Uncovering Implicit Inferences for Improved Relational Argument Mining

Ameer Saadat-Yazdi
School of Informatics
University of Edinburgh
ameer.saadat@ed.ac.uk

Jeff Z. Pan
School of Informatics
University of Edinburgh
j.z.pan@ed.ac.uk

Nadin Kökciyan
School of Informatics
University of Edinburgh
nadin.kokciyan@ed.ac.uk

Abstract

Argument mining seeks to extract arguments and their structure from unstructured texts. Identifying relations (such as attack, support, and neutral) between argumentative units is a challenging task because two units may be related to each other via implicit inferences. These inferences often rely on external commonsense knowledge to discover how one argumentative unit relates to another. State-of-the-art methods, however, rely on predefined knowledge graphs, and thus might not cover target pairs of argumentative units well. We introduce a new generative approach to finding inference chains that connect these pairs by making use of the Commonsense Transformer (COMET). We evaluate our approach on three datasets for both the two-label (attack/support) and three-label (attack/support/neutral) tasks. Our approach significantly outperforms the state-of-the-art, by 2-5% in F1 score, on two out of the three datasets with minor improvements on the remaining one.

1 Introduction

In argumentation, such as in a debate, it is important to understand the key arguments put forward and how the arguments relate to each other based on a specific context. Ideally, we would like to empower systems to extract arguments and the relations between arguments automatically. Such systems could be used to aggregate opinions (Cocarascu et al., 2019), participate in debates (Slonim et al., 2021), or they could assist humans in making informed decisions by considering different points of view.

Toulmin (1958) defines an argument as a claim backed by one or more grounds (or premises) based on a warrant (i.e., the inference link between the grounds and the claim). For example, consider the following two sentences: *Drunk driving hurts innocent people. Therefore, drunk driving is wrong.* Based on Toulmin's model, we have the claim *drunk driving is wrong*, supported on the grounds that *drunk driving hurts innocent people*. In what follows, we will call the claim and grounds *argumentative units* (AUs). To justify the claim, the author relies on the reader's ability to uncover the implicit warrant that *hurting innocent people is wrong* which connects the grounds to the claim. While Toulmin's warrants model only arguments in which the units are connected by support relations, we extend the definition to also allow for attacks and neutral relations. We do this to enable us to apply the notion of warrants to the relational argument mining task, as proposed by Carstens and Toni (2015), to better identify attack/support/neutral relations between pairs of AUs.

Finding the correct warrant is a challenging task that requires deep reasoning, which pre-trained language models struggle with (Helwe et al., 2021). However, by iteratively generating knowledge, we can chain together inferences to get a series of causes and effects that connect two AUs together similar to chain-of-thought reasoning (Wei et al., 2022). We propose the use of the Commonsense Transformer (COMET) (Hwang et al., 2021) which is equipped with social commonsense to generate warrants as a series of inferences. The knowledge contained in COMET is qualitatively different from other knowledge sources in that it contains normative commonsense. This depends upon the culture and beliefs of various groups compared with static commonsense knowledge graphs (KGs), such as ConceptNet (Speer et al., 2017), which focus on globally accepted knowledge. For this reason, we suggest that COMET is better able to identify the reasoning behind human arguments and that a generative approach based on COMET is better suited to finding warrants compared to a static knowledge graph as proposed by Paul et al. (2020).

Section 4 introduces our approach to test this hypothesis, where we articulate an algorithm,

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2484–2495

ARGCON (Section 4.2), to generate warrants between a pair of argumentative units and use these warrants as additional input to the classifier of argument relations to evaluate whether this improves the model's ability to distinguish between attack, support or neutral relations. We then compare the use of three external knowledge sources: ConceptNet$^+$ the commonsense knowledge extracted by Paul et al., the knowledge generated by ARGCON, to obtain novel commonsense inferences (Section 4.2), and ARGCON + LINK which enriches the generated knowledge with additional relations via a link prediction model (Section 4.3).

In Section 5, we compare our work to (Paul et al., 2020), which is the only other work we are aware of to address the task of relational argument mining by uncovering commonsense relations. We show that our method for generating warrants with COMET outperforms their method with a 1-2% increase in F1 scores on two datasets. We further evaluate our model on a three-class dataset, which includes an additional neutral relation, and witness a 5% improvement of the F1 score over a RoBERTa baseline. Hence, we set a new benchmark for the three-class task in relational argument mining. We share our findings in Section 6 and conclude with our future directions in Section 7.

2 Background

In this paper, we investigate the effectiveness of COMET (Bosselut et al., 2019) for generating implicit warrants that link arguments together. COMET is a generative model which has been fine-tuned on a human-authored commonsense knowledge graph ATOMIC (Sap et al., 2019). By training on ATOMIC, COMET extends the knowledge captured by the knowledge graph to new domains and is able to synthesize new commonsense knowledge based on knowledge captured during pre-training. The model is able to generate knowledge along multiple relations- a few examples are shown in Table 1. COMET takes as input a sentence and a relation along which to generate inferences. For example, given an input *PersonX's house was foreclosed* and relation *CausedBy*, COMET would generate *PersonX failed to pay their mortgage*.

