turn, *Text:1* and *Text:2* supports *Text:3*. *Text:4* by user 1 attacks *Text:3*. *Text:4* is also supported by *Text:5* from user 1. Hence, there exists the following relationship pairs: (*Text:1*, *Text:3*), (*Text:2*, *Text:3*), (*Text:1*, *Text:4*), and (*Text:4*, *Text:5*)

Task Instructions: Read the below conversation between multiple participants, and mark whether there can exist any support OR attack relationship between the 7 pairs of text. Note that:

1. For each extracted pair there are 3 options:
   1. **True:** Indicating that there exists a relationship between the text pairs.
   2. **False:** Indicating that there does **NOT** exists any relationship between the text pairs.
   3. **Not In Conversation:** Out of the 7 presented pairs, there can be a few pairs which are not from the current converation. Such pairs must be marked as "Not In Conversation".
2. For each pair, **only 1** checkbox should be ticked. Initially all the checkboxes are ticked, and you will have to **uncheck the irrelevant boxes**.

Please make sure that you follow the above mentioned instructions correctly, in order for the annotation to be considered as valid.

Figure 7: Instructions provided for evaluating intra/inter-turn relation identification in AMT.



Conversation:

USER 0: I think the abortion debate focuses too much on women's rights, and not enough on fetal rights CMV i believe that the real question when debating abortion is at what point does the fetus have rights (ie to not be aborted). any discussion before establishing this is idiotic in my opinion. of course, if the fetus doesn't have any rights the women should be able to get an abortion! its their body they should be in control of their health. therefore, i believe that all this talk about "women's rights" really misses the point. both sides should be defending why/why not, at a certain point in gestation, a fetus does or does not have rights. simply saying "its the woman's body she should be able to do what she wants" does not acknowledge the fact that at some point the rights of the fetus needs to be taken into account. and where that point is, is actually the crux of the debate. note: i'm very pro women's health, and womens rights. i just believe that the debate isn't really a matter of womens rights, its a matter of at what point does the fetus have rights.
USER 1: the people touting women's rights are doing so because they generally believe that a woman's rights to her body trumps the fetus's right to her body, that's the whole point. they're not failing to acknowledge the fetus's rights, they're arguing that it's rights don't supersede the mother's.
USER 0: sry i should have specified i was talking about elective abortion. if the mother's rights do supersede the fetus' then shouldn't an elective abortion be permissible at any point in gestation (which is not what i think they are arguing)?

Pairs to annotate:

- **Text 1:** any discussion before establishing this is idiotic in my opinion.
  Text 2: the people touting women's rights are doing so because they generally believe that a woman's rights to her body trumps the fetus's right to her body, that's the whole point.
  ☑ True ☑ False ☑ Not In Conversation
- **Text 1:** of course, if the fetus doesn't have any rights the women should be able to get an abortion!
  Text 2: the people touting women's rights are doing so because they generally believe that a woman's rights to her body trumps the fetus's right to her body, that's the whole point.
  ☑ True ☑ False ☑ Not In Conversation
- **Text 1:** therefore, i believe that all this talk about "women's rights" really misses the point.
  Text 2: the people touting women's rights are doing so because they generally believe that a woman's rights to her body trumps the fetus's right to her body, that's the whole point.
  ☑ True ☑ False ☑ Not In Conversation
- **Text 1:** both sides should be defending why/why not, at a certain point in gestation, a fetus does or does not have rights.
  Text 2: the people touting women's rights are doing so because they generally believe that a woman's rights to her body trumps the fetus's right to her body, that's the whole point.
  ☑ True ☑ False ☑ Not In Conversation
- **Text 1:** simply saying "its the woman's body she should be able to do what she wants" does not acknowledge the fact that at some point the rights of the fetus needs to be taken into account.
  Text 2: the people touting women's rights are doing so because they generally believe that a woman's rights to her body trumps the fetus's right to her body, that's the whole point.
  ☑ True ☑ False ☑ Not In Conversation

Figure 8: Relation identification sample from AMT.