Classifying Argumentative Relations Using Logical Mechanisms and Argumentation Schemes

Yohan Jo[1] **Seojin Bang**[1] **Chris Reed**[2] **Eduard Hovy**[1]

[1]School of Computer Science, Carnegie Mellon University, United States
[2]Centre for Argument Technology, University of Dundee, United Kingdom
[1]{yohanj,seojinb,ehovy}@andrew.cmu.edu, [2]c.a.reed@dundee.ac.kr

Abstract

While argument mining has achieved significant success in classifying argumentative relations between statements (support, attack, and neutral), we have a limited computational understanding of logical mechanisms that constitute those relations. Most recent studies rely on black-box models, which are not as linguistically insightful as desired. On the other hand, earlier studies use rather simple lexical features, missing logical relations between statements. To overcome these limitations, our work classifies argumentative relations based on four logical and theory-informed mechanisms between two statements, namely, (i) factual consistency, (ii) sentiment coherence, (iii) causal relation, and (iv) normative relation. We demonstrate that our operationalization of these logical mechanisms classifies argumentative relations without directly training on data labeled with the relations, significantly better than several unsupervised baselines. We further demonstrate that these mechanisms also improve supervised classifiers through representation learning.

1 Introduction

There have been great advances in argument mining—classifying the argumentative relation between statements as support, attack, or neutral. Recent research has focused on training complex neural networks on large labeled data. However, the behavior of such models remains obscure, and recent studies found evidence that those models may rely on spurious statistics of training data (Niven and Kao, 2019) and superficial cues irrelevant to the meaning of statements, such as discourse markers (Opitz and Frank, 2019). Hence, in this work, we turn to an *interpretable* method to investigate *logical relations* between statements, such as causal relations and factual contradiction. Such relations have been underemphasized in earlier studies (Feng and Hirst, 2011; Lawrence and Reed, 2016), possibly because their operationalization was unreliable then. Now that computational semantics is fast developing, our work takes a first step to computationally investigate how logical mechanisms contribute to building argumentative relations between statements and to classification accuracy with and without training on labeled data.

To investigate what logical mechanisms govern argumentative relations, we hypothesize that governing mechanisms should be able to classify the relations without directly training on relation-labeled data. Thus, we first compile a set of rules specifying logical and theory-informed mechanisms that signal the support and attack relations (§3). The rules are grouped into four mechanisms: factual consistency, sentiment coherence, causal relation, and normative relation. These rules are combined via probabilistic soft logic (PSL) (Bach et al., 2017) to estimate the optimal argumentative relations between statements. We operationalize each mechanism by training semantic modules on public datasets so that the modules reflect real-world knowledge necessary for reasoning (§4). For normative relation, we build a necessary dataset via rich annotation of the normative argumentation schemes *argument from consequences* and *practical reasoning* (Walton et al., 2008), by developing a novel and reliable annotation protocol (§5).

Our evaluation is based on arguments from kialo.com and debatepedia.org. We first demonstrate that the four logical mechanisms explain the argumentative relations between statements effectively. PSL with our operationalization of the mechanisms can classify the relations without

Transactions of the Association for Computational Linguistics, vol. 9, pp. 721–739, 2021. https://doi.org/10.1162/tacl_a_00394
Action Editor: Vincent Ng. Submission batch: 11/2020; Revision batch: 1/2021; Published 8/2021.

direct training on relation-labeled data, outperforming several unsupervised baselines (§7). We analyze the contribution and pitfalls of individual mechanisms in detail. Next, to examine whether the mechanisms can further inform supervised models, we present a method to learn vector representations of arguments that are ''cognizant of'' the logical mechanisms (§8). This method outperforms several supervised models trained without concerning the mechanisms, as well as models that incorporate the mechanisms in different ways. We illustrate how it makes a connection between logical mechanisms and argumentative relations. Our contributions are:

- An interpretable method based on PSL to investigate logical and theory-informed mechanisms in argumentation computationally.
- A representation learning method that incorporates the logical mechanisms to improve the predictive power of supervised models.
- A novel and reliable annotation protocol, along with a rich schema, for the argumentation schemes *argument from consequences* and *practical reasoning*. We release our annotation manuals and annotated data.[1]

2 Related Work

There has been active research in NLP to understand different mechanisms of argumentation computationally. Argumentative relations have been found to be associated with various statistics, such as discourse markers (Opitz and Frank, 2019), sentiment (Allaway and McKeown, 2020), and use of negating words (Niven and Kao, 2019). Further, as framing plays an important role in debates (Ajjour et al., 2019), different stances for a topic emphasize different points, resulting in strong thematic correlations (Lawrence and Reed, 2017).

Such thematic associations have been exploited in stance detection and dis/agreement classification. Stance detection (Allaway and McKeown, 2020; Stab et al., 2018; Xu et al., 2018) aims to classify a statement as pro or con with respect to a topic, while dis/agreement classification (Chen et al., 2018; Hou and Jochim, 2017; Rosenthal and McKeown, 2015) aims to decide whether two statements are from the same or opposite stance(s) for a given topic. Topics are usually discrete, and models often learn thematic correlations between a topic and a stance (Xu et al., 2019). Our work is slightly different as we classify the *direct* support or attack relation between two *natural* statements.

The aforementioned correlations, however, are byproducts rather than core mechanisms of argumentative relations. In order to decide whether a statement supports or attacks another, we cannot ignore the *logical* relation between them. Textual entailment was found to inform argumentative relations (Choi and Lee, 2018) and used to detect arguments (Cabrio and Villata, 2012). Similarly, there is evidence that the opinions of two statements toward the same concept constitute their argumentative relations (Gemechu and Reed, 2019; Kobbe et al., 2020). Causality between events also received attention, and causality graph construction was proposed for argument analysis (Al-Khatib et al., 2020). Additionally, in argumentation theory, Walton's argumentation schemes (Walton et al., 2008) specify common reasoning patterns people use to form an argument. This motivates our work to investigate logical mechanisms in four categories: factual consistency, sentiment coherence, causal relation, and normative relation.

Logical mechanisms have not been actively studied in argumentative relation classification. Models based on hand-crafted features have used relatively simple lexical features, such as n-grams, discourse markers, and sentiment agreement and word overlap between two statements (Stab and Gurevych, 2017; Habernal and Gurevych, 2017; Persing and Ng, 2016; Rinott et al., 2015). Recently, neural models have become dominant approaches (Chakrabarty et al., 2019; Durmus et al., 2019; Eger et al., 2017). Despite their high accuracy and finding of some word-level interactions between statements (Xu et al., 2019; Chen et al., 2018), they provide quite limited insight into governing mechanisms in argumentative relations. Indeed, more and more evidence suggests that supervised models learn to overly rely on superficial cues, such as discourse markers (Optiz and Frank, 2019), negating words (Niven and Kao, 2019), and sentiment (Allaway and McKeown, 2020) behind the scenes. We instead use an interpretable method based on PSL to examine logical mechanisms (§7) and then show

[1]The annotations, data, and source code are available at: https://github.com/yohanjo/tacl_arg_rel.

evidence that these mechanisms can inform supervised models in intuitive ways (§8).

Some research adopted argumentation schemes as a framework, making comparisons with discourse relations (Cabrio et al., 2013) and collecting and leveraging data at varying degrees of granularity. At a coarse level, prior studies annotated the presence of particular argumentation schemes in text (Visser et al., 2020; Lawrence et al., 2019; Lindahl et al., 2019; Reed et al., 2008) and developed models to classify different schemes (Feng and Hirst, 2011). However, each scheme often accommodates both support and attack relations between statements, so classifying those relations requires semantically richer information within the scheme than just its presence. To that end, Reisert et al. (2018) annotated individual components within schemes, particularly emphasizing *argument from consequences*. Based on the logic behind this scheme, Kobbe et al. (2020) developed an unsupervised method to classify the support and attack relations using syntactic rules and lexicons. Our work extends these studies by including other normative schemes (*practical reasoning* and property-based reasoning) and annotating richer information.