Some successful applications of COMET include the generation of potential outcomes of events for abductive reasoning (Bhagavatula et al., 2020), question answering (Branco et al., 2021) and text generation (Guan et al., 2020). Closely related to our work, Chakrabarty et al. (2021) shows that COMET can be used to generate logically sound premises for incomplete arguments. While previous works have considered using COMET for single-hop inference, we found that for our task of relational argument mining, often multiple inference steps are needed to get from the first argumentative unit to the second; we address this challenge in Section 4.2.

3 Related Work

Various approaches have been proposed to infer argumentative structure from unstructured text. These include Long-Short Term Memory (LSTM) models (Cocarascu and Toni, 2017; Paul et al., 2020), pre-trained transformers (Ruiz-Dolz et al., 2021) and logic programming (Jo et al., 2021). Pre-trained transformers, in particular, have been shown to perform exceptionally well on this task with no additional feature engineering (Ruiz-Dolz et al., 2021; Fromm et al., 2019); this suggests that the introduction of external knowledge encoded within the transformers due to pre-training on large corpora is necessary to make significant progress on this task. Of these works, only Fromm et al. (2019) makes use of external knowledge to identify the stance of arguments towards a given topic. This differs from our task in that we wish to identify the relationship between pairs of AUs and so have the added challenge of finding paths that link the AUs together.

Commonsense knowledge has proven repeatedly to aid in tasks such as natural language inference (Wang et al., 2019) and question answering (Lv et al., 2020). The majority of these works focus mainly on deriving knowledge from ConceptNet (Speer et al., 2017); however, a key limitation of knowledge graphs (Pan et al., 2016) is their inherent incompleteness.

Recent trends in argument mining have shown a move towards incorporating external knowledge for various argument-mining tasks. Two works explore the use of automatically constructed knowledge graphs extracted from Wikidata and Google search, which, however, fail to achieve better results than BERT with no additional knowledge (Fromm et al., 2019; Li et al., 2021). On the other hand, a work using LSTM-based models for knowledge-enhanced argument mining with ConceptNet$^+$, shows that the introduction of knowledge indeed improves performance upon their baselines (Paul et al., 2020).

One of the key limitations of existing works is their use of static knowledge graphs which are unable to handle arbitrary pairs of AUs due to their limited coverage. To address this issue, we use COMET to generate conmmonsense inferences on-the-fly to connect AUs and therefore generalize to new domains.

4 Our Approach

In this section, we describe the design of our model as well as our procedures for obtaining knowledge on the fly with COMET and enriching the knowledge with additional links.

4.1 KE-RoBERTa

Given a pair of Argumentative Units, AUs, we aim to predict the relation between these AUs by the use of external knowledge relevant to this pair. Ruiz-Dolz et al. (2021) show that RoBERTa is a strong baseline for the classification of argument relations when compared to other transformer-based architectures. Hence, we introduce KE-RoBERTa, which is a knowledge-enhanced RoBERTa-based model as depicted in Figure 1.

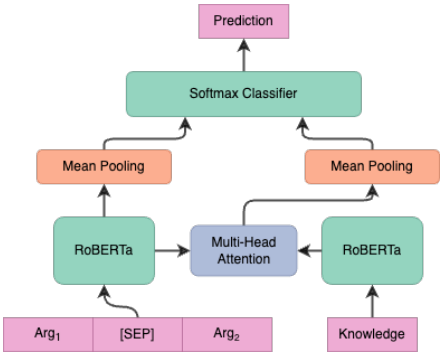

Figure 1: KE-RoBERTa, our proposed Siamese model for Commonsense Knowledge injection with Arg_1 and Arg_2 being the input AUs

In order to inject knowledge into RoBERTa, we use a Siamese model to separately encode the two AUs as well as the commonsense knowledge much like the S-BERT model (Reimers and Gurevych, 2019); in other words, the parameters of the two RoBERTa blocks are shared. The model then combines the two representations by concatenating them together. We also introduce a multi-head attention mechanism (Vaswani et al., 2017) to attend to relevant information and block out noise. We use mean pooling to combine individual token representations before concatenating the result. This allows us to use a fix-sized feed-forward network to compute predictions from the combined AU and knowledge representations.

The notion of knowledge is loosely defined in the literature but can be viewed as a set of beliefs about the world that approximate truth. In our case this knowledge is represented as the set of all knowledge graph paths that could connect two AUs. While there are many ways to represent knowledge graph paths, in this work, we convert each knowledge graph relation into its natural language equivalent and concatenate the nodes and relations together to form a sentence in natural language. For each pair of AUs we then concatenate all possible paths together in order to input the knowledge into a transformer. The source of knowledge graph paths we refer to in the following sections as ConceptNet$^+$ combines knowledge from ConceptNet, and WordNet, with additional links predicted by neural networks; ConceptNet$^+$ is proposed by Paul et al. and the extracted knowledge was provided to us by the authors. Since the authors did not experiment on the three class task we did not have ConceptNet$^+$ relations for M-Arg.

