Graph-based Argument Quality Assessment

Ekaterina Saveleva and **Volha Petukhova** and **Marius Mosbach** and **Dietrich Klakow**
Spoken Language Systems Group, Saarland Informatics Campus
Saarland University, Saarbrücken, Germany
{esaveleva,v.petukhova,mmosbach,dietrich.klakow}@lsv.uni-saarland.de

Abstract

The paper presents a novel discourse-based approach to argument quality assessment defined as a graph classification task, where the depth of reasoning (argumentation) is evident from the number and type of detected discourse units and relations between them. We successfully applied state-of-the-art discourse parsers and machine learning models to reconstruct argument graphs with the identified and classified discourse units as nodes and relations between them as edges. Then Graph Neural Networks were trained to predict the argument quality assessing its acceptability, relevance, sufficiency and overall cogency. The obtained accuracy ranges from 74.5% to 85.0% and indicates that discourse-based argument structures reflect qualitative properties of natural language arguments. The results open many interesting prospects for future research in the field of argumentation mining.

1 Introduction

Argumentation modelling and mining are steadily gaining attention of the broad natural language processing and engineering community. In many studies and applications, assessment of the argument quality plays an important role. The ability to construct good arguments and engage in argumentative discussions is assessed by argumentation systems focusing on training hypothetical reasoning, creating and structuring arguments (Ashley et al., 2007), preventing opinion manipulation, detecting inconsistent arguments in online discussions and addressing different standpoints, attacking or supporting claims with evidence (DebateGraph[1] and TruthMapping[2]) as well as on the use of multimodal rhetorical devices (Petukhova et al., 2017a). Assessment of argument quality, its organization, clarity, adherence and strength, are approached by several authors as sub-tasks in the evaluation of written essays (Stab and Gurevych, 2014; Persing and Ng, 2015; Wachsmuth et al., 2016; Stab and Gurevych, 2017). Online content is searched to filter or weight the validity of statements and factoids (Rowe and Butters, 2009), to identify fake news and false claims (Popat et al., 2018) and to detect opinion manipulation (Cambria et al., 2010). While the acceptability of an argument in the presence of other supporting or attacking arguments has been addressed (Dung, 1995; Cayrol and Lagasquie-Schiex, 2005), 'local' argument quality still deserves our attention – an argument built on a certain set of conditions, is logically strong, rhetorically convincing, socially undistorted by virtue of its intrinsic properties.

In this paper, we present a novel approach to assessing the structural strength and inferential weakness of arguments as merits of argument cogency. The approach relies on the discourse-based reconstruction of argumentation schemes. For this, we apply state-of-the-art discourse parsers and machine learning models to reconstruct argument graphs where the identified discourse units are represented as nodes and the classified discourse relations between them as edges. A Graph Neural Network (GNN) model is built to predict the quality (*low vs high*) of the reconstructed argument graphs in terms of argument acceptability, relevance, sufficiency and overall cogency.

The paper is structured as follows. Section 2 defines the conceptual framework within which the study is performed. We provide the definition of an argument and elaborate on its internal structure. In Section 3, we survey related work on argument quality assessment. Section 4 presents the argument graph reconstruction approach. The performed GNN-based quality assessment experiments are discussed and results are reported in

[1] http://debategraph.org/
[2] https://www.truthmapping.com/

Proceedings of Recent Advances in Natural Language Processing, pages 1268–1280
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_143

Section 5. Section 6 summarises our findings and outlines directions for future research.

2 Argument and Its Structure

An argument may be considered as an atomic entity without an internal structure. For instance, an argument is defined as an overall position held by a person towards an idea or attitude, e.g. a stance in 'favour' or 'against' a certain motion (Somasundaran and Wiebe, 2009). A structured argumentation model is an essential element for the tasks aiming at understanding and emulating of human inference, investigating patterns of reasoning, focusing at extraction and validity assessment of arguments. A simple argument structure is then defined as consisting of *a claim* that is supported by *evidence(s)* (Mochales and Moens, 2011; Aharoni et al., 2014). A claim is an assertion that an argument aims to prove, i.e. a claim is a *conclusion* whose merit must be established. Evidence comprises propositions which give reasons or grounds for drawing the conclusion.

This general argument definition has been translated into several discourse-based schemes for analysing and evaluating natural language arguments (Teufel, 1999; Palau and Moens, 2009). An argument is modelled as a group of Argumentative Discourse Units (ADUs) – text segments corresponding to propositions that are argumentatively relevant and have their own argumentative function (Peldszus and Stede, 2013). An EDU can function as a claim, as an evidence or as a conclusion. An ADU can be identified as a collection of several Elementary Discourse Units (EDUs) which correspond to clauses in written discourse and to dialogue acts in spoken discourse (Petukhova et al., 2016). Discourse relations such as *Justification*, *Motivation*, *Cause*, and *Exemplification* can be used to identify how propositions are related to each other, inferring the type of support that is expressed. A claim may be summarized or re-stated in a conclusion. Figure 1 depicts a general discourse-based argument structure.