3 Rules

We first compile rules that specify evidence for the support and attack relations between **claim** C and **statement** S (Table 1).[2] These rules are combined via PSL (Bach et al., 2017) to estimate the optimal relation between C and S.[3]

We will describe individual rules in four categories: factual consistency, sentiment coherence, causal relation, and normative relation, followed by additional chain rules.

3.1 Factual Consistency

A statement that supports the claim may present a fact that naturally entails the claim, while an attacking statement often presents a fact

		Rules
Factual Consist.	R1	$FactEntail(S,C) \rightarrow Support(S,C)$
	R2	$FactContradict(S,C) \rightarrow Attack(S,C)$
	R3	$FactConflict(S,C) \rightarrow Attack(S,C)$
Senti Cohe.	R4	$SentiConflict(S,C) \rightarrow Attack(S,C)$
	R5	$SentiCoherent(S,C) \rightarrow Support(S,C)$
Causal Relation		***CAUSE-TO-EFFECT REASONING***
	R6	$Cause(S,C) \rightarrow Support(S,C)$
	R7	$Obstruct(S,C) \rightarrow Attack(S,C)$
		EFFECT-TO-CAUSE REASONING
	R8	$Cause(C,S) \rightarrow Support(S,C)$
	R9	$Obstruct(C,S) \rightarrow Attack(S,C)$
Normative Relation		***ARGUMENT FROM CONSEQUENCES***
	R10	$BackingConseq(S,C) \rightarrow Support(S,C)$
	R11	$RefutingConseq(S,C) \rightarrow Attack(S,C)$
		PRACTICAL REASONING
	R12	$BackingNorm(S,C) \rightarrow Support(S,C)$
	R13	$RefutingNorm(S,C) \rightarrow Attack(S,C)$
Relation Chain	R14	$Support(S,I) \wedge Support(I,C) \rightarrow Support(S,C)$
	R15	$Attack(S,I) \wedge Attack(I,C) \rightarrow Support(S,C)$
	R16	$Support(S,I) \wedge Attack(I,C) \rightarrow Attack(S,C)$
	R17	$Attack(S,I) \wedge Support(I,C) \rightarrow Attack(S,C)$
Const-raints	C1	Neutral(S,C) = 1
	C2	Support(S,C)+Attack(S,C)+Neutral(S,C) = 1

Table 1: PSL rules. (S: statement, C: claim).

contradictory or contrary to the claim. For example:

> **Claim:** *Homeschooling deprives children and families from interacting with people with different religions, ideologies, or values.*
>
> **Support Statement:** *Homeschool students have few opportunities to meet diverse peers they could otherwise see at normal schools.*
>
> **Attack Statement:** *Homeschool students can interact regularly with other children from a greater diversity of physical locations, allowing them more exposure outside of their socio-economic group.*

This logic leads to two rules:

R1: FactEntail$(S,C) \rightarrow$ Support(S,C),

R2: FactContradict$(S,C) \rightarrow$ Attack(S,C)

s.t. FactEntail$(S,C) = P(S$ entails $C)$,

FactContradict$(S,C) = P(S$ contradicts $C)$.

In our work, these probabilities are computed by a textual entailment module (§4.1).

[2] We do not assume that claim-hood and statement-hood are intrinsic features of text spans; we follow prevailing argumentation theory in viewing claims and statements as roles determined by virtue of relationships between text spans.

[3] Predicates in the rules are probability scores, and PSL aims to estimate the scores of Support(S,C), Attack(S,C), and Neutral(S,C) for all (S,C). The degree of satisfaction of the rules are converted to a loss, which is minimized via maximum likelihood estimation.

In argumentation, it is often the case that an attacking statement and the claim are not strictly contradictory nor contrary, but the statement contradicts only a specific part of the claim, as in:

> **Claim:** *Vegan diets are healthy.*
>
> **Attack Statement:** *Meat is healthy.*

Formally, let $(A^S_{i,0}, A^S_{i,1}, \cdots)$ denote the ith relation tuple in S, and $(A^C_{j,0}, A^C_{j,1}, \cdots)$ the jth relation tuple in C. We formulate the conflict rule:

$$\textbf{R3: } \text{FactConflict}(S, C) \rightarrow \text{Attack}(S, C)$$

$$\textit{s.t. } \text{FactConflict}(S, C) = \max_{i,j,k} P(A^S_{i,k} \text{ contradicts } A^C_{j,k}) \prod_{k' \neq k} P(A^S_{i,k'} \text{ entails } A^C_{j,k'}).$$

We use Open IE 5.1 to extract relation tuples, and the probability terms are computed by a textual entailment module (§4.1).

3.2 Sentiment Coherence

When S attacks C, they may express opposite sentiments toward the same target, whereas they may express the same sentiment if S supports C (Gemechu and Reed, 2019). For example:

> **Claim:** *Pet keeping is morally justified.*
>
> **Attack Statement:** *Keeping pets is hazardous and offensive to other people.*
>
> **Support Statement:** *Pet owners can provide safe places and foods to pets.*

Let (t^S_i, s^S_i) be the ith expression of sentiment $s^S_i \in \{\text{pos}, \text{neg}, \text{neu}\}$ toward target t^S_i in S, and (t^C_j, s^C_j) the jth expression in C. We formulate two rules:

$$\textbf{R4: } \text{SentiConflict}(S, C) \rightarrow \text{Attack}(S, C),$$
$$\textbf{R5: } \text{SentiCoherent}(S, C) \rightarrow \text{Support}(S, C)$$

$$\textit{s.t. } \text{SentiConflict}(S, C) = \max_{i,j} P(t^S_i = t^C_j) \left\{ P(s^S_i = \text{pos}) P(s^C_j = \text{neg}) + P(s^S_i = \text{neg}) P(s^C_j = \text{pos}) \right\},$$
$$\text{SentiCoherent}(S, C) = \max_{i,j} P(t^S_i = t^C_j) \left\{ P(s^S_i = \text{pos}) P(s^C_j = \text{pos}) + P(s^S_i = \text{neg}) P(s^C_j = \text{neg}) \right\}.$$

In this work, targets are all noun phrases and verb phrases in C and S. $P(t^S_i = t^C_j)$ is computed by a textual entailment module (§4.1), and $P(s^S_i)$ and $P(s^C_j)$ by a target-based sentiment classifier (§4.2).

3.3 Causal Relation

Reasoning based on causal relation between events is used in two types of argumentation: *argument from cause to effect* and *argument from effect to cause* (Walton et al., 2008). In cause-to-effect (C2E) reasoning, C is derived from S because the event in S may cause that in C. If S causes (obstructs) C then S is likely to support (attack) C. For example:

> **Claim:** *Walmart's stock price will rise.*
>
> **Support Statement:** *Walmart generated record revenue.*
>
> **Attack Statement:** *Walmart had low net incomes.*

This logic leads to two rules:

$$\textbf{R6: } \text{Cause}(S, C) \rightarrow \text{Support}(S, C),$$
$$\textbf{R7: } \text{Obstruct}(S, C) \rightarrow \text{Attack}(S, C),$$

$$\textit{s.t. } \text{Cause}(S, C) = P(S \text{ causes } C),$$
$$\text{Obstruct}(S, C) = P(S \text{ obstructs } C).$$

Effect-to-cause (E2C) reasoning has the reversed direction; S describes an observation and C is a reasonable explanation that may have caused it. If C causes (obstructs) S, then S is likely to support (attack) C, as in:

> **Claim:** *St. Andrew Art Gallery is closing soon.*
>
> **Support Statement:** *The number of paintings in the gallery has reduced by half for the past month.*
>
> **Attack Statement:** *The gallery recently bought 20 photographs.*

$$\textbf{R8: } \text{Cause}(C, S) \rightarrow \text{Support}(S, C),$$
$$\textbf{R9: } \text{Obstruct}(C, S) \rightarrow \text{Attack}(S, C).$$

The probabilities are computed by a causality module (§4.3).