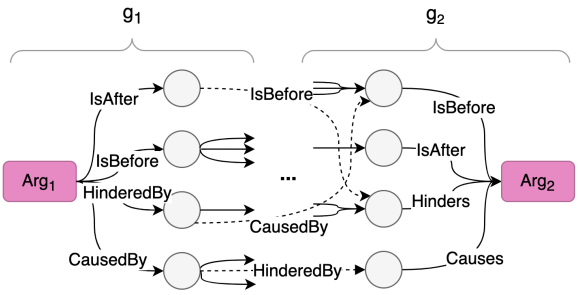

Figure 2: Knowledge graph generation by ARGCON with link inference enabled. The figure shows how two graphs, g_1 and g_2, are generated from each AU, Arg_1 & Arg_2, and then merged. Dashed lines show additional relations found by LINK.

4.2 ARGCON: A new way to generate knowledge with COMET

All previous works using COMET seek to generate a possible outcome, a_2, from a known event, a_1, and a relation. For this task, we seek to invert this and find a path to connect a_1 and a_2. What makes this more challenging is that AUs are often connected by multi-hop inferences and so multiple intermediary outcomes o_i may be needed to connect a_1 and a_2. The types of inference we wish to generate can be represented as:

$$a_1 \xrightarrow{r_1} o_1 \xrightarrow{r_2} o_2 \cdots o_{n-1} \xrightarrow{r_n} a_2$$

where r_i is a relation connecting events/outcomes, and neither the o_i nor r_i are known in advance. To do this, we propose a new technique using an improved version of COMET, trained on ATOMIC$^{20}_{20}$, a larger and more extensive version of the original ATOMIC knowledge graph (Hwang et al., 2021), to generate deep commonsense inference chains that connect two AUs together.

Relation	Inverse
Causes	*CausedBy*
HinderedBy	*Hinders*
isAfter	*isBefore*
isBefore	*isAfter*

Table 1: ATOMIC$^{20}_{20}$ relations (R) used to generate knowledge from COMET with corresponding inversions

ATOMIC$^{20}_{20}$ consists of several classes of relations, including 'causes for a person to perform an action', 'effects of an action on others', and 'event-centred'. In this paper, we focus on 'event-centred' relations to generate knowledge from, which is a subset of the relations contained in ATOMIC$^{20}_{20}$. In Table 1, we show the four relations selected. In Section 5, we will report our results on three existing datasets that include AUs focusing on event-centred knowledge.

Algorithm 1 provides pseudocode for our knowledge generation method *Argument Connector* (ARGCON). Our algorithm takes three inputs, a pair of AUs, (a_1 and a_2) as well as a boolean flag ($wlink$) to indicate whether additional link prediction should be performed as described in Section 4.3. The algorithm then returns the paths that connect a_1 to a_2.

Starting from the pair of AUs, we first initialise two single-node graphs, g_1 and g_2, from a_1 and a_2 (lines 1-2). For each of these graphs, we take the set of their nodes N, and for each node n and each relation r we generate new nodes n' using COMET (lines 8-13). We then update N to contain the set of nodes at the current depth i (line 14) and repeat this process d times until each graph has a tree with a maximum depth of d (line 4). Note that graphs g_1 and g_2 now contain only paths from a_1 and a_2, meaning that a_2 cannot be reached from any other node. In order to get paths from a_1 to a_2, we need to flip the relations in g_2 replace them with their corresponding inverse, as shown in Table 1, and then merge the two graphs wherever they share a common node (line 15). We now have a single merged graph g_m containing a_1 and a_2. If $wlink$ is set to `true`, the LINK model will be used to infer additional relations that connect g_1 to g_2 (line 16-18) which are then appended to g_m. The algorithm returns all paths that connect a_1 and a_2 (line 19).

Algorithm 1: ARGCON$(a_1, a_2, wlink)$

Data: R, the set of relations to consider
Data: d, the depth of a tree
Data: COMET, commonsense transformer
Data: LINK, our LINK model