ADUs reflect different ways to provide support for a claim, i.e. links between them express the level of support that evidence provides to the claim and the level of their sufficiency to draw a valid conclusion. Figure 2 provides an example of an argument. EDUs (solid-line rectangles) are combined by means of discourse relations into ADUs (dotted-line rectangles) which are connected to each other by support links. Evidence may either together (*linked support*, e.g. Evidence 2.1 and 2.2) or independently (*multiple support*, e.g. Evidence 1, 2 and 4) support a conclusion. A premise may provide support for another premise and indirectly support a conclusion (*serial support*, e.g. Evidence 3 and 2). A special form of lending support to a claim is that of providing examples (*example support*, e.g. Evidences 4.1 and 4.2).[3]

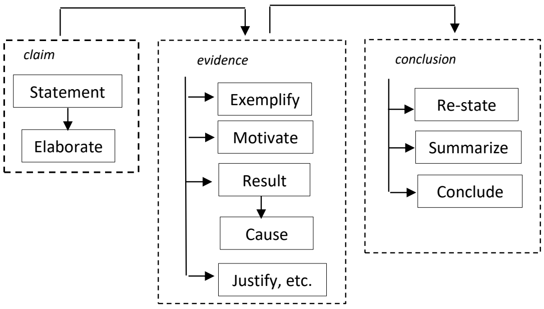


Figure 1: Argument structure observed in spoken debate arguments, adapted from Petukhova et al. (2017a). Solid-line rectangles represent EDUs and dotted-line rectangles represent ADUs.

3 Related Work

Clear properties of a good argument and successful argumentation are not easy to define. Wachsmuth et al. (2017) proposed a unified taxonomy of argumentation quality assessment that resulted from an extensive analysis of numerous existing approaches. The assessment comprises three quality dimensions: cogency, effectiveness and reasonableness. Argument quality assessment aims at answering the question how *logical*, *persuasive* or *convincing* the given argument is, and how *rhetorically appealing* it is for the targeted audience.

Evaluation of argument cogency is based on the **truthfulness** and **logical** coherence of arguments. An argument is cogent if it has acceptable premises that are relevant and sufficient to support the conclusion (Johnson and Blair, 2006; Govier, 2013). A premise is *acceptable* if it is rationally worthy of being believed to be true (Wachsmuth et al., 2017). According to Govier (2013), a premise is locally acceptable if it is supported by a cogent sub-argument or another cogent argument; it is a matter of common knowledge, testimony or expert view (appeal to authority). A statement A is positively *relevant* to another statement B if and only if the truth of A counts in favour of the truth of B. This means that

[3] For a discussion on other types of support links we refer to Palau and Moens (2009) and Peldszus and Stede (2013).

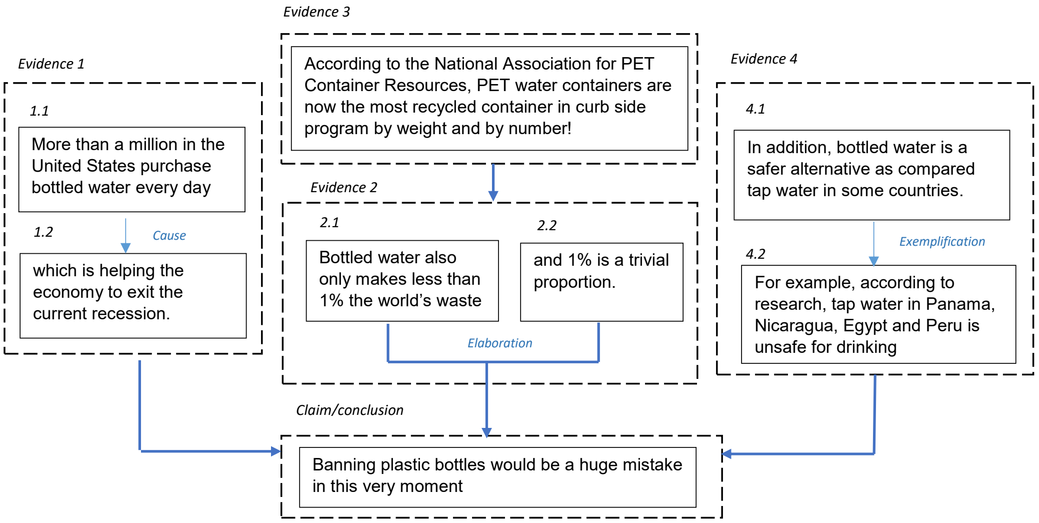

Evidence 3
According to the National Association for PET Container Resources, PET water containers are now the most recycled container in curb side program by weight and by number!
Evidence 1
1.1
More than a million in the United States purchase bottled water every day
1.2
Cause
which is helping the economy to exit the current recession.
Evidence 2
2.1
Bottled water also only makes less than 1% the world's waste
2.2
and 1% is a trivial proportion.
Elaboration
Evidence 4
4.1
In addition, bottled water is a safer alternative as compared tap water in some countries.
Exemplification
4.2
For example, according to research, tap water in Panama, Nicaragua, Egypt and Peru is unsafe for drinking
Claim/conclusion
Banning plastic bottles would be a huge mistake in this very moment

Figure 2: Argument example from Dagstuhl-15512 ArgQuality Corpus (Wachsmuth et al., 2017) annotated with core ISO 24617-8 core discourse relations (Bunt and Prasad, 2016) and support links observed.

A provides some evidence for B, or some reason to believe that B is true. An argument is locally *sufficient* if all premises together provide sufficient reasons to accept the conclusion. The preconditions of the argument sufficiency are rooted in its local acceptability and its local relevance (Govier, 2013). The local sufficiency of an argument is often called *inferential* sufficiency and it holds if one of the following logical patterns is applicable: deductive entailment, conductive support, inductive support and analogy.

