Journal of E-Technology

Print ISSN: 0976-3503
Online ISSN: 0976-2930

DLINE JOURNALS

JET 2024; 15 (2)
https://doi.org/10.6025/jet/2024/15/2/51-63

Arguments Taxonomy System using Linguistic and Knowledge-based Features

Jonathan Kobbe, Ioana Hulpus, Heiner Stuckenschmidt
University of Mannheim, Germany
{jonathan@informatik.uni-mannheim.de}
{ioana@informatik.uni-mannheim.de} {heiner@informatik.uni-mannheim.de}
Juri Opitz, Maria Becker, Anette Frank
Heidelberg University, Germany
{opitz@cl.uni-heidelberg.de} {mbecker@cl.uni-heidelberg.de}
{frank@cl.uni-heidelberg.de}

ABSTRACT

Classifying Arguments Argument relations classification is a way of classifying the type of relationship between two argument units. Current models mainly rely on surface-level language features such as discourse markers, modal, or adverbial to classify the relationship. However, a model that primarily relies on language features to classify an argument can be easily misled by the style rather than the content of the argument, particularly when a weak argument is masked by strong language. This paper examines the challenges and potential advantages of knowledge-based argument analysis in advancing the current state of argument analysis towards a deeper, knowledge-driven comprehension and representation of arguments. We propose an Arguments Classification System that uses linguistic and knowledge-based features to classify Arguments. We start with a Neural Baseline Model for classifying a Pair of Arguments based on the Siamese Network and expand it with a set of Features derived from two additional background knowledge sources: ConceptNet and DBpedia.

Received: 20 November 2024
Revised: 13 February 2024
Accepted: 22 February 2024

Keywords: Argument Structure Analysis, Background Knowledge, Argumentative Functions, Argument Classification, Commonsense Knowledge Relations

1. Introduction

Attack and support are two important relations that can hold between argumentative units.

Consider the following two argumentative units (1) and (2) that are given in response to the topic (0) Smoking should be allowed in every restaurant:

(1) Smoking is a significant health hazard.

(2) Combustion processes always produce toxins.

Both (1) and (2) have a negative stance towards the topic (0), and at the same time they stand in a support relation themselves: (2) supports (1). In textual discourse, this relationship is often indicated with discourse markers, e.g., because (i.e., (1) because (2)), or therefore (i.e., (2), therefore (1)). Similarly, attack relations are frequently marked with discourse markers, e.g., A, however, B, etc. Although in the given example, the argumentative units (1) and (2) have no words in common and do not include discourse markers, a human can easily determine the support relation between them. This can be done for instance by recognizing relations that connect the two units like the fact that smoking generally involves a combustion process and that toxins are detrimental to health.

While accessing such knowledge is seamless for humans, it is much more challenging for machines. State-of-the-art machine learning systems for argument analysis (for instance [27] or [1]) mainly rely on the exploitation of shallow linguistic markers (such as adverbials, discourse connectors or punctuation) and largely ignore background knowledge and common sense reasoning as evidences for classifying argumentative relations. We argue that for building reliable systems, world knowledge and common sense reasoning should be core criteria and evidences for determining whether an argumentative unit A attacks or supports B. Rather than solving the argumentative relation classification or argumentation structure reconstruction task by using only linguistic indicators that characterize the rhetorics of the argument, we emphasize the need of systems that are able to capture the underlying logics of an argument by analyzing its content.

Clearly, this is a challenging task, as it requires appropriate knowledge sources and reasoning capacities. However, exploiting the knowledge relations that hold between argument units carries an immense potential of explaining, in interpretable ways, why an argument holds (or does not hold), when presenting supporting or attacking evidence. We therefore use the opportunity brought by the current advances in the Linked Open Data movement, and investigate the potential of external, structured knowledge bases such as ConceptNet and DBpedia, for providing the required background knowledge. Specifically, we propose a series of knowledge-based features for argumentative relation classification and analyze their impact as compared to surface-linguistic features as used in current state-of-the-art models. Starting with a linear regression classifier, we proceed to a stronger Siamese neural network system that encodes pairs of argumentative units to classify their relation. This system, when enriched with knowledge-based features, yields considerable performance improvements over the non-enriched version, and thus offers clear indications for the prospects of knowledge-enhanced argument structure analysis.