1 $g_1 \leftarrow \mathsf{makeGraph}(a_1)$
2 $g_2 \leftarrow \mathsf{makeGraph}(a_2)$
3 **forall** $g \in \{g_1, g_2\}$ **do**
4 **for** $i = 0$ **to** d **do**
5 **if** $i = 0$ **then**
6 $N \leftarrow g.nodes()$
7 $N' \leftarrow []$
8 **forall** $n \in N$ **do**
9 **forall** $r \in R$ **do**
10 $n' \leftarrow \mathsf{genNode}(n, r, \mathsf{COMET})$
11 **if** $n' \neq NULL$ *and* $n' \neq n$ **then**
12 $g \leftarrow g.addtriple(n, r, n')$
13 $N' \leftarrow N'.append(n)$
14 $N \leftarrow N'$
15 $g_m \leftarrow \mathsf{mergeGraphs}(g_1, \mathsf{flip}(g_2))$
16 **if** $wlink$ **then**
17 $new\_edges \rightarrow$ LINK(g_1, g_2)
18 $g_m.append(new\_edges)$
19 **return** $\mathsf{computePaths}(a_1, a_2, g_m)$

4.3 LINK Prediction Between Nodes

It may not be always possible to find a path between the AUs by using only the merging operation described above. We introduce an additional inference step to identify the relationships between the nodes of the two knowledge graphs.

To achieve this, e trained a link prediction model (LINK) to identify the most likely relations that exist between two nodes, which are provided as inputs. LINK is a BERT model fine-tuned on the ATOMIC$^{20}_{20}$ training set to predict the relation between two nodes of the knowledge graph. In order to identify unrelated node pairs, we sample 4000 pairs that are not connected in ATOMIC$^{20}_{20}$ and assign them a 'None' label. Table 2 shows the training and validation performance of LINK when trained to predict the relations in Table 1 and the 'None' relation.

	Accuracy	Precision	Recall	F1
val	0.91	0.89	0.88	0.88
test	0.90	0.85	0.85	0.85

Table 2: Validation and testing performance of LINK model for COMET link prediction after 5 epochs.

Given a pair of graphs and their corresponding node sets, N_1, N_2, generated by COMET, we use LINK to predict the most likely relationships between the two graphs. Explicitly, we pass all node pairs $(a, b) \in N_1 \times N_2$ to LINK and add the predicted relation to g_{merged} (line 15 in Algorithm 1, $wlink$ is set to 1 to enable further inference) before computing paths. Figure 2 depicts our method when link inference is enabled.

5 Evaluation

In this section, we describe the datasets we have chosen, our choices of parameters for knowledge extraction, our training setup, and the results of our experiments on three datasets.

5.1 Datasets and Knowledge

We consider three datasets for our experiments: the Student Essay corpus (Opitz and Frank, 2019), Debatepedia (Paul et al., 2020), and the M-Arg Presidential Debate corpus (Mestre et al., 2021). We may refer to the datasets as Essay, Debate, and M-Arg for brevity.

The first two datasets were annotated with two labels (attack/support). Student Essay is a collection of argumentative essays written in English by second-language speakers, while Debatepedia contains arguments selected from a wiki of controversial topics and pro/con arguments[1] where controversial topics are discussed by multiple authors. These are the datasets used by Paul et al. (2020) which we use for comparison and we have taken the same dataset splits as used in their experiments. M-Arg consists of transcripts from five presidential debates in 2020; sentence pairs in this dataset were annotated with three labels (attack/support/neutral). Table 3 shows how the data has been split.

5.2 Experimental Setup

As a baseline, we trained a RoBERTa model[2], with no additional knowledge. We trained our models to minimize the cross-entropy loss on the labeled data with a batch size of 10 and a learning rate of $1e^{-5}$. The models were trained using the same hyperparameters for 10 epochs, and the model with the highest F1 validation score was selected for testing. Our code and the data can be found on GitLab[3].

We used ARGCON with a maximum depth limit of $d = 3$ to generate a list of paths from Arg_1 to Arg_2. Each path is given as lists of nodes and relations, which, by using a simple set of rules, we convert into natural language and concatenate the result. This is common practice when using COMET and the reader can refer to the code-base for more details on how this was done.

		A	S	N
Debate	train	3250	3236	-
	val	1088	1075	-
	test	1035	1127	-
Essay	train	273	2797	-
	val	130	1012	-
	test	91	1009	-
M-Arg	train	94	302	2887
	val	11	40	359
	test	15	42	342

Table 3: Distribution of labels (**A**ttack/**S**upport/**N**eutral) across datasets. While Debatepedia is balanced, the other two datasets show drastic imbalances making the task of detecting minority classes more challenging.

5.3 ARGCON vs ARGCON+LINK

We first investigate the proportion of AU pairs for which ARGCON and ARGCON+LINK could generate knowledge for (Table 4). We observe that nearly half of the pairs are not covered by ARGCON while ARGCON+LINK gives an improvement of 30% on Debatepedia and Student Essay and 40% on M-Arg.

5.4 Relational Argument Mining Results

Table 5 shows the results from the classification experiments; the first two rows, (Paul et al., 2020) and RoBERTa, are existing approaches. We also apply RoBERTa to the ConceptNet$^+$ extracted by Paul et al. to compare the performance against our COMET-based knowledge extraction methods. The results show that the common-

[1] The wiki has since migrated to https://idebate.org/debatabase.