Argument quality is found to correlate with the argument's actual **persuasive** success. Persuasion is defined as a process of encouraging people to do or believe something through argument. Here, many factors are relevant, including psychological effects of argument memorisation, replication and reviewing (Kumkale and Albarracín, 2004). Certain argumentation patterns are acknowledged as more persuasive than others, however they may differ in different domains. Hornikx (2008) experimentally investigated lay people's expectations about the persuasiveness of anecdotal, statistical, causal, and expert evidence, and compared these expectations with the actual persuasiveness of these evidence types. Van Eemeren and Grootendorst (2004) defined symptomatic (sign), comparison (resemblance) and causal (consequence) argumentation, and specified what argumentative patterns are more suitable/persuasive for what communicative types in various domains. For persuasive essays, different quality dimensions of argumentation were studied such as essay's organization (Persing et al., 2010), thesis clarity (Persing and Ng, 2013), prompt adherence (Persing and Ng, 2014) and argument strength (Persing and Ng, 2015). These studies exploit a complex feature-rich approach to predict a score for each essay based on its content or style along with all of these categories. The study of Persing and Ng (2017) looks at the argument persuasiveness from a different point of view: it does not try to estimate how persuasive an argument is but attempts to explain why an argument is experienced as unpersuasive. Research has also targeted various interactive aspects, e.g. capturing the interactions between participants on argument level (Ji et al., 2018) and providing feedback regarding the argument persuasiveness (Ke et al., 2018).

Many studies explore the aspect of argument **convincingness**[4] . In contrast to cogency, which is based on the truthfulness and logical coherence of arguments, convincingness is related to subjective perception by the audience (Wei et al., 2016). Experiments were performed to detect more convincing arguments (Habernal and Gurevych, 2016; Simpson and Gurevych, 2018) and evidence (Gleize et al., 2019).

The **rhetorical force** of an argument should not be underestimated. Due to the use of powerful rhetorical devices, even a not very cogent argument may be perceived as convincing (Petukhova et al., 2017b; Hirschberg, 2002). People generally associate certain speech, personality and interaction features with what they think is a persuasive argument.

[4] It should be noted here that persuasiveness and convincingness of an argument are terms that are often used interchangeably.

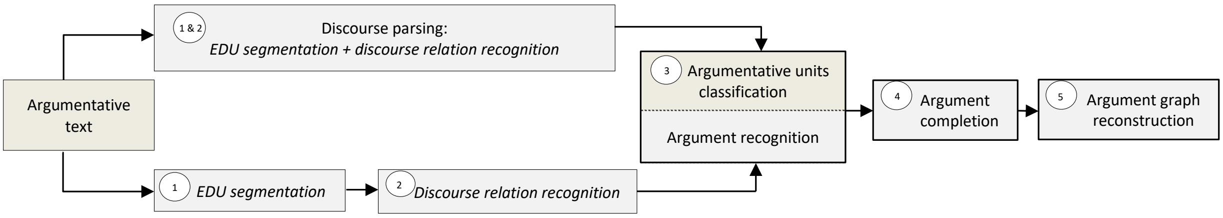

1 & 2
Discourse parsing:
EDU segmentation + discourse relation recognition
Argumentative text
1
EDU segmentation
2
Discourse relation recognition
3
Argumentative units classification
Argument recognition
4
Argument completion
5
Argument graph reconstruction

Figure 3: Argument graph reconstruction pipeline.

More broadly, the persuasion literature of the last decades has shown that an argument that has higher perceived competence (e.g., evidence-based expert knowledge) and/or higher warmth (e.g., more likeable and trustworthy) is more convincing (Petty and Cacioppo, 1986; Albarracín et al., 2019).

The study of Wachsmuth et al. (2016) suggests an argument quality assessment approach that focuses solely on the argument structure, and defines statistical patterns in the structure of essays and define novel features that are evaluated in argumentation-related essay scoring tasks. The present study investigates the structural properties of a cogent argument and assesses its inferential strength, i.e. structures of inference (argumentation schemes) predicted from the associated amount, depth and type of evidence provided to the claim.

4 Argument Graph Reconstruction

We define *argument graph reconstruction* to involve: (1) segmentation of a text into EDUs; (2) discourse relation detection and classification between them; (3) identification and classification of ADUs based on the classified discourse relations; and (4) argument completion, i.e. reconstruction of the implicit units to achieve a complete argument structure, see Figure 3.

We performed *joint* and *two-stage* segmentation and classification of EDUs. For the joint segmentation and classification, the full PDTB parser developed by Lin et al. (2010) was applied. We observed that the parser failed to identify many EDU spans.[5] Rather low overall F1 scores of 21.20% for exact segment boundaries match and 5-class discourse relation classification were achieved on the Penn Discourse Tree Bank 1.0 corpus (PDTB 1.0, Prasad et al. (2005)), see the right part of Table 1. However, we observed, that in case of the correct span identification, relations classification was reasonably accurate. Misclassified cases mostly belonged to the implicit discourse relations as they were more difficult to classify then the explicit ones, a well known problem reported in the literature.

The two-stage segmentation and classification was performed applying the BiLSTM-CRF based segmentation model NeuralEDUSeg developed by Wang et al. (2018) and the XLNet-large discourse relations classification model by Yang et al. (2019). A segmentation performance of 68.55% in terms of F1 score was achieved when testing on the PDTB 1.0 and PDTB 2.0 datasets (PDTB 2.0, Prasad et al. (2008)).[6] For discourse relation recognition with the XLNet model designed by Yang et al. (2019), we first carried out a binary classification to establish whether any relation exists between the identified units, i.e. discriminate between the `Rel` class which includes any type of discourse relations and `NoRel` types. The former comprises explicitly marked (`Explicit`), implicitly marked (`Implicit`) and alternatively lexicalized (`AltLex`) discourse relations, the later includes `EntRel` relation which is not a discourse relation between clauses but an entity-based coherence relation. Subsequently, we performed five-class top-level (L1) and ten-class fine-grained (L2) relations classification. Table 4 in Appendix I provides an overview of the PDTB discourse relation and their distribution in the PDTB 1.0 and the newer PDTB 2.0 corpora.