3.4 Normative Relation

In argumentation theory, Walton's argumentation schemes specify common reasoning patterns used in arguments (Walton et al., 2008). We focus on two schemes related to normative arguments, whose claims suggest that an action or situation be brought about. Normative claims are one of the most common proposition types in argumentation (Jo et al., 2020) and have received much attention in the literature (Park and Cardie, 2018).

Argument from Consequences: In this scheme, the claim is supported or attacked by a positive or negative consequence, as in:

> **Claim:** *Humans should stop eating animal meat.*
>
> **Support Statement:** *The normalizing of killing animals for food leads to a cruel mankind.* (S1)
>
> **Attack Statement:** *Culinary arts developed over centuries may be lost.* (S2)

In general, an argument from consequences may be decomposed into two parts: (i) whether S is a positive consequence or a negative one; and (ii) whether the source of this consequence is consistent with or facilitated by C's stance (S2), or is contrary to or obstructed by it (S1).

Logically, S is likely to support C by presenting a positive (negative) consequence of a source that is consistent with (contrary to) C's stance. In contrast, S may attack C by presenting a negative (positive) consequence of a source that is consistent with (contrary to) C's stance. Given that S describes consequence Q of source R, this logic leads to:

R10: $\text{BackingConseq}(S, C) \rightarrow \text{Support}(S, C)$,
R11: $\text{RefutingConseq}(S, C) \rightarrow \text{Attack}(S, C)$

s.t. $\text{BackingConseq}(S, C) =$
$P(S \text{ is a consequence}) \times$
$\{P(Q \text{ is positive}) \cdot P(R \text{ consistent with } C)$
$+ P(Q \text{ is negative}) \cdot P(R \text{ contrary to } C)\}$,
$\text{RefutingConseq}(S, C) =$
$P(S \text{ is a consequence}) \times$
$\{P(Q \text{ is negative}) \cdot P(R \text{ consistent with } C)$
$+ P(Q \text{ is positive}) \cdot P(R \text{ contrary to } C)\}$.

Practical Reasoning: In this scheme, the statement supports or attacks the claim by presenting a goal to achieve, as in:

> **Claim:** *Pregnant people should have the right to choose abortion.*
>
> **Support Statement:** *Women should be able to make choices about their bodies.* (S3)
>
> **Attack Statement:** *Our rights do not allow us to harm the innocent lives of others.* (S4)

The statements use a normative statement as a goal to justify their stances. We call their target of advocacy or opposition (underlined above) a **norm target**. Generally, an argument of this scheme may be decomposed into: (i) whether S advocates for its norm target (S3) or opposes it (S4), as if expressing positive or negative sentiment toward the norm target; and (ii) whether the norm target is a situation or action that is consistent with or facilitated by C's stance, or that is contrary to or obstructed by it.[4]

Logically, S is likely to support C by advocating for (opposing) a norm target that is consistent with (contrary to) C's stance. In contrast, S may attack C by opposing (advocating for) a norm target that is consistent with (contrary to) C's stance. Given that S has norm target R, this logic leads to:

R12: $\text{BackingNorm}(S, C) \rightarrow \text{Support}(S, C)$,
R13: $\text{RefutingNorm}(S, C) \rightarrow \text{Attack}(S, C)$

s.t. $\text{BackingNorm}(S, C) =$
$P(S \text{ is normative}) \times$
$\{P(S \text{ advocates for } R) \cdot P(R \text{ consistent with } C)$
$+ P(S \text{ opposes } R) \cdot P(R \text{ contrary to } C)\}$,
$\text{RefutingNorm}(S, C) =$
$P(S \text{ is normative}) \times$
$\{P(S \text{ opposes } R) \cdot P(R \text{ consistent with } C)$
$+ P(S \text{ advocates for } R) \cdot P(R \text{ contrary to } C)\}$.

The probabilities are computed by modules trained on our annotation data (§5).

[4] Both harming innocent lives and making choices about their bodies are facilitated by the right to choose abortion ('consistent').

3.5 Relation Chain

A chain of argumentative relations across arguments may provide information about the plausible relation within each argument. Given three statements S, I, and C, we have four chain rules:

R14: $\text{Support}(S, I) \wedge \text{Support}(I, C) \rightarrow \text{Support}(S, C)$,
R15: $\text{Attack}(S, I) \wedge \text{Attack}(I, C) \rightarrow \text{Support}(S, C)$,
R16: $\text{Support}(S, I) \wedge \text{Attack}(I, C) \rightarrow \text{Attack}(S, C)$,
R17: $\text{Attack}(S, I) \wedge \text{Support}(I, C) \rightarrow \text{Attack}(S, C)$.

For each data split, we combine two neighboring arguments where the claim of one is the statement of the other, whenever possible. The logical rules R1–R13 are applied to these ''indirect'' arguments.

3.6 Constraints

C and S are assumed to have the neutral relation (or the attack relation for binary classification) if they do not have strong evidence from the rules mentioned so far (Table 1 **C1**). In addition, the probabilities of all relations should sum to 1 (**C2**).

4 Modules

In this section, we discuss individual modules for operationalizing the PSL rules. For each module, we fine-tune the pretrained uncased BERT-base (Devlin et al., 2019). We use the Transformers library v3.3.0 (Wolf et al., 2020) for high reproducibility and low development costs. But any other models could be used instead.

Each dataset used is randomly split with a ratio of 9:1 for training and test. Cross-entropy and Adam are used for optimization. To address the imbalance of classes and datasets, the loss for each training instance is scaled by a weight inversely proportional to the number of its class and dataset.

4.1 Textual Entailment

A textual entailment module is used for rules about factual consistency and sentiment coherence (R1–R5). Given a pair of texts, it computes the probabilities of entailment, contradiction, and neutral.

Our training data include two public datasets: MNLI (Williams et al., 2018) and AntSyn (Nguyen et al., 2017) for handling antonyms and synonyms. An NLI module combined with the word-level entailment handles short phrases better without hurting accuracy for sentence-level entailment. Since AntSyn does not have the neutral class, we add 50K neutral word pairs by randomly pairing two words among the 20K most frequent words in MNLI; without them, a trained model can hardly predict the neutral relation between words. The accuracy for each dataset is in Table 2 rows 1–4.

		Dataset (Classes, N)	Accuracy
Textual Entailment (R1–R5)	1	MNLI (ent/con/neu, 412,349)	F1=82.3
	2	AntSyn (ent/con, 15,632)	F1=90.2
	3	Neu50K (neu, 50,000)	R=97.5
	4	**MicroAvg (ent/con/neu, 477,981)**	**F1=84.7**
Sentiment Classification (R4–R5)	5	SemEval17 (pos/neg/neu, 20,632)	F1=64.5
	6	Dong (pos/neg/neu, 6,940)	F1=71.4
	7	Mitchell (pos/neg/neu, 3,288)	F1=62.5
	8	Bakliwal (pos/neg/neu, 2,624)	F1=69.7
	9	Norm (pos/neg, 632)	F1=100.0
	10	**MicroAvg (pos/neg/neu, 34,116)**	**F1=69.2**
Causality (R6–R9)	11	PDTB (cause/else, 14,224)	F1=68.1
	12	PDTB-R (cause/else 1,791)	F1=75.7
	13	BECauSE (cause/obstruct, 1,542)	F1=46.1
	14	BECauSE-R (else, 1,542)	R=86.5
	15	CoNet (cause, 50,420)	R=88.6
	16	CoNet-R (else, 50,420)	R=91.7
	17	WIQA (cause/obstruct, 31,630)	F1=88.2
	18	WIQA-P (else, 31,630)	R=90.2
	19	**MicroAvg (cause/obstr/else, 183,119)**	**F1=87.7**
Normative Relation (R10–R13)	20	JustType (conseq/norm, 1,580)	F1=90.2
	21	ConseqSenti (pos/neg, 824)	F1=71.8
	22	NormType (adv/opp, 758)	F1=91.1
	23	RC-Rel (consist/contra/else, 1,924)	F1=70.1

Table 2: F1-scores and recall of modules.