[2] https://huggingface.co/roberta-base

[3] https://git.ecdf.ed.ac.uk/s1707343/commonsense-argmining/-/tree/master

		ARGCON	ARGCON+LINK
Debate	train	0.56	0.85
	val	0.53	0.83
	test	0.54	0.83
Essay	train	0.42	0.71
	val	0.37	0.68
	test	0.42	0.70
M-Arg	train	0.20	0.69
	val	0.23	0.74
	test	0.22	0.72

Table 4: Coverage of ARGCON vs. ARGCON+LINK.

sense enhanced KE-RoBERTa models improve upon RoBERTa in almost all cases. In addition, ARGCON and ARGCON+LINK give the best results on the Debate and Essay datasets respectively. For significance testing we used the Almost Stochastic Order test (Dror et al., 2019; Del Barrio et al., 2018) implemented in Python using the `deep-significance` library (Ulmer et al., 2022). We compared all models against the baseline based on the sample predictions on a single trial, with a confidence level of $\alpha = 0.05$ (before adjusting for all pair-wise comparisons using the Bonferroni correction). We consider the results significant when $\epsilon_{\text{min}} < \tau$ with $\tau = 0.3$.

We conducted further experiments on the M-Arg dataset to evaluate our knowledge extraction method on a three-label (attack, support, and neutral) task. The results in Table 6 confirm that our methods do indeed perform better than RoBERTa with no external knowledge.

6 Analysis and Discussion

In order to interpret the results better, we analyze the inference chains generated by ARGCON, and we also investigate the performance of the models across different topics. In Table 7, we provide some examples from the data to discuss our findings. For clarity, we have kept the table concise and omitted information not relevant to the points being made.

6.1 Performance Across Different Topics

The performance of the model varies across different topics. Figure 3 depicts the macro F1 scores of the KE-RoBERTa model for each topic in the M-Arg dataset. We observe significant improvements across some of the topics such as 'Taxes', 'Racism', 'Families' and 'Climate Change'. COMET has been exposed to these topics in the training set and so is better able to provide relevant knowledge.

One topic in which RoBERTa outperforms both knowledge-enhanced models is 'Why They Should Be Elected'. Our analysis of the class level precision (Appendix A) indicates that the drop in precision in Table 6 is actually due to this missidentification of support relations when we add external knowledge, we discuss this in Subsection 6.4.

6.2 Noisy Connections

The injected knowledge could be detrimental in some cases. The COMET generated knowledge can introduce noise by generating irrelevant relationships between two arguments. Consider Ex1 in Table 7, where ARGCON makes an incorrect prediction while the RoBERTa model correctly predicts a support relation. Here, the generic word 'good' is used to link Arg_1 and Arg_2. While useful, the analysis shows that the attetion mechanism in KE-RoBERTa does not solve the issue.

ARGCON *can sometimes infer relations when there are in fact none present.* The existence of a commonsense inference path between the arguments causes the model to favor the support relation over the neutral one. In Ex3 of Table 7, the first argument attacks the speaker's opponent while the second deflects the argument by addressing something unrelated. However, COMET connects the first sentence to the second on the basis that 'the country is in decline' supports 'being weaker'.

6.3 ARGCON vs. ConceptNet$^+$

The nature of the injected knowledge can change the outcome. Ex4 in Table 7 shows a comparison of knowledge extracted with ARGCON+LINK and ConceptNet$^+$. In the example the model provided with input from ARGCON+LINK predicts the correct relation while ConceptNet$^+$ fails. By examining the knowledge we observe that ARGCON+LINK is able to deduce that the first argument hinders the ability of parents to teach their children and therefore prevents students from learning skills. This allows the model to make the correct prediction. However, ConceptNet$^+$ can only deduce that a parent has a child that can learn.

As seen in this example, ConceptNet$^+$ provides some information about causality (learning causes intelligence) but this information is limited and often too simplistic to provide meaningful information for understanding the underpinning arguments.

Model	KG	Student Essay			Debatepedia		
		P	R	F1	P	R	F1
(Paul et al., 2020)	ConceptNet+	0.56	0.63	0.60	0.64	0.64	0.64
RoBERTa	-	0.64	0.68	0.66	0.75	0.75	0.75
KE-RoBERTa	ConceptNet+	0.66	0.70	0.68	0.74	0.74	0.74
	ARGCON	0.66	0.69	0.67	**0.76**	**0.76**	**0.76**
	ARGCON+LINK	**0.70**	**0.71**	**0.70***	0.75	0.75	0.75

Table 5: Macro-averaged F1, precision, recall scores of our experiments on the test sets for the 2-class task. Asterisk indiciates significant stochastic dominance over the baseline.

Model	KG	M-Arg		
		P	R	F1
RoBERTa	-	**0.57**	0.41	0.44
KE-RoBERTa	ARGCON	0.48	0.51	0.48*
	ARGCON+LINK	0.54	**0.48**	**0.49***

Table 6: Macro-averaged F1, precision, recall scores of our experiments on the test sets for the 3-class task. Asterisk indicates significant stochastic dominance over the baseline.