Our contributions are as follows: (i) we propose features that extract background know?ledge from two complementary knowledge resources: ConceptNet and DBpedia and analyze their respective impact on the task; (ii) we show that a neural system enriched with back?ground knowledge obtains considerable performance gains over the non-enriched baseline. In sum, our work is one of the first to shows positive impact of background knowledge on argument classification.

2. Related Work

2.1. Argument Structure Analysis

Stab and Gurevych, (2014) [26] propose an approach for (1) identifying argument components and (2) classifying the relation between pairs of argument components as either supportive or non-supportive. They propose several features, including structural features (e.g. number of tokens of the argument component, token ratio between covering sentence and argument component), lexical features (n-grams, verbs, adverbs, modals), syntactic features (e.g. production rules as proposed by [13]), contextual features (e.g. number of punctuations and number of tokens of the covering sentence), and further indicators such as discourse markers and pronouns, which are fed into a SVM classifier. When trained on the corpus of student essays *that we also use in this work [25], the system obtains F1-scores of up to 0.726 for identifying argument components and 0.722 for distinguishing support from non? support relations. Following up the task of argument structure analysis, Stab and Gurevych, (2017)[27] propose an end-to-end approach where they first identify argument components using sequence labeling at the token level. For detecting argumentation structures, they then apply a model which jointly distinguishes argument component types (major claim, claim, premise) and argumentative relations (linked vs. not linked) using Integer Linear Programming. Finally, the stance recognition model differentiates between support and attack relations using a SVM classifier with lexical, sentiment, syntactic and structural features (similar to the features used in their previous work [26]) as well as PDTB discourse relations and combined word embeddings. They evaluate their model on the student essay corpus and the Microtext corpus [19], achieving F1 scores of 0.68 and 0.75 respectively on the task of stance classification (support vs. attack). Similar to Stab and Gurevych [26, 27], Persing and Ng (2016) [21] propose an End-to-End system for identifying argument components and the relations that occur between them in the student essay corpus. Their baseline system is a pipeline which first extracts argument components heuristically and then distinguishes firstly between argumentative and non-argumentative spans and subsequently between attack vs. support vs. not related relations. For both classifiers they apply maximum entropy classification, using the same features as Stab and Gurevych [26, 27]. This baseline system is outperformed by a joint model which uses global consistency constraints to perform joint inference over the outputs of the single pipeline tasks in an ILP framework, achieving F1 scores of up to 38.8% for the relation identification task.*

The features used in these approaches are partly also used in our Baseline system (e.g. sentiment, token and punctuation statistics, modal verbs). Nonetheless, in this work we take a step further, by leveraging external knowledge bases such as DBpedia and ConceptNet in addition to our linguistic feature set.

Nguyen and Litman (2016)[16] also address the task of argumentative relation classification based on the student essay corpus. They adapt Stab and Gurevych's (2014) [26] system by adding contextual features extracted from surrounding sentences of source and target components as well as from topic information of the writings. For identifying attack relations, they achieve up to 0.33 F1 scores, and for support relations 0.94 F1 scores, which shows that contextual features are helpful for the task of relation classification. In contrast, we aim for an approach that is agnostic of the context in which the argument units originally occur.

Most existing work on argument analysis focuses on classifying relations between argument units in monologic argumentation, partly due to the used /available datasets. Since our aim is to assess pairs of argument units regardless of whether they belong to the same monologue, we create a new dataset, sourcing pairs of argumentative units from Debatepedia1. In this regard, our work is comparable to Hou and Jochim's (2017) [9], who learn to predict for pairs of argument units stemming from different texts in Debatepedia whether they are in agreement or disagreement with each other. They apply various models including an attention-based LSTM, a textual entailment system, and classification models trained by logistic regression. Their best performing system utilizes the mutual support relations between argumentative relation classification and stance classification jointly and achieves an accuracy of 65.5%, which confirms that there is a close relationship between argumentative relation classification and stance classification.