Since class distributions were unbalanced in all classification settings, *re-sampling* was performed: *up-sampling* of the under-represented `NoRel` class in binary classification by adding synthetic samples. For this random EDUs from different textual units were combined. In addition to this, *down-sampling* of the majority classes in the multi-class settings was performed. For the training and evaluation procedures, we fine-tuned each encoder model following the suggestions of Mosbach et al. (2021) and trained it for 10 epochs using a learning rate of

[5] The same observation was made by Hewett et al. (2019).

[6] PDTB 2.0 is the PDTB 1.0 corpus extended with annotations of implicit relations for the entire corpus, senses of all connectives and attribution of object type, scopal polarity and determinacy. Thus, for the purpose of this study, differences between PDTB 1.0 and PDTB 2.0 are not relevant.

Joined segmentation & classification		**Two-stage segmentation & classification**							
PDTB 1.0 data		**PDTB 2.0 data**				**Dagstuhl data**			
Full parsing		***EDU segmentation***		***PDTB relation recognition***		***EDU segmentation***		***PDTB relation recognition***	
Scenario	F1 (in %)	Scenario	F1 (in %)	Scenario	Accuracy (in %)	Scenario	F1 (in %)	Scenario	Accuracy (in %)
exact segment match & 5-class classification	21.20	exact match	68.55	5-class	66.37	exact match	47.94	5-class	60.22
				10-class	53.64			10-class	50.48

Table 1: Performance overview on the *joined* EDU segmentation and 5-class discourse relation classification task with Lin et al. (2010) parser in terms of F1 scores (in %) on the PDTB 1.0 corpus (*left*); and on the *two-stage* segmentation and classification tasks performing EDU segmentation with the NeuralEDUSeg model (Wang et al., 2018) in terms of F1 scores (in %) on the PDTB 2.0 corpus, and 5- and 10-class discourse relation classification with the fine-tuned XLNet-large model (Yang et al., 2019) in terms of accuracy (in %) on the DagStuhl corpus (*right*).

0.00001 and a batch-size of eight. Accuracy was observed to drop with a higher number of classes to learn from 66.37% (five classes) to 53.64% (ten classes), see the middle part of Table 1.

To demonstrate the applicability of the approach beyond PDTB, we applied the two-stage segmentation and classification procedure and fine-tuned models on the argumentative *Dagstuhl15512 ArgQuality* corpus which is a collection of 304 argumentative texts annotated according to 15 argument quality criteria (Wachsmuth et al., 2017). The output was manually examined and corrected. Table 1 reports the performance of the NeuralEDUSeg and XLNet-large models on the manually corrected argumentative Dagstuhl corpus. The resulting *Dagstuhl* corpus of argumentative units annotated with discourse relations contains the same number of 304 arguments as the original *Dagstuhl15512 ArgQuality*, but segmented into 2,222 EDU pairs. The evaluated models showed a reasonable segmentation, F1 score of 47.94% for exact segment match, and discourse relation recognition (accuracy ranging from 50.48% to 60.22%) performance on argumentative discourse data.

We observed that some argument components, often claims, are implicit, see also Wachsmuth et al. (2017). Without the claim or conclusion, an argument structure is incomplete. Therefore, we reconstructed a claim for every topic in the corpus, either 'for' or 'against' stance it may present. The reconstructed claim is a simple sentence corresponding to a single EDU, see Table 5 in Appendix II for selected examples.

The identified Dagstuhl arguments are of different length and have various, often complex discourse-based structures distinguishable through diverse linking patterns and number of evidences provided by an arguer to support a claim. Figure 4 provides an example of the identified discourse-based argumentation scheme. The upper node represents the claim *Books are better than TV* which was supported by seven evidence statements, six of them connected to the claim by means of *Contingency.Cause* relation and one by *Expansion.Instantiation*. Five of the evidence statements correspond to one EDU, whereas the other two are more complex and consist of two EDUs.

Finally, argument structures were represented as graphs where the detected EDUs spans are represented as nodes and the classified discourse relations – as edges visualising number and level of supporting evidence through links to the claim and premises. 2,278 reconstructed arguments have 303 structures (argumentative schemes) specifying 172 unique reasoning patterns. Figure 6 in Appendix III provides the most frequent examples of the reconstructed discourse-based argument structures.

5 Argument Quality Assessment Using Graph Neural Networks

Our main assumption is that arguments constructed to follow certain patterns and containing particular discourse relations are of higher quality, i.e. inferentially stronger, than others. Thus, the amount and type of evidence matter. For example, a widely used structure of debate arguments is known as the **ARE**. ARE comprises a claim of an **A***rgument* supported by a **R***eason* and an **E***vidence*, see also Petukhova et al. (2016). Another commonly used argument structuring technique is called **chunking** (Johnson, 2009). Here, arguers generalise from a claim (*chunking up*), provide a specific example (*chunking down*) or draw analogies (*chunking side-*