On the other hand, COMET is well equipped to handle more complex causal relations due to the nature of the data it was trained on. This supports our claim that the nature of COMET is better suited to providing useful commonsense for argument mining tasks compared to ConceptNet+.

6.4 Uncovering Hidden Biases in Arguments

In argument understanding, the ability to capture and make some of the implicit biases explicit can be helpful. Ex2 in Table 7 shows an argument made by the speaker to attack [PERSON X]'s candidacy. The argument made relies on the implicit reasoning that [PERSON X] did not fight [COUNTRY Y] because he is a communist. While there are many controversial assumptions being made when generating this knowledge, the COMET knowledge reflects the beliefs underlying the argument pair.

One of the key differences between fact-based and commonsense knowledge graphs is that the latter are more likely to capture cultural and normative biases. This is amplified by the fact that COMET is built on a pre-trained BART model, which captures biases present in the training corpus.

7 Conclusion

In this paper, we introduced a new method, ARGCON, to investigate the role of the commonsense knowledge encoded in COMET for the task of relational argument mining. In particular, ARGCON can reveal implicit commonsense reasoning chains (i.e., warrants) and these chains provide useful information for classifying argument relations such as attack, support, and neutral. Our experiments on three different datasets show that the chains generated by this method outperform existing approaches. Our analysis shows that the performance of our models varies given the topic under discussion and in some cases ARGCON creates noisy connections as a result of injected knowledge. We also observe that our model performs well in its ability to handle complex causal relationships and reason about social norms and biases.

In this work, we only consider pairs of AUs without taking into account any additional context. However, incorporating the text surrounding the pair of arguments is often necessary to be able to determine their relationship. In future work, we would like to incorporate such contextual information into our approach and explore the impact this has on the final performance. Another direction that we want to explore is considering other relations of COMET knowledge such as *intents*, *needs*, and *wants* to investigate whether this kind of social commonsense improves classification accuracy. Another avenue of investigation we would like to pursue is the role that external knowledge plays in increasing explainability and improving the diagnosis of end-to-end model predictions. Beyond argument mining, we believe that ARGCON is useful for other inference tasks such as recognizing textual entailment and question answering, we hope to experiment with other datasets to test this.

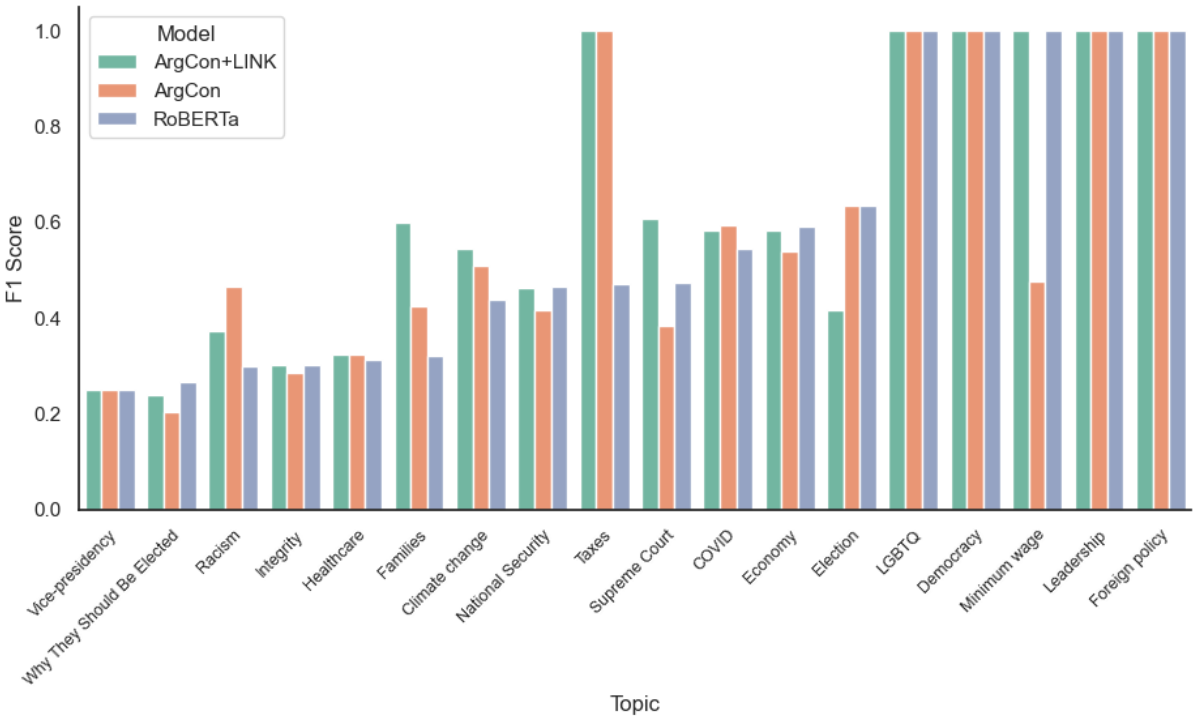

Figure 3: Comparison of F1 scores of our models on various topics in the test set of the M-Arg dataset.