4.2 Target-Based Sentiment Classification

A sentiment classifier is for rules about sentiment coherence (R4–R5). Given a pair of texts T_1 and T_2, it computes the probability of whether T_1 has positive, negative, or neutral sentiment toward T_2.

Our training data include five datasets for target-based sentiment classification: SemEval17 (Rosenthal et al., 2017), entities (Dong et al., 2014), open domain (Mitchell et al., 2013), Irish politics (Bakliwal et al., 2013), and our annotations of positive/negative norms toward norm targets (§5.1). These annotations highly improve classification of sentiments expressed through advocacy and opposition in normative statements. Pretraining on general sentiment resources–subjectivity lexicon (Wilson et al., 2005) and sentiment140 (Go et al., 2009)–also helps (Table 2 rows 5–10).

Corpus	Corpus-Specific Labels	Our Label (N)
PDTB	Temporal.Asynchronous	Cause (1,255)
	Temporal.Synchrnonous	Cause (536)
	Comparison, Expansion	Else (12,433)
PDTB-R†	Temporal.Asynchronous	Else (536)
	Temporal.Synchronous	Cause (1,255)
BECauSE	Promote	Cause (1,417)
	Inhibit	Obstruct (142)
BECauSE-R†	Promote, Inhibit	Else (1,613)
WIQA	RESULTS_IN	Cause (12,652)
	NOT_RESULTS_IN	Obstruct (18,978)
WIQA-P‡	RESULSTS_IN, NOT_RESULTS_IN	Else (31,630)
ConceptNet	Causes, CausesDesire, HasFirstSubevent, HasLast-Subevent, HasPrerequisite	Cause (50,420)
ConceptNet-R†	Causes, CausesDesire, HasFirstSubevent, HasLast-Subevent, HasPrerequisite	Else (50,420)

Table 3: Mapping between corpus-specific labels and our labels for the causality module. †The order of two input texts are reversed. ‡The second input text is replaced with a random text in the corpus.

4.3 Causality

A causality module is used for rules regarding causal relations (R6–R9). Given an input pair of texts T_1 and T_2, it computes the probability of whether T_1 causes T_2, obstructs T_2, or neither.

Our training data include four datasets about causal and temporal relations between event texts. PDTB 3.0 (Webber et al., 2006) is WSJ articles annotated with four high-level discourse relations, and we map the sub-relations of 'Temporal' to our classes.[5] BECauSE 2.0 (Dunietz et al., 2017) is news articles annotated with linguistically marked causality. WIQA (Tandon et al., 2019) is scientific event texts annotated with causality between events. ConceptNet(Speer et al., 2017) is a knowledge graph between phrases, and relations about causality are mapped to our classes. To prevent overfitting to corpus-specific characteristics, we add adversarial data by swapping two input texts (PDTB-R, BECauSE-R, ConceptNet-R) or pairing random texts (WIQA-P). The mapping between corpus-specific labels and ours is in Table 3, and the module accuracy in Table 2 rows 11–19.

[5]We use explicit relations only for pretraining, since they often capture linguistically marked, rather than true, relations between events. We also exclude the Contingency relations as causal and non-causal relations (e.g., justification) are mixed.

4.4 Normative Relation

All the modules here are trained on our annotations of normative argumentation schemes (§5).

$P(S$ is a consequence / norm) (R10–R13): Given a statement, one module computes the probability that it is a consequence, and another module the probability of a norm. Both modules are trained on all claims and statements in our annotations, where all claims are naturally norms, and each statement is annotated as either norm or consequence (Table 2 row 20).

$P(Q$ is positive / negative) (R10–R11): Given a statement assumed to be a consequence, this module computes the probability of whether it is positive or negative. It is trained on all statements annotated as consequence (Table 2 row 21).

$P(S$ advocates / opposes) (R12–R13): Given a statement assumed to be a norm, this module computes the probability of whether it is advocacy or opposition. It is trained on all claims, plus statements annotated as norm (Table 2 row 22).

$P(R$ consistent / contrary to $C)$ (R10–R13): For a pair of S and C, the module computes the probability of whether R (the norm target or the source of consequence in S) and C's stance are consistent, contrary, or else. In our annotations, R and C are 'consistent' if both (1a and 3a in Figure 1) are advocacy or opposition, and 'contrary' otherwise. To avoid overpredicting the two classes, we add negative data by pairing C with a random statement in the annotations. The module is pretrained on MNLI and AntSyn (Table 2 row 23).

5 Annotation of Normative Argumentation Schemes

In this section, we discuss our annotation of the argumentation schemes *argument from consequences* and *practical reasoning* (Figure 1). The resulting annotations are used to train the modules in §4.4 that compute the probability terms in R10–R13.

For each pair of normative claim C and statement S, we annotate the following information: (1a) Whether C advocates for or opposes its norm target, and (1b) the norm target T (Figure 1 TASK 1); (2a) Whether S uses a norm, consequence, or property for justification, and (2b) the justification J (Figure 1 TASK 2); (3a) Whether J's focus is on

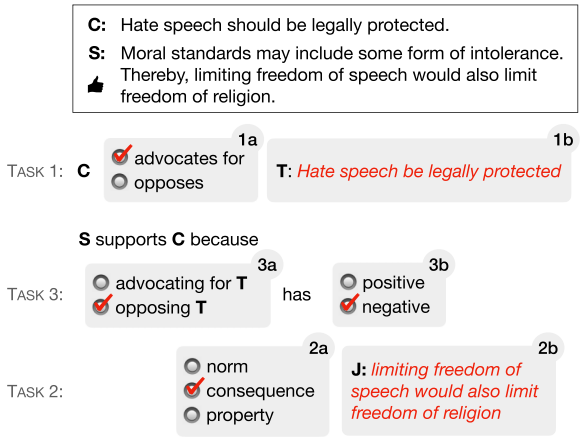


Figure 1: Example annotations (checks and italic) of the normative argumentation schemes. It depends on the argument whether S supports or attacks C.

advocating for T or opposing T, and (3b) whether J is positive or negative (Figure 1 TASK 3).[6]

Our annotation schema is richer than existing ones (Lawrence and Reed, 2016; Reisert et al., 2018). Due to the increased complexity, however, our annotation is split into three pipelined tasks. For this annotation, we randomly sampled 1,000 arguments from Kialo whose claims are normative (see §6 and Table 4 for details).

5.1 Task 1. Norm Type/Target of Claim

For each C, we annotate: (1a) the norm type—advocate, oppose, or neither—toward its norm target; and (1b) the norm target T. Advocacy is often expressed as ‘‘should/need T’’, whereas opposition as ‘‘should not T’’, ‘‘T should be banned’’; ‘neither’ is noise (2.8%) to be discarded. T is annotated by rearranging words in C (Figure 1 TASK 1).

There are 671 unique claims in the annotation set. The first author of this paper wrote an initial manual and trained two undergraduate students majoring in economics, while resolving disagreements through discussion and revising the manual. In order to verify that the annotation can be conducted systematically, we measured inter-annotator agreement (IAA) on 200 held-out claims. The annotation of norm types achieved

[6]This annotation schema provides enough information for the classifiers in §4.4. $P(S$ is a consequence / norm) is from (2a), and both $P(Q$ is positive / negative) and $P(S$ advocates / opposes) are from (3b). $P(R$ consistent / contrary to $C)$ can be obtained by combining (1a) and (3a): ‘consistent’ if both advocate or both oppose, and ‘contrary’ otherwise.

		Kialo			Debatepedia		
		Annotation	Fit	Val	Test	Fit	Val	Test
Normative	Sup	480	4,621	1,893	6,623	6,598	229	356
	Att	520	5,383	2,124	7,623	4,502	243	351
	Neu	–	9,984	4,000	14,228	–	–	–
Non-normative	Sup	–	4,953	10,135	21,138	3,302	243	178
	Att	–	5,043	9,848	20,197	3,278	253	152
	Neu	–	10,016	20,000	40,947	–	–	–

Table 4: Numbers of arguments in datasets.