The relation between our task of argumentative relation classification and the task of stance classification has also been discussed by Peldszus and Stede (2015) [18] and by Afantenos et. al (2018) [1]. Compared to the binary distinction (support vs. attack) in our work and in Hou and Jochim (2017) [9] (agree vs. disagree), the annotation of their argumentation structure is more fine-grained and contains several aspects. The structure follows the scheme outlined by Peldszus and Stede (2013) [17], where the different aspects are (1) finding the central claim of the text, (2) predicting the relation between that claim and the other segments, (3) predicting

[1] *http://www.debatepedia.org*

the relation between the other segments, (4) identifying the argumentative role of each segment, and (5) predicting the argumentative function of each relation. Similar to Hou and Jochim (2017) [9], they show that joint predictions - in this case the prediction of all these levels in the evidence graph - help to improve the classification on single levels.

Menini and Tonelli (2016) [15] also address the task of distinguishing agreement vs. disagreement relations of argument components in a dialogic setting, investigating documents from political campaigns and Debatepedia. They introduce three main categories of features: sentiment-based features (e.g. the sentiment of the statements and sentiment of the topic), semantic features (e.g word embeddings, cosine similarity and entailment), and surface features (e.g. the lexical overlap and the use of negations). Using all features jointly as input to an SVM classifier, they achieve up to 83 % accuracy on the political campaign dataset and 74 % accuracy on Debatepedia.

2.2. Background Knowledge for Argument Analysis

External knowledge resources have been leveraged as supporting information for various tasks in NLP, including Argument Analysis. Potash et al. (2017) [22] assess the feasibility of integrating Wikipedia articles when predicting convincingness of arguments and find that they can provide meaningful external knowledge. Habernal et al. (2018) [7] claim that comprehending arguments requires significant language understanding and complex reasoning over world knowledge, especially commonsense knowledge. Incorporating external knowledge is therefore viewed as essential for solving the SemEval Argument Reasoning Comprehension Task (2018 Task 12, [7]) [2]

This can be confirmed by the results of the participating systems: *The best performing system, proposed by Choi and Lee [6], is a network transferring inference knowledge to the argument reasoning comprehension task. It makes use of the SNLI dataset [4] and benefits from similar information in both datasets. This system outperforms all other systems by more than 10%. Besides pretrained word embeddings (e.g. contexualized embeddings, [11]) and a sentiment polarity dictionary [5], none of the other published systems takes into account external knowledge resources for solving the task.*

Following up on the observation about the usefulness of external knowledge for argument?ative reasoning, the approach of Botschen et al. (2018) [3] leverages event knowledge from FrameNet and fact knowledge from Wikidata to solve the Argument Reasoning Comprehension task. They extend the baseline model of Habernal et al. (2018) [7], an intra-warrant attention model that only uses conventional pretrained word embeddings as input, with embeddings for frames and entities derived from FrameNet and Wikipedia, respectively. They conclude that external world knowledge might not be enough to improve argumentative reasoning. However, motivated by the promising results of Becker et al. 2017 [2] who have shown that commonsense knowledge that is useful for understanding Microtext arguments can be mapped to relation types covered by ConceptNet, we analyze additional knowledge bases, specifically ConceptNet for commonsense knowledge and DBpedia for world knowledge.

3. Knowledge Graph Features

For exploiting background knowledge, we designed features based on two knowledge graphs: ConceptNet[3] *and DBpedia* [4]*. We expect ConceptNet to contain valuable information about common sense knowledge while DBpedia captures encyclopedic knowledge. The core idea is to connect pairs of argumentative units via relations in the knowledge graphs and to use the relation types and the extracted paths as features. The intuition is that certain types of paths or relations, like e.g. the Antonym relation in ConceptNet, occur more often in disagreeing and therefore attacking pairs of statements than in supporting ones and vice versa.*

[2] Given an argument consisting of a claim and a reason, the task is to select one out of two potential inferential licenses, called warrants, that explains the reasoning underlying the argument.