Ex1	**Arg**$_1$	Congress will not ultimately cut Medicare to pay for reform.
	Arg$_2$	The 2010 US health care reform legislation is a good idea.
	ArgCon+Link	[Arg$_1$], isAfter, PersonY is not a member of the government, isAfter, PersonY is a lawyer, Causes, good, CausedBy, [Arg$_2$]
	Label: Support	**RoBERTa Prediction**: Support, **ArgCon+Link Prediction**: Attack
Ex2	**Arg**$_1$	[PERSON X] never fought it.
	Arg$_2$	[PERSON X] has been a cheerleader for communist [COUNTRY Y] over the last several decades.
	ArgCon	[Arg$_2$], HinderedBy, [PERSON X] is a good person., HinderedBy, [PERSON X] is a communist., Hinders, [Arg$_2$]
	Label: Support	**ArgCon Prediction**: Support
Ex3	**Arg**$_1$	Under this president, we become weaker, sicker, poor, more divided and more violent.
	Arg$_2$	With regard to being weaker, the fact is that I've gone head to head with [PERSON X] and made it clear to him we're not going to take any of his stuff.
	ArgCon	[Arg$_2$], HinderedBy, under this president, the country is in decline, HinderedBy, with regard to being weaker, Hinders, [Arg$_2$]
	Label: Neutral	**ArgCon Prediction**: Support
Ex4	**Arg**$_1$	Parents are usually too busy with their daily jobs to teach their children the life skills
	Arg$_2$	Students can learn practical skills after school hours from their parents
	ArgCon+Link	[Arg$_2$], Causes, parents are lazy, HinderedBy, parents are teaching their children, CausedBy, [Arg$_2$]
	ConceptNet$^+$	parent Antonym child used for learning, learning Causes intelligence related to mind HasA parent, parent Antonym child CapableOf learn
	Label: Attack	**ArgCon+Link Prediction**: Attack, **ConceptNet$^+$ Prediction**: Support

Table 7: Examples of knowledge and model predictions from the data: Ex1 is taken from the Debate dataset, Ex2 and Ex3 are taken from M-Arg, and Ex4 is taken from Student Essay. One interesting finding is that due to inductive biases in ArgCon *HinderedBy* and *Hinders* do not always reflect negative relations between nodes. The tags [PERSON X] and [COUNTRY Y] have been used to remove mentions of named individuals and countries in this table to avoid causing political offense.

Acknowledgements

The work was supported by the Edinburgh-Huawei Joint Lab and Huawei's grant CIENG4721/LSC. We would like to thank Paul et al. for providing their datasets and code-base allowing us to compare our work and improve upon their results.

Limitations

One limitation comes from our reliance on manually-constructed knowledge graphs. Given that different cultures and communities have varying notions of what is and is not commonsense, what is left out between arguments will also depend on the culture. For example, a knowledge graph constructed by English speakers in the UK, may not be able to provide useful knowledge about arguments about politics in the US due to differences in cultural values.

Similarly, our approach does not apply to low-resource languages that do not have sufficiently large commonsense knowledge graphs to train a transformer model such as COMET. Machine Translation of existing knowledge graphs may also prove insufficient due to cultural differences.

Ethics Statement

The datasets we use in this paper are publicly available. These datasets are anonymized and do not include sensitive information. A source of ethical concern is our reliance on generative models to provide warrants for arguments. These models, being trained on large web-based corpora may potentially generate problematic or biased outputs.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Ruben Branco, António Branco, João António Rodrigues, and João Ricardo Silva. 2021. Shortcutted Commonsense: Data Spuriousness in Deep Learning of Commonsense Reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1521, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lucas Carstens and Francesca Toni. 2015. Towards relation based Argumentation Mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO. Association for Computational Linguistics.

Tuhin Chakrabarty, Aadit Trivedi, and Smaranda Muresan. 2021. Implicit Premise Generation with Discourse-aware Commonsense Knowledge Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6247–6252, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oana Cocarascu, Antonio Rago, and Francesca Toni. 2019. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 1261–1269, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

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

Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. 2018. An optimal transportation approach for assessing almost stochastic order. In *The Mathematics of the Uncertain*, pages 33–44. Springer.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2773–2785. Association for Computational Linguistics.