Krippendorff’s α of 0.81. To measure IAA for annotation of T, we first aligned words between each annotation and the claim, obtaining a binary label for each word in the claim (1 if included in the annotation). As a result, we obtained two sequences of binary labels of the same length from the two annotators and compared them, achieving an F1-score of 0.89. The high α and F1-score show the validity of the annotations and annotation manual. All disagreements were resolved through discussion afterward.[7]

5.2 Task 2. Justification Type of Premise

For each pair of C and S, we annotate: (2a) the justification type of S—norm, consequence, property, or else; and (2b) the justification J. The justification types are defined as follows:

- **Norm:** J states that some situation or action should be achieved (practical reasoning).
- **Consequence:** J states a potential or past outcome (argument from consequences).
- **Property:** J states a property that (dis)qualifies C’s stance (argument from consequence).

The difference between consequence and property is whether the focus is on extrinsic outcomes or intrinsic properties, such as feasibility, moral values, and character (e.g., ‘‘Alex shouldn’t be the team leader because he is dishonest’’). We consider both as argument from consequences

[7]These annotations are used for the sentiment classifiers in §4.2, too. For example, ‘‘the lottery should be banned’’ is taken to express negative sentiment toward the lottery. Such examples are underrepresented in sentiment datasets, resulting in inaccurate sentiment classification for normative statements.

because property-based justification has almost the same logic as consequence-based justification. The 'else' type is rare (3.4%) and discarded after the annotation.

The process of annotation and IAA measurement is the same as Task 1, except that IAA was measured on 100 held-out arguments due to a need for more training. For justification types, Krippendorff's α is 0.53—moderate agreement. For justification J, the F1-score is 0.85. The relatively low IAA for justification types comes from two main sources. First, a distinction between consequence and property is fuzzy by nature, as in ''an asset tax is the most fair system of taxing citizens''. This difficulty has little impact on our system, however, as both are treated as argument from consequences.

Second, some statements contain multiple justifications of different types. If so, we asked the annotators to choose one that they judge to be most important (for training purposes). They sometimes chose different justifications, although they usually annotated the type correctly for the chosen one.

5.3 Task 3. Justification Logic of Statement

Given C with its norm target T, and S with its justification J, we annotate: (3a) whether the consequence, property, or norm target of J is regarding advocating for T or opposing T; and (3b) whether J is positive or negative. J is positive (negative) if it's a positive (negative) consequence/property or expresses advocacy (opposition).

This task was easy, so only one annotator worked with the first author. Their agreement measured on 400 heldout arguments is Krippendorff's α of 0.82 for positive/negative and 0.78 for advocate/oppose.

5.4 Analysis of Annotations

We obtained 962 annotated arguments with claims of advocacy (70%) and opposition (30%), and statements of consequence (54%), property (32%), and norm (14%). Supporting statements are more likely to use a positive justification (62%), while attacking statements a negative one (68%), with significant correlations ($\chi^2 = 87, p < .00001$). But 32–38% of the time, they use the opposite sentiment, indicating that sentiment alone cannot determine argumentative relations.

6 Data

6.1 Kialo

Our first dataset is from kialo.com, a collaborative argumentation platform covering contentious topics. Users contribute to the discussion of a topic by creating a statement that either supports or attacks an existing statement, resulting in an argumentation tree for each topic. We define an **argument** as a pair of parent and child statements, where the parent is the **claim** and the child is the **support or attack statement**. Each argument is labeled with support or attack by users and is usually self-contained, not relying on external context, anaphora resolution, or discourse markers.

We scraped arguments for 1,417 topics and split into two subsets. **Normative arguments** have normative claims suggesting that a situation or action be brought about, while **non-normative arguments** have non-normative claims. This distinction helps us understand the two types of arguments better. We separated normative and non-normative claims using a BERT classifier trained on Jo et al.'s (2020) dataset of different types of statements (AUC=98.8%), as binary classification of normative statement or not. A claim is considered normative (non-normative) if the predicted probability is higher than 0.97 (lower than 0.4); claims with probability scores between these thresholds (total 10%) are discarded to reduce noise.

In practice, an argument mining system may also need to identify statements that seem related but do not form any argument. Hence, we add the same number of ''neutral arguments'' by pairing random statements within the same topic. To avoid paired statements forming a reasonable argument accidentally, we constrain that they be at least 9 statements apart in the argumentation tree, making them unlikely to have any support or attack relation but still topically related to each other.

Among the resulting arguments, 10K are reserved for fitting; 20% or 30% of the rest (depending on the data size) are used for validation and the others for test (Table 4). We increase the validity of the test set by manually discarding non-neutral arguments from the neutral set. We also manually inspect the normativity of claims, and if they occur in the fitting or validation sets too, the corresponding arguments are assigned to

the correct sets according to the manual judgments. For normative arguments, we set aside 1,000 arguments for annotating the argumentation schemes (§5).

The data cover the domains economy (13%), family (11%), gender (10%), crime (10%), rights (10%), God (10%), culture (10%), entertainment (7%), and law (7%), as computed by LDA. The average number of words per argument is 49 (45) for normative (non-normative) arguments.

6.2 Debatepedia

The second dataset is Debatepedia arguments (Hou and Jochim, 2017). A total of 508 topics are paired with 15K pro and con responses, and we treat each pair as an **argument** and each topic and response as **claim** and **statement**, respectively.

One important issue is that most topics are in question form, either asking if you agree with a stance (''yes'' is pro and ''no'' is con) or asking to choose between two options (the first is pro and the second is con). Since our logical mechanisms do not handle such questions naturally, we convert them to declarative claims as follows. The first type of questions are converted to a claim that proposes the stance (e.g., ''Should Marijuana be legalized?'' to ''Marijuana should be legalized''), and the second type of questions to a claim that prefers the first option (e.g., ''Mission to the Moon or Mars?'' to ''Mission to the Moon is preferred to Mars''). The first author and an annotator converted all topics independently and then resolved differences.

We split the arguments into **normative** and **non-normative** sets as we do for Kialo, manually verifying all claims. There is no neutral relation. We use the original train, validation, and test splits (Table 4). Debatepedia claims are shorter and less diverse than Kialo claims. They focus mostly on valuation, while Kialo includes mostly factual claims.

7 Experiment 1. Probabilistic Soft Logic

The goal here is to see how well the logical mechanisms alone can explain argumentative relations.

7.1 PSL Settings

We use the PSL toolkit v2.2.1.[8] The initial weights of the logical rules R1–R13 are set to 1. The importance of the chain rules R14–R17 may be different, so we explore {1, 0.5, 0.1}. The weight of C1 serves as a threshold for the default relation (i.e., neutral for Kialo and attack for Debatepedia), and we explore {0.2, 0.3}; initial weights beyond this range either ignore or overpredict the default relation. C2 is a hard constraint. The optimal weights are selected by the objective value on the validation set (this does not use true relation labels).

7.2 Baselines

We consider three baselines. **Random** assigns a relation to each argument randomly. **Sentiment** assigns a relation based on the claim and statement's agreement on sentiment: support if both are positive or negative, attack if they have opposite sentiments, and neutral otherwise. We compute a sentiment distribution by averaging all target-specific sentiments from our sentiment classifier (§4.2). **Textual entailment** assigns support (attack) if the statement entails (contradicts) the claim, and neutral otherwise (Cabrio and Villata 2012). We use our textual entailment module (§4.1). For Debatepedia, we choose between support and attack whichever has a higher probability.

7.3 Results

Tables 5a and 5b summarize the accuracy of all models for Kialo and Debatepedia, respectively. Among the baselines, sentiment (row 2) generally outperforms textual entailment (row 3), both significantly better than random (row 1). Sentiment tends to predict the support and attack relations aggressively, missing many neutral arguments, whereas textual entailment is conservative and misses many support and attack arguments. PSL with all logical rules R1–R13 (row 4) significantly outperforms all the baselines with high margins, and its F1-scores are more balanced across the relations.