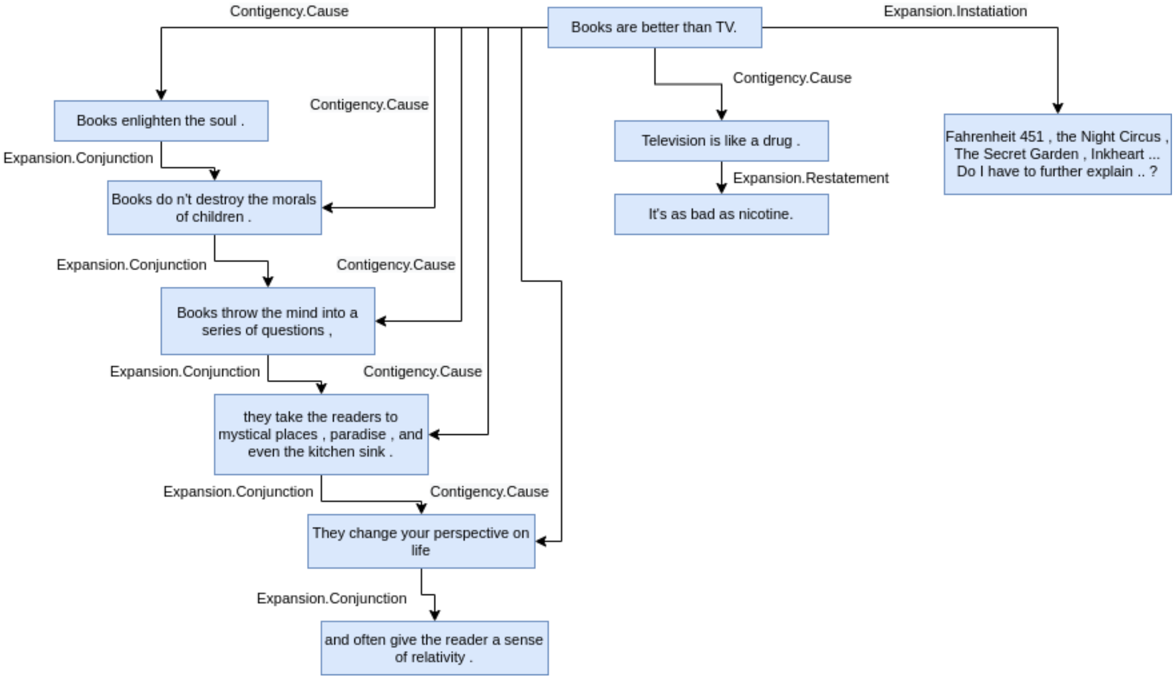

Contigency.Cause
Books are better than TV.
Expansion.Instatiation
Contigency.Cause
Books enlighten the soul .
Contigency.Cause
Expansion.Conjunction
Television is like a drug .
Fahrenheit 451 , the Night Circus , The Secret Garden , Inkheart ... Do I have to further explain .. ?
Books do n't destroy the morals of children .
Expansion.Restatement
It's as bad as nicotine.
Expansion.Conjunction
Contigency.Cause
Books throw the mind into a series of questions ,
Expansion.Conjunction
Contigency.Cause
they take the readers to mystical places , paradise , and even the kitchen sink .
Expansion.Conjunction
Contigency.Cause
They change your perspective on life
Expansion.Conjunction
and often give the reader a sense of relativity .

Figure 4: Example of a discourse-based argument structure identified in the Dagstuhl15512 ArgQuality corpus.

ways). Thus, an argument which contains *Cause* or *Instantiation* relations making a claim to be justified and explained, is expected to be of high quality. A large variety of identified discourse relations may indicate that an argument is very elaborate. However, very lengthy arguments are often difficult to comprehend and may be rhetorically less persuasive.

Particular structures can be important to compute local sufficiency. For instance, conductive reasoning may be expressed by the number of first-level evidence supporting the main claim independently (multiple support) and together (linked support) connected by means of *Expansion* relations. For inductive support, Exemplification discourse relations can be analysed as evidence providing an example support. Finally, the argumentation depth can be relevant for the argument sufficiency assessment and can be computed by looking at evidence which is linked to other evidence statements providing a serial support. Obviously, not only linking patterns but also evidence content would impact the argument quality.

Arguments in the Dagstuhl15512 ArgQuality corpus were annotated by seven independent annotators across 15 quality dimensions including four for argument cogency: acceptability, relevance, sufficiency and overall cogency. Quality scores from 1 (low) to 3 (high) were assigned. A fair inter-annotator agreement for all cogency dimensions was reached ranging from .44 to .47 in terms of Krippendorffs α (Wachsmuth et al., 2017). Distribution of the annotated quality classes resulted in a rather unbalanced training set, in particular for the sufficiency dimension, therefore we combined the minority class with the adjacent one defining a binary classification task predicting arguments of a *lower* and of a *higher* quality, see Table 2 for distributions.

To assess the argument cogency, we employed a Graph Neural Network (GNN) model which is able to generalize over manifold structures. Errica et al. (2019) presents an overview of GNNs models for graph classification, e.g DGCNN, DiffPool, ECC, GIN, GraphSAGE. However, none of these models exploit edge features required for our application so it incorporates discourse relation information.

5.1 Architecture Overview

For our experiments, we use the Graph Attention Network (GAT) model by Veličković et al. (2018). Initial inputs to the model include the node feature matrix X^0 as presented in Figure 5 (left). The matrix is fed to the GNN layer which is able to handle binary edge features, i.e. in our case the model can only handle the existence or absence of a relation between two EDUs. The single-head and multi-head attention mechanisms used within the GNN layer are illustrated on Figure 5 (center and right). As a result, a new node matrix X^1 is

	overall		training set		validation set		test set	
class	**lower**	**higher**	**lower**	**higher**	**lower**	**higher**	**lower**	**higher**
cogency	143 (47.2%)	160 (52.8%)	114 (47.1%)	128 (52.9%)	14 (46.7%)	16 (53.3%)	14 (45.2%)	17 (54.8%)
acceptability	71 (23.5%)	232 (76.5%)	57 (23.6%)	185 (76.4%)	8 (26.7%)	22 (73.3%)	6 (19.4%)	25 (80.6%)
relevance	183 (60.4%)	120 (39.6%)	146 (60.3%)	96 (39.7%)	20 (66.7%)	10 (33.3%)	17 (54.8%)	14 (45.2%)
sufficiency	167 (55.1%)	136 (44.9%)	133 (55%)	109 (45%)	17 (56.7%)	13 (43.3%)	17 (54.8%)	14 (45.2%)

Table 2: Class distribution in the training, validation and test sets in the Dagstuhl15512 ArgQuality corpus.