Michael Fromm, Evgeniy Faerman, and Thomas Seidl. 2019. TACAM: Topic And Context Aware Argument Mining. *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 99–106. ArXiv: 1906.00923.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Chadi Helwe, Chloé Clavel, and Fabian M. Suchanek. 2021. Reasoning with transformer-based models:

Deep learning, but shallow reasoning. In *3rd Conference on Automated Knowledge Base Construction*.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (Comet-) Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):6384–6392.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard Hovy. 2021. Classifying Argumentative Relations Using Logical Mechanisms and Argumentation Schemes. *Transactions of the Association for Computational Linguistics*, 9:721–739.

Weichen Li, Patrick Abels, Zahra Ahmadi, Sophie Burkhardt, Benjamin Schiller, Iryna Gurevych, and Stefan Kramer. 2021. Topic-guided knowledge graph construction for argument mining. In *2021 IEEE International Conference on Big Knowledge (ICBK)*, pages 315–322.

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-Based Reasoning over Heterogeneous External Knowledge for Commonsense Question Answering. *arXiv:1909.05311 [cs]*. ArXiv: 1909.05311.

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

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, Florence, Italy. Association for Computational Linguistics.

J.Z. Pan, G. Vetere, J.M. Gomez-Perez, and H. Wu. 2016. *Exploiting Linked Data and Knowledge Graphs for Large Organisations*. Springer.

Debjit Paul, Juri Opitz, Maria Becker, Jonathan Kobbe, Graeme Hirst, and Anette Frank. 2020. Argumentative Relation Classification with Background Knowledge. *Computational Models of Argument*, pages 319–330. Publisher: IOS Press.

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 (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ramon Ruiz-Dolz, Jose Alemany, Stella Heras, and Ana Garcia-Fornes. 2021. Transformer-Based Models for Automatic Identification of Argument Relations: A Cross-Domain Evaluation. *IEEE Intelligent Systems*, pages 1–1. Conference Name: IEEE Intelligent Systems.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

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*, pages 4444–4451.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2022. deep-significance-easy and meaningful statistical significance testing in the age of neural networks. *arXiv preprint arXiv:2204.06815*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael Witbrock. 2019. Improving Natural Language Inference Using External Knowledge in the Science Questions Domain. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7208–7215. Number: 01.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022.

Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

A Class level performance of our model

Table 8 shows the breakdown of our models' performance across the different classes. To begin with, the performance on Debatepedia is mixed, with a slight improvement in the F1 score of the support class due to ARGCON. However, there is no clear winner as all models perform similarly to the baseline. This suggests the need for further investigation to identify the cause of the performance bottleneck which is not mirrored by the other 2 datasets.

In the Student Essay, we observe that knowledge indeed helps to identify the minority attack class, both in terms of precision and recall with ARGCON+LINK providing the greatest contribution to the precision of the attack class and a slight improvement in the recall of support. This gives strong evidence to support our hypothesis that the knowledge provided by our methods is effective for distinguishing between relation types.

We see a similar phenomenon emerge in the M-Arg dataset, where there is a consistent improvement in the identification of the smallest attack class, as we go from no knowledge to ARGCON to ARGCON+LINK. We also notice a significant, 33% increase in the recall of supporting AU pairs. However, as we introduce knowledge to our model, we see that the recall of neutral pairs decreases alongside the precision of support. This is what was observed in Subsection 6.4 due to the appearance of supporting inference chains when we generate knowledge with ARGCON.

To summarise, the improvement of our model on the Student Essay and M-Arg dataset provides evidence to support continued investigation and improvement of commonsense-aided relational argument mining with COMET. In particular, we see that external commonsense knowledge significantly improves the identification of minority classes in imbalanced datasets which are particularly common to the domain of argumentation.

	Knowledge	**Attack**			**Support**			**Neutral**		
		P	R	F1	P	R	F1	P	R	F1
Debate	-	0.71	0.80	0.75	0.79	0.71	0.75	-	-	-
	ConceptNet$^+$	0.72	0.75	0.74	0.76	0.73	0.75	-	-	-
	ARGCON	0.73	0.77	0.75	0.78	0.75	**0.76**	-	-	-
	ARGCON+LINK	0.72	0.79	0.75	0.78	0.71	0.75	-	-	-
Essay	-	0.33	0.45	0.38	0.95	0.92	0.93	-	-	-
	ConceptNet$^+$	0.37	0.48	0.42	0.95	0.93	0.94	-	-	-
	ARGCON	0.36	0.44	0.40	0.95	0.93	0.94	-	-	-
	ARGCON+LINK	0.44	0.47	**0.46**	0.95	0.95	**0.95**	-	-	-
M-Arg	-	0.00	0.00	0.00	0.43	0.31	0.36	0.90	0.94	**0.92**
	ARGCON	0.14	0.07	0.09	0.37	0.60	**0.46**	0.92	0.88	0.90
	ARGCON+LINK	0.25	0.13	**0.17**	0.32	0.64	0.43	0.93	0.83	0.88

Table 8: Micro-F1 score comparison of the models in our experiments. Results with no knowledge represent our RoBERTa baseline.