To examine the contribution of each logical mechanism, we conducted ablation tests (rows 5–8). The most contributing mechanism is clearly normative relation across all settings, without which F1-scores drop by 2.6–4.8 points (row 8).

[8] https://psl.linqs.org/wiki/2.2.1/.

	Normative Arguments						Non-normative Arguments					
	ACC	AUC	F1	$F1_{sup}$	$F1_{att}$	$F1_{neu}$	ACC	AUC	F1	$F1_{sup}$	$F1_{att}$	$F1_{neu}$
1 Random	33.5	50.2	32.6	27.8	30.1	39.9	33.4	49.9	32.5	28.7	28.8	40.0
2 Sentiment	40.8	64.1	40.7	40.6	39.1	42.4	43.7	61.1	42.2	40.0	35.2	51.5
3 Text Entail	51.8	61.8	36.7	12.8	30.4	67.0	52.1	62.8	38.6	18.4	31.0	66.4
4 PSL (R1–R13)	54.0‡	73.8‡	52.1‡	47.0‡	43.6‡	65.7‡	57.0‡	76.0‡	54.0‡	50.1‡	42.6‡	69.3‡
5 \ Fact	55.1‡	74.3‡	52.4‡	47.1‡	41.6‡	68.4‡	58.6‡	77.1‡	55.1‡	50.5‡	42.2‡	72.7‡
6 \ Sentiment	**62.1**‡	77.6‡	57.5‡	49.1‡	45.8‡	**77.7**‡	61.3‡	77.8‡	56.7‡	50.3‡	44.1‡	75.7‡
7 \ Causal	54.4‡	73.1‡	52.3‡	45.4‡	45.4‡	66.0‡	57.6‡	76.1‡	54.3‡	48.7‡	43.4‡	70.7‡
8 \ Normative	51.8‡	68.6‡	49.4‡	44.3‡	40.4†	63.4‡	54.7‡	70.3‡	51.4‡	47.0‡	40.3‡	66.8‡
9 \ Sentiment + Chain	61.9‡	**77.7**‡	**57.7**‡	**49.3**‡	**46.2**‡	77.6‡	**61.5**‡	**78.0**‡	**57.2**‡	**50.8**‡	**44.7**‡	**76.1**‡

(a) Kialo

	Normative Arguments					Non-normative Arguments				
	ACC	AUC	F1	$F1_{sup}$	$F1_{att}$	ACC	AUC	F1	$F1_{sup}$	$F1_{att}$
1 Random	47.7	49.4	50.2	49.0	51.4	53.0	54.6	52.4	53.7	51.1
2 Sentiment	59.3	63.9	59.2	61.0	57.4	69.1	73.4	68.5	72.7	64.3
3 Text Entail	52.2	55.8	49.4	37.6	61.2	70.6	74.2	70.5	69.0	72.0
4 PSL (R1–R13)	**63.9***	**68.3***	**63.9***	63.8	64.0†	73.0	76.1	73.0	74.2	71.7
5 \ Fact	63.4*	67.1	63.4*	**64.0**	62.7*	71.8	75.6	71.7	73.2	70.3
6 \ Sentiment	63.1*	67.2	63.1*	62.7	63.5*	70.9	74.0	70.9	71.6	70.2
7 \ Causal	62.4*	66.3	62.1*	58.6	**65.5***	**74.5**	**78.7**	**74.5**	**75.4**	**73.6**
8 \ Normative	61.0	64.7	61.0	60.3	61.6*	68.2	72.4	68.2	68.3	68.1

(b) Debatepedia

Table 5: PSL accuracy. $p < \{0.05^{\star}, 0.01^{\dagger}, 0.001^{\ddagger}\}$ with paired bootstrap compared to the best baseline.

This indicates that our operationalization of *argument from consequences* and *practical reasoning* can effectively explain a prevailing mechanism of argumentative relations.

Quite surprisingly, normative relation is highly informative for non-normative arguments as well for both datasets. To understand how this mechanism works for non-normative arguments, we analyzed arguments for which it predicted the correct relations with high probabilities. It turns out that even for non-normative claims, the module often interprets negative sentiment toward a target as an opposition to the target. For the following example,

> **Claim:** *Schooling halts individual development.*
>
> **Attack Statement:** *Schooling, if done right, can lead to the development of personal rigor ...*

the module implicitly judges the ‘‘schooling’’ in the claim to be opposed and thus judges the ‘‘schooling’’ in the statement (the source of consequence) to be contrary to the claim’s stance while having positive sentiment (i.e., R11 applies). This behavior is reasonable, considering how advocacy and opposition are naturally mapped to positive and negative norms in our annotation schema (§5.3).

The utility of normative relation for non-normative arguments is pronounced for Debatepedia. Excluding this mechanism leads to a significant drop of F1-scores by 4.8 points (Table 5b row 8). One possible reason is that most claims in the non-normative set of Debatepedia are valuation; that is, they focus on whether something is good or bad, or preferences between options. As discussed above, valuation can be handled by this mechanism naturally. And in such arguments, causal relation may provide only little and noisy signal (row 7).

Sentiment coherence is the second most contributing mechanism. For Kialo, including it in the presence of normative relation is rather disruptive (Table 5a row 6). This may be because the two mechanisms capture similar (rather than complementary) information, but sentiment coherence provides inaccurate information conflicting with that captured by normative relation. Without normative relation, however, sentiment coherence

contributes substantially more than factual consistency and causal relation by 4.4–5.9 F1-score points (not in the table). For Debatepedia, the contribution of sentiment coherence is clear even in the presence of normative relation (Table 5b row 6).

Factual consistency and causal relation have high precision and low recall for the support and attack relations. This explains why their contribution is rather small overall and even obscure for Kialo in the presence of normative relation (Table 5a rows 5 and 7). However, without normative relation they contribute 0.7–1.1 F1-score points for Kialo (not in the table). For Debatepedia, factual consistency contributes 0.5–1.3 points (Table 5b row 5), and causal relation 1.8 points to normative arguments (row 7). Their contributions show different patterns in a supervised setting, however, as discussed in the next section.

To apply the chain rules (R14–R17) for Kialo, we built 16,328 and 58,851 indirect arguments for the normative and non-normative sets, respectively. Applying them further improves the best performing PSL model (Table 5a row 12). It suggests that there is a relational structure among arguments, and structured prediction can reduce noise in independent predictions for individual arguments.

There is a notable difference in the performance of models between the three-class setting (Kialo) and the binary setting (Debate). The binary setting makes the problem easier for the baselines, reducing the performance gap with the logical mechanisms. When three relations are considered, the sentiment baseline and the textual entailment baseline suffer from low recall for the neutral and support/attack relations, respectively. But if an argument is guaranteed to belong to either support or attack, these weaknesses seem to disappear.

7.4 Error Analysis

We conduct an error analysis on Kialo. For the mechanism of normative relation, we examine misclassifications in normative arguments by focusing on the 50 support arguments and 50 attack arguments with the highest probabilities of the opposite relation. Errors are grouped into four types: R-C consistency/contrary (60%), consequence sentiment (16%), ground-truth relation (8%), and else (16%). The first type is mainly due to the model failing to capture antonymy relations, such as *collective presidency* ↔ *unitary presidency* and *marketplace of ideas* ↔ *deliver the best ideas*. Integrating advanced knowledge may rectify this issue. The second type of error often arises when a statement has both positive and negative words, as in ‘‘student unions could *prevent* professors from *intentionally failing students* due to personal factors’’.

For the other mechanisms, we examine non-normative arguments that each mechanism judged to have strong signal for a false relation. To that end, for each predicate in R1–R9, we choose the top 20 arguments that have the highest probabilities but were misclassified. Many errors were simply due to the misclassification of the classification modules, which may be rectified by improving the modules’ accuracy. But we also found some blind spots of each predicate. For instance, FactEntail often fails to handle concession and scoping.