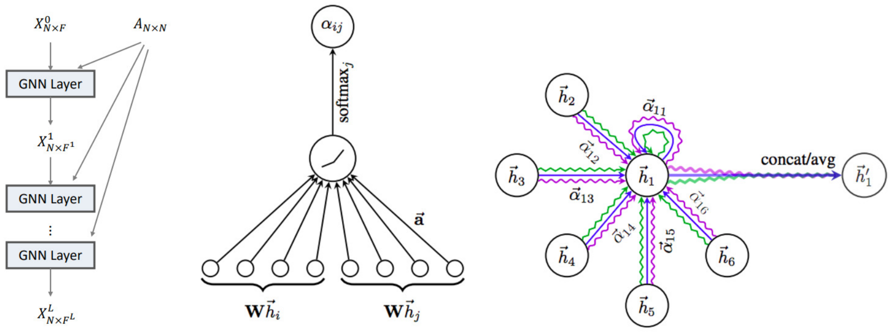


Figure 5: **Left**: Our model architecture. Tensor X is a matrix representation of the node features of the graph with dimensions NxF where N is the number of nodes and F is the number of features (Gong and Cheng, 2019). **Center**: The attention mechanism employed by GAT layer (Veličković et al., 2018). **Right**: an illustration of multi-head attention (with K = 3 heads) by node 1 on its neighbourhood (Veličković et al., 2018).

produced. The procedure is repeated for every subsequent layer. Our network consists of three such layers. For graph classification, an average pooling layer is applied to the first dimension of X^{L}, i.e. the feature matrix is reduced to a single vector representing the whole graph. Subsequently, the fully connected layer is applied to the vector whose outputs are used as logits for the final classification.

The model described above allows incorporating various features in the node feature matrix X^0. We experimented with three different settings. In the first design, no node features were available. Hence, that model predicts argument quality based solely on the graph structure. In the second design, we encoded the texts corresponding to each node using GloVe embeddings[7] taking the mean of all word embeddings of an EDU to obtain the EDU representation. In the third setting, we used BERT embeddings of an EDU as node features using the DistilBERT pre-trained model optimized for Semantic Textual Similarity task (Reimers and Gurevych, 2019). The vector dimensionality for each of the settings described above was 300, 300 and 768 dimensions, respectively.

5.2 Experimental Results

We trained a model for each quality dimension (cogency, acceptability, relevance, sufficiency) varying the number of epochs from 100 to 1000. Subsequently, we determined the best training setting based on the validation set accuracy. We observed that the performance on quality dimensions applying different either GloVe or BERT node features depends on the number of training epochs. Furthermore, we learned that the models with the GloVe and BERT node features reduced loss much faster than the models with uniform node features which indicates a possible overfitting. Therefore, we experimented with dropout rates from 0.2 to 0.6 applying it to the output of the last hidden layer. Test accuracy of BERT-based models improved for all dimensions except for acceptability. 10 training runs for each quality dimension showed that the test accuracy for certain models did not vary much with different dropout rates, the results were also proven not to be statistically significant according to the paired t-test. Table 3 reports the highest mean test accuracy achieved across five runs.

All trained quality assessment models outperformed their corresponding baselines indicating that the graph structure and node features incorporate useful information. Models generally were proven to be resistant to class imbalance. Resampling or weighted training and testing did not result in better performance, which indicates that more sophisticated methods are required to improve results. Further findings suggest that there is no ‘universal’ training setting for various quality dimensions: to achieve acceptable performance some models require longer training and different sets

[7] http://nlp.stanford.edu/data/glove.42B.300d.zip

quality dimension		majority classifier	uniform features	GloVe features	BERT features
cogency	accuracy	54.8	**72.5**	68.0	63.0
	epochs		400	800	800
	dropout		-	0.2	0.4
acceptability	accuracy	80.6	**85.0**	84.5	**85.0**
	epochs		300	600	600
	dropout		-	0.3	-
relevance	accuracy	54.8	69.5	**74.5**	74.0
	epochs		300	200	1000
	dropout		-	0.2	0.5
sufficiency	accuracy	54.8	64.0	**74.5**	69.0
	epochs		500	200	900
	dropout		-	0.4	0.6

Table 3: Classification accuracies on Dagstuhl15512 ArgQuality argument quality assessment applying a graph classification model. The results are reported for each quality dimension (cogency, acceptability, relevance, sufficiency) and using different node features (uniform, GloVe, BERT).

of node features. Surprisingly, sophisticated node features do not always lead to a better model performance. For instance, for the cogency dimension the model without node features significantly outperformed the GloVe- and BERT-based models which may suggest that the argument structure alone is sufficient to accurately predict argument cogency. This confirms our initial assumption. However, we are cautious with this conclusion and emphasise that further experiments on larger datasets are needed. Training on such small training (242 instances), validation (30 instances) and test (30 instances) sets is a challenging task causing oscillation in the validation and test accuracies. Model instability was also caused, in our view, by a large variety of graph structures which can possibly be resolved by graph pruning or graph unification.