> **Claim:** *Fourth wave feminists espouse belief in equality.*
>
> **Attack Statement:** *It is belief in equality of outcome not opportunity that fourth wave feminists are espousing with quotas and beneficial bias.*

For SentiConsist, a statement can have the same ground of value as the claim without supporting it:

> **Claim:** *The education of women is an important objective to improve the overall quality of living.*
>
> **Attack Statement:** *Education of both men and women will have greater effects than that of women alone. Both must play a role in improving the quality of life of all of society’s members.*

The statement attacks the claim while expressing the same sentiment toward the same target (underlined).

8 Experiment 2. Representation Learning

Supervised models are good at capturing various associations between argumentative relations and data statistics. Here, we examine if our logical mechanisms can further inform them. We describe a simple but effective representation learning method, followed by baselines and experiment results.

8.1 Method

Our logical mechanisms are based on textual entailment, sentiment classification, causality classification, and four classification tasks for normative relation (§4). We call them **logic tasks**. We combine all minibatches across the logic tasks using the same datasets from §4 except the heuristically made negative datasets. Given uncased BERT-base, we add a single classification layer for each logic task and train the model on the minibatches for five epochs in random order. After that, we fine-tune it on our fitting data (Table 4), where the input is the concatenation of statement and claim. Training stops if AUC does not increase for 5 epochs on the validation data. We call our model **LogBERT**.

8.2 Baselines

The first goal of this experiment is to see if the logical mechanisms improve the predictive power of a model trained without concerning them. Thus, our first baseline is **BERT** fine-tuned on the main task only. This method recently yielded the (near) best accuracy in argumentative relation classification (Durmus et al., 2019; Reimers et al., 2019).

In order to see the effectiveness of the representation learning method, the next two baselines incorporate logical mechanisms in different ways. **BERT+LX** uses latent cross (Beutel et al., 2018) to directly incorporate predicate values in R1–R13 as features; we use an MLP to encode the predicate values, exploring (i) one hidden layer with D=768 and (ii) no hidden layers. BERT+LX consistently outperforms a simple MLP without latent cross. **BERT+MT** uses multitask learning to train the main and logic tasks simultaneously.

Lastly, we test two recent models from stance detection and dis/agreement classification. **TGA Net** (Allaway and McKeown, 2020) takes a statement-topic pair and predicts the statement's stance. It encodes the input using BERT and weighs topic tokens based on similarity to other topics. In our task, claims serve as ''topics''. We use the published implementation, exploring {50, 100, 150, 200} for the number of clusters and increasing the max input size to the BERT input size. **Hybrid Net** (Chen et al., 2018) takes a quote-response pair and predicts whether the response agrees or disagrees with the quote. It encodes the input using BiLSTM and uses self- and cross-attention between tokens. In our task, claims and statements serve as ''quotes'' and ''responses'', respectively.

8.3 Results

Tables 6a (Kialo) and 6b (Debatepedia) summarize the accuracy of each model averaged over 5 runs with random initialization. For non-normative arguments, the causality task is excluded from all models as it consistently hurts them for both datasets.

Regarding the baselines, TGA Net (row 1) and Hybrid Net (row 2) underperform BERT (row 3). TGA Net, in the original paper, handles topics that are usually short noun phrases. It weighs input topic tokens based on other similar topics, but this method seems not as effective when topics are replaced with longer and more natural claims. Hybrid Net encodes input text using BiLSTM, whose performance is generally inferior to BERT.

BERT trained only on the main task is competitive (row 3). BERT+LX (row 4), which incorporates predicate values directly as features, is comparable to or slightly underperforms BERT in most cases. We speculate that predicate values are not always accurate, so using their values directly can be noisy. LogBERT (row 6) consistently outperforms all models except for non-normative arguments in Debatepedia (but it still outperforms BERT). While both BERT+MT and LogBERT are trained on the same logic tasks, BERT+MT (row 5) performs consistently worse than LogBERT. The reason is likely that logic tasks have much larger training data than the main task, so the model is not optimized enough for the main task. On the other hand, LogBERT is optimized solely for the main task after learning useful representations from the logic tasks, which seem to lay a good foundation for the main task.

We examined the contribution of each logic task using ablation tests (not shown in the tables). Textual entailment has the strongest contribution across settings, followed by sentiment classification. This contrasts the relatively small contribution of factual consistency in Experiment 1. Moreover, the tasks of normative relation have the smallest contribution for normative arguments and the causality task for non-normative arguments in both datasets. Three of the normative relation tasks take only a statement as input, which is inconsistent with the main task. This inconsistency might cause these tasks

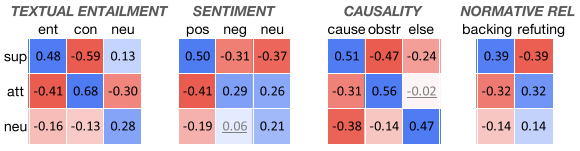
		Normative Arguments						Non-normative Arguments					
		ACC	AUC	F1	$F1_{sup}$	$F1_{att}$	$F1_{neu}$	ACC	AUC	F1	$F1_{sup}$	$F1_{att}$	$F1_{neu}$
1	TGA Net	71.5	88.3	62.2	43.5	54.3	88.7	76.6	90.8	69.8	62.9	53.9	92.5
2	Hybrid Net	66.8	78.2	56.2	42.9	42.4	83.4	71.8	82.2	65.7	55.6	51.4	90.2
3	BERT	79.5	92.4	73.3	60.5	65.2	**94.2**	83.8	94.6	79.2	72.3	68.8	96.6
4	BERT+LX	79.2	92.1	72.7	58.7	65.6*	93.8	83.7	94.6	79.2	70.8	69.9‡	**96.9**‡
5	BERT+MT	79.3	92.6*	73.4	**63.8**‡	63.6	92.7	83.6	94.7	79.2	71.8	69.7‡	96.1
6	LogBERT	**80.0**‡	**92.8**‡	**74.3**‡	63.6‡	**66.2**‡	93.2	**84.3**‡	**95.0**‡	**80.2**‡	**73.1**‡	**71.4**‡	96.1

(a) Kialo

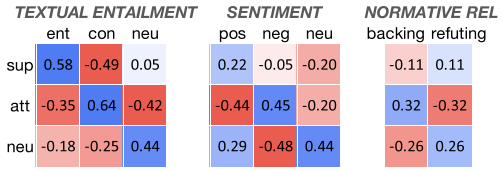
		Normative Arguments					Non-normative Arguments				
		ACC	AUC	F1	$F1_{sup}$	$F1_{att}$	ACC	AUC	F1	$F1_{sup}$	$F1_{att}$
1	TGA Net	66.1	75.0	65.4	69.8	60.9	66.5	74.3	65.9	70.1	61.7
2	Hybrid Net	67.2	70.1	67.2	68.1	66.3	59.7	62.6	58.8	64.5	53.2
3	BERT	79.1	88.3	79.4	79.8	79.0	80.7	87.6	80.7	81.4	79.9
4	BERT+LX	78.4	88.1	78.4	79.2	77.5	**81.6**	**88.8**	**81.5**	**82.3**	**80.8**
5	BERT+MT	79.6	88.2	79.6	80.0	79.1	77.6	86.3	77.5	78.9	76.0
6	LogBERT	**81.0***	**88.8**	**80.7***	**81.1***	**80.4***	81.2	88.3	80.8	81.7	80.0

(b) Debatepedia

Table 6: Accuracy of supervised models. $p < \{0.05^{\star}, 0.001^{\ddagger}\}$ with paired bootstrap compared to BERT.



(a) Normative arguments.



(b) Non-normative arguments.

Figure 2: Pearson correlation coefficients between argumentative relations and logic tasks from LogBERT. All but underlined values have $p < 0.0001$.

to have only small contributions in representation learning. The small contribution of the causality task in both Experiments 1 and 2 suggests large room for improvement in how to effectively operationalize causal relation in argumentation.

To understand how LogBERT makes a connection between the logical relations and argumentative relations, we analyze ‘‘difficult’’ arguments in Kialo that BERT misclassified but LogBERT classified correctly. If the correct decisions by LogBERT were truly informed by its logic-awareness, the decisions may have correlations with its (internal) decisions for the logic tasks as well, for example, between attack and textual contradiction. Figure 2 shows the correlation coefficients between the probabilities of argumentative relations and those of the individual classes of the logic tasks, computed simultaneously by LogBERT (using the pretrained classification layers for the logic tasks). For sentiment, the second text of an input pair is the sentiment target, so we can interpret each class roughly as the statement’s sentiment toward the claim. For normative relation, we computed the probabilities of backing (R10+R12) and refuting (R11+R13).

The correlations are intuitive. The support relation is positively correlated with textual entailment, positive sentiment, ‘cause’ of causality, and ‘backing’ of normative relation, whereas the attack relation is positively correlated with textual contradiction, negative sentiment, ‘obstruct’ of causality, and ‘refuting’ of normative relation. The neutral relation is positively correlated with the neutral classes of the logic tasks. The

only exception is the normative relation for non-normative arguments. A possible reason is that most claims in non-normative arguments do not follow the typical form of normative claims, and that might affect how the tasks of normative relation contribute for these arguments.

LogBERT's predictive power comes from its representation of arguments that makes strong correlations between the logical relations and argumentative relations. Though LogBERT uses these correlations, it does not necessarily *derive* argumentative relations *from* the logic rules. It is still a black-box model with some insightful explainability.

9 Conclusion

We examined four types of logical and theory-informed mechanisms in argumentative relations: factual consistency, sentiment coherence, causal relation, and normative relation. To operationalize normative relation, we built rich annotation schema and data for the argumentation schemes *argument from consequences* and *practical reasoning*, too.

Evaluation on arguments from Kialo and Debatepedia revealed the importance of these mechanisms in argumentation, especially normative relation and sentiment coherence. Their utility was further verified in a supervised setting via our representation learning method. Our model learns argument representations that make strong correlations between logical relations and argumentative relations in intuitive ways. Textual entailment was found to be particularly helpful in the supervised setting.

Some promising future directions are to probe fine-tuned BERT to see if it naturally learns logical mechanisms and to improve PSL with more rules.

Acknowledgments

We thank the reviewers and action editor for valuable comments. This research was supported in part by the Kwanjeong Educational Foundation, CMU's GuSH Research grant, Volkswagen Stiftung under grant 92 182, ESRC under ES/V003901/1, and EPSRC under EP/N014871/1.

References

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. 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 2922–2932. https://doi.org/10.18653/v1/D19-1290

Khalid Al-Khatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. 2020. End-to-end argumentation knowledge graph construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7367–7374. https://doi.org/10.1609/aaai.v34i05.6231

Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8913–8931. https://doi.org/10.18653/v1/2020.emnlp-main.717

Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2017. Hinge-Loss Markov random fields and probabilistic soft logic. *Journal of Machine Learning Research*, 18(109):1–67.

Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O Brien, Lamia Tounsi, and Mark Hughes. 2013. Sentiment analysis of political tweets: Towards an accurate classifier. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 49–58.

Alex Beutel, Paul Covington, Sagar Jain, Can Xu, Jia Li, Vince Gatto, and Ed H. Chi. 2018. Latent cross: Making use of context in recurrent recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 46–54. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3159652.3159727

Elena Cabrio, Sara Tonelli, and Serena Villata. 2013. From discourse analysis to argumentation schemes and back: relations and differences.

In *Computational Logic in Multi-Agent Systems*, pages 1–17. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-40624-9_1

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. 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 2926–2936. https://doi.org/10.18653/v1/D19-1291

Di Chen, Jiachen Du, Lidong Bing, and Ruifeng Xu. 2018. Hybrid neural attention for agreement/disagreement inference in online debates. In *Proceedings of the 2018 Conference of Empirical Methods in Natural Language Processing*, pages 665–670. https://doi.org/10.18653/v1/D18-1069

HongSeok Choi and Hyunju Lee. 2018. GIST at SemEval-2018 Task 12: A network transferring inference knowledge to Argument Reasoning Comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 773–777. https://doi.org/10.18653/v1/S18-1122

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54. https://doi.org/10.3115/v1/P14-2009

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE Corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104. https://doi.org/10.18653/v1/W17-0812

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. Determining relative argument specificity and stance for complex argumentative structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4630–4641. https://doi.org/10.18653/v1/P19-1456

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. https://doi.org/10.18653/v1/P17-1002

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

Debela Gemechu and Chris Reed. 2019. Decompositional argument mining: A general purpose approach for argument graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 516–526. https://doi.org/10.18653/v1/P19-1049

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report*.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1): 125–179. https://doi.org/10.1162/COLI_a_00276

Yufang Hou and Charles Jochim. 2017. Argument relation classification using a joint inference model. In *Proceedings of the 4th Workshop on Argument Mining*, pages 60–66. https://doi.org/10.18653/v1/W17-5107

Yohan Jo, Elijah Mayfield, Chris Reed, and Eduard Hovy. 2020. Machine-aided annotation for fine-grained proposition types in argumentation. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*, pages 1008–1018.

Jonathan Kobbe, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Unsupervised stance detection for arguments from consequences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 50–60, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.4

John Lawrence and Chris Reed. 2016. Argument mining using argumentation scheme structures. In *Proceedings of the Sixth International Conference on Computational Models of Argument*, pages 379–390.

John Lawrence and Chris Reed. 2017. Mining argumentative structure from natural language text using automatically generated premise-conclusion topic models. In *Proceedings of the 4th Workshop on Argument Mining*, pages 39–48. https://doi.org/10.18653/v1/W17-5105

John Lawrence, Jacky Visser, and Chris Reed. 2019. An online annotation assistant for argument schemes. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 100–107. https://doi.org/10.18653/v1/W19-4012

Anna Lindahl, Lars Borin, and Jacobo Rouces. 2019. Towards assessing argumentation annotation—a first step. In *Proceedings of the 6th Workshop on Argument Mining*, pages 177–186. https://doi.org/10.18653/v1/W19-4520

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Distinguishing antonyms and synonyms in a pattern-based neural network. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 76–85, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-1008

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664. https://doi.org/10.18653/v1/P19-1459

Juri Opitz and Anette Frank. 2019. Dissecting content and context in argumentative relation analysis. In *Proceedings of the 6th Workshop on Argument Mining*, pages 25–34. https://doi.org/10.18653/v1/W19-4503

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)*.

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. https://doi.org/10.18653/v1/N16-1164

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008.

Language resources for studying argument. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco. European Language Resources Association (ELRA),

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578. https://doi.org/10.18653/v1/P19-1054

Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. 2018. Feasible annotation scheme for capturing policy argument reasoning using argument templates. In *Proceedings of the 5th Workshop on Argument Mining*, pages 79–89, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5210

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence—an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1050

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518. https://doi.org/10.18653/v1/S17-2088

Sara Rosenthal and Kathy McKeown. 2015. I Couldn't Agree More: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 168–177. Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-4625

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659. https://doi.org/10.1162/COLI_a_00295

Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674. https://doi.org/10.18653/v1/D18-1402

Niket Tandon, Bhavana Dalvi Mishra, Keisuke Sakaguchi, Antoine Bosselut, and Peter Clark. 2019. WIQA: A dataset for ''What if...'' reasoning over procedural text. 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 6076–6085. https://doi.org/10.18653/v1/D19-1629

Jacky Visser, John Lawrence, Chris Reed, Jean Wagemans, and Douglas Walton. 2020. Annotating argument schemes. *Argumentation*, pages 1–39. https://doi.org/10.1007/s10503-020-09519-x, PubMed: 33678987

Douglas Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*, Cambridge University Press. https://doi.org/10.1017/CBO9780511802034

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2006. The penn discourse treebank 3.0 annotation manual.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding

through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. https://doi.org/10.18653/v1/N18-1101

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics. https://doi.org/10.3115/1220575.1220619

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–783. Melbourne, Australia. Association for Computational Linguistics.

Chang Xu, Cecile Paris, Surya Nepal, and Ross Sparks. 2019. Recognising agreement and disagreement between stances with reason comparing networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4665–4671.