6 Conclusions and Future Work

We presented an approach to the assessment of argument quality, in particular its cogency, evaluating the structural strength of the argumentation schemes applied by an arguer. Argumentation schemes were represented as graphs reconstructed by applying the NeuralEDUSeg model developed by Wang et al. (2018) to segment a text into elementary discourse units and the fine-tuned XLNet-large model of Yang et al. (2019) to classify discourse relations between the identified units. Both segmentation and classification models showed reasonable performance in processing argumentative texts: F1 scores of 47.94% on the segmentation task were achieved; discourse relation classification accuracy ranges from 50.48% to 60.22% depending on the classification scenario (5- vs 10-class discourse relation classification). Parsed argumentative texts subsequently were used to reconstruct discourse-based argumentation structures as graphs of varying complexity reflecting reasoning patterns that emulate human inferencing. Given a graph structure, the argument acceptability, relevance, sufficiency and overall cogency were predicted. The trained models incorporated not only linking structures but also claim and evidence content as node features computed from GloVe and BERT embeddings. We tentatively concluded that overall argument cogency may be predicted based on the argument structure alone without computing sophisticated node text-based features. To ensure that this observation is not an artefact of our data, it needs to be tested on a larger argument set from various domains.

Limitations of the presented study call for further improvements. The next major step is to incorporate edge features which contain important information about the type of relations between claim and evidence units. Further, we intend to explore new discourse processing tools and experiment with mapping between different discourse analysis frameworks, e.g. Rhetorical Structure Theory (RST, Mann and Thompson (1988) and ISO 24617-8 DR-Core (ISO, 2016). Additional argumentative corpora will be explored as well as the assessment of the other quality dimensions.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the first workshop on argumentation mining*, pages 64–68.

Dolores Albarracín, Aashna Sunderrajan, Sophie Lohmann, Man-Pui Sally Chan, and Duo Jiang. 2019. The psychology of attitudes, motivation, and persuasion. *The handbook of attitudes, volume 1: Basic principles*, pages 3–44.

Kevin Ashley, Niels Pinkwart, Collin Lynch, and Vincent Aleven. 2007. Learning by diagramming Supreme Court oral arguments. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, pages 271–275, Stanford, California. ACM.

Harry Bunt and Rashmi Prasad. 2016. ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th joint ACL-ISO workshop on interoperable semantic annotation (ISA-12)*, pages 45–54.

Erik Cambria, Praphul Chandra, Avinash Sharma, and Amir Hussain. 2010. Do not feel the trolls. In *Proceedings of the 3rd International Workshop on Social Data on the Web*, Shanghai, China.

Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 378–389. Springer.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

Federico Errica, Marco Podda, Davide Bacciu, and Alessio Micheli. 2019. A fair comparison of graph neural networks for graph classification. *arXiv preprint arXiv:1912.09893*.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? Choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. Association for Computational Linguistics.

Liyu Gong and Qiang Cheng. 2019. Exploiting edge features for graph neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9211–9219.

Trudy Govier. 2013. *A practical study of argument*. Cengage Learning.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1589–1599.

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.

Julia Hirschberg. 2002. The pragmatics of intonational meaning. In *Speech Prosody 2002, International Conference*.

Jos Hornikx. 2008. Comparing the actual and expected persuasiveness of evidence types: How good are lay people at selecting persuasive evidence? *Argumentation*, 22(4):555–569.

ISO. 2016. *ISO 24617-8, Semantic annotation framework (SemAF) Part 8, Semantic relations in discourse.* ISO Central Secretariat, Geneva.

Lu Ji, Zhongyu Wei, Xiangkun Hu, Yang Liu, Qi Zhang, and Xuan-Jing Huang. 2018. Incorporating argument-level interactions for persuasion comments evaluation using co-attention model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3703–3714.

Ralph Henry Johnson and J Anthony Blair. 2006. *Logical self-defense*. Idea.

Steven Johnson. 2009. *Winning Debates: A Guide to Debating in the Style of the World Universities Debating Championships*. G - Reference,Information and Interdisciplinary Subjects Series. International Debate Education Association, Brussels, Belgium.

Zixuan Ke, Winston Carlile, Nishant Gurrapadi, and Vincent Ng. 2018. Learning to give feedback: Modeling attributes affecting argument persuasiveness in student essays. In *IJCAI*, pages 4130–4136.

G Tarcan Kumkale and Dolores Albarracín. 2004. The sleeper effect in persuasion: a meta-analytic review. *Psychological bulletin*, 130(1):143.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281.

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

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. In *International Conference on Learning Representations (ICLR)*.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552.

Isaac Persing and Vincent Ng. 2017. Why can't you convince me? Modeling weaknesses in unpersuasive arguments. In *IJCAI*, pages 4082–4088.

Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion*, pages 1–24. Springer.

Volha Petukhova, Andrei Malchanau, and Harry Bunt. 2016. Modelling argumentative behaviour in parliamentary debates: data collection, analysis and test case. In M. Baldoni, C. Baroglio, F. Bex, F. Grasso, N. Green, M. Namazi-Rad, M.-R.and Numao, and M.T. Suarez, editors, *Principles and Practice of Multi-Agent Systems. Lecture Notes in Artificial Intelligence*, pages 26–46. Springer, Berlin.

Volha Petukhova, Tobias Mayer, Andrei Malchanau, and Harry Bunt. 2017a. Virtual debate coach design: assessing multimodal argumentation performance. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 41–50.

Volha Petukhova, Manoj Raju, and Harry Bunt. 2017b. Multimodal markers of persuasive speech: Designing a virtual debate coach. In *INTERSPEECH*, pages 142–146.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. 2005. The Penn Discourse TreeBank as a resource for natural language generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, pages 25–32.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, Hong Kong, China.

Matthew Rowe and Jonathan Butters. 2009. Assessing trust: contextual accountability. In *Proceedings of the First Workshop on Trust and Privacy on the Social and Semantic Web*, Heraklion, Greece.

Edwin Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable Bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 226–234. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. 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.

Simone Teufel. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh.

Frans Van Eemeren and Rob Grootendorst. 2004. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Li, and Yoshua Bengio. 2018. Graph attention networks. In *In Proceedings of the International Conference on Learning Representations*.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

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.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d‘Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

Appendix I: PDTB discourse relation tagsets and corpus distribution

binary	**L1 top-level relations**	**L2 fine-grained relations**	**# Instances**		
Rel	Expansion	*Conjunction*	8763	15116	35136
		Restatement	3326		
		Instantiation	1735		
		List	627		
		Alternative	531		
		Expansion	118		
		Exception	16		
	Comparison	*Contrast*	5947	7958	
		Concession	1425		
		Comparison	553		
		Pragmatic contrast	21		
		Pragmatic concession	12		
	Contingency	*Cause*	6203	7710	
		Condition	1359		
		Pragmatic cause	78		
		Pragmatic condition	68		
		Contingency	2		
	Temporal	*Asynchronous*	2739	4352	
		Synchrony	1607		
		Temporal	6		
NoRel		*NoRel*	5464	5464	5464

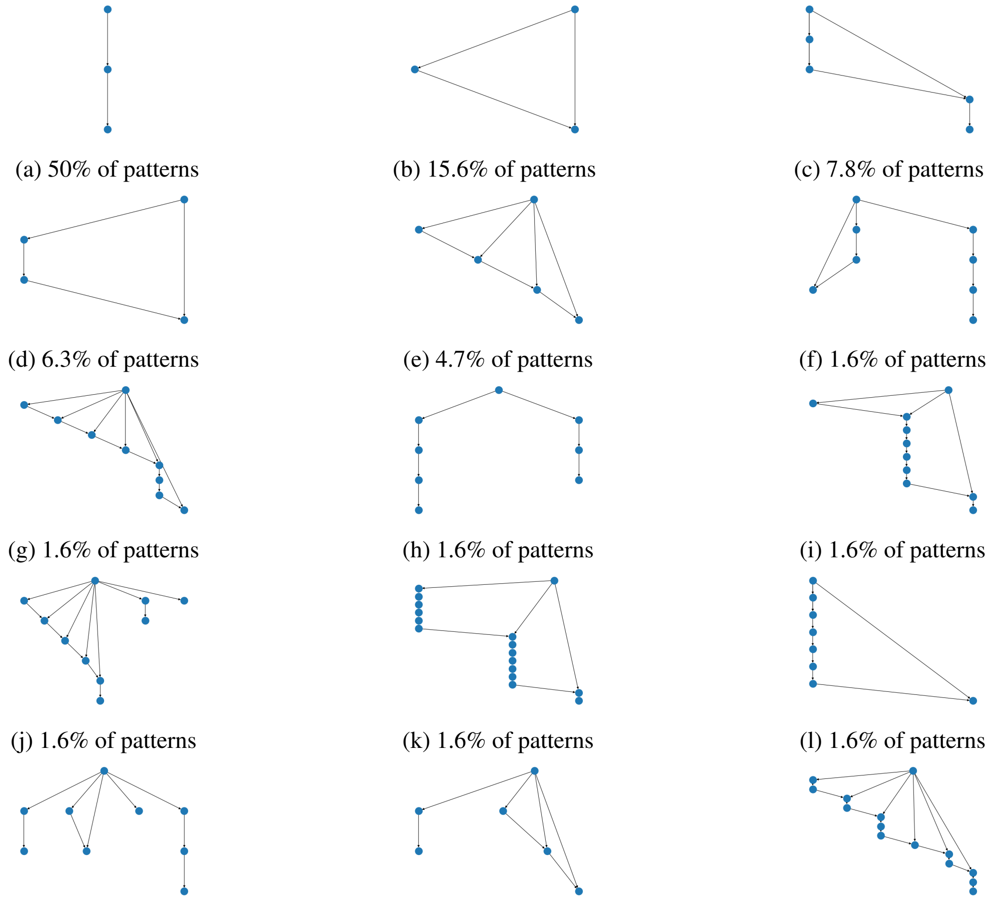
Table 4: The PDTB binary, top-level (L1) and fine-grained (L2) discourse relations and their distribution in PDTB 1.0 and 2.0 datasets. L2 relations marked italics were used for 10-class classification with XLNet.

Appendix II: Implicit Claim Reconstruction

Topic	**Claim**
Ban of plastic bottles	The consumption of water bottles should not be banned.
	The consumption of water bottles should be allowed only in the case of emergency.
Christianity or atheism	I choose atheism over Christianity and do not believe in God.
	I choose Christianity over atheism and do believe in God.
Evolution vs. creation	The world was created by God.
	The evolution is the beginning of life.
Personal pursuit or advancing the common good?	Advancing the common good is better than personal pursuit.
	Personal pursuit is better than advancing the common good.
Should physical education be mandatory in schools?	Physical education should not be mandatory in schools.
	Physical education should be mandatory in schools.
Is TV better than books?	Books are better than TV.
	TV is better than books.

Table 5: Examples of the claims reconstructed based on the corresponding Dagstuhl15512 ArgQuality topics.

Appendix III: Argument Graphs Examples



(a) 50% of patterns

(b) 15.6% of patterns

(c) 7.8% of patterns

(d) 6.3% of patterns

(e) 4.7% of patterns

(f) 1.6% of patterns

(g) 1.6% of patterns

(h) 1.6% of patterns

(i) 1.6% of patterns

(j) 1.6% of patterns

(k) 1.6% of patterns

(l) 1.6% of patterns

(m) 1.6% of patterns

(n) 1.6% of patterns

(o) 1.6% of patterns

Figure 6: Examples of the unique discourse-based argument schemes reconstructed from the Dagstuhl15512 ArgQuality corpus and their relative frequencies (in %).