[3] http://conceptnet.io

[4] http://dbpedia.org

Given two argumentative units, we first proceed to link them to the external knowledge bases. Section 5.2 provides the entity linking details. Once the two argumentative units are linked, we represent them as sets A and B of their linked entities. We then pair all the elements in A to those in B. For each such pair $(x, y) \in A \times B, x = y$, we extract all the paths from x to y up to length three within the knowledge base. Figure 1 shows a graph consisting of such paths extracted from ConceptNet. As one can see in the graph, each path consists of nodes connected by directed edges labeled with relation types. As mentioned above, we assume that those relation types contain valuable information. For that reason, we design two kinds of features that rely on them: First, we check how often a certain relation type occurs along all paths between all pairs $(x, y) \in A \times B, x = y$ and divide that number by the total count of edges. This way, each relation type is a numerical feature on its own and all those features together sum up to 1. Second, we specifically exploit the paths. Since there are too many paths to create one feature per path, we group them via patterns. Each pattern is a multiset of relation types. For example, given the pattern [Synomym,RelatedTo,RelatedTo], the graph in Figure 1 contains two paths between mathtwo paths between math and computer that instantiate this pattern:

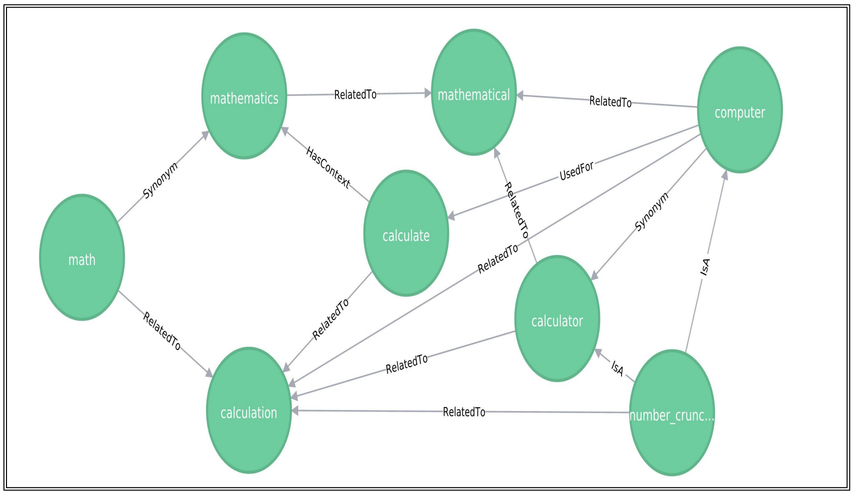

Figure 1. Connection between math and computer in ConceptNet, generated using Neo4j[5]

$$math \xrightarrow{Synonym} mathematics \xrightarrow{RelatedTo} mathematical \xleftarrow{RelatedTo} computer$$

$$math \xrightarrow{RelatedTo} calculation \xleftarrow{RelatedTo} calculator \xleftarrow{Synonym} computer$$

Each such path pattern corresponds to a numerical feature whose value is the number of its instantiations divided by the total number of paths. As some of the relation type-based and path-based features described above occur only rarely, we only use those features that occur in at least one percent of all the instances in the training data.

Besides exploiting the relation types and paths, we also hypothesize that the length and number of paths are useful for classification, as they provide an indication to the relatedness of A and B [10]. To account for this, we additionally compute (i) one feature representing the total number of paths divided by |A| · |B|, (ii) three features representing the number of paths of a certain length i (i ∊ {1, 2, 3}) divided by the total number of paths, (iii) one feature representing the total number of identical entities in A and B divided by |A| · |B| and (iv) one feature with the count of all the different nodes along all paths divided by |A| · |B| again.

[5] *https://neo4j.com/*

4. Neural Network Model

We design a Siamese neural network model for argumentative relation classification (NN). The architecture of the model is displayed in Figure 2. It consists of one Bi-LSTM [8], which is used to embed two argumentative units A and B into a common vector space. More precisely, sequences of word embeddings[6], $(e(w_1^A), ..., e(w_n^A))$ and $(e(w_1^B), ..., e(w_m^B))$ are fed through the Bi-LSTM to induce representations $emb(A), emb(B) \in \mathbb{R}^{2h}$, where h is the number of the two LSTM's hidden units (we concatenate the last states of the forward and backward pass of each LSTM). Based on the argument representations emb(A) and emb(B) we then compute a representation for the relation holding between these units by computing the difference vector between their representations emb(A) and emb(B): r(A, B) = emb(B) - emb(A). The obtained representation for the relation can be further enriched by adding, e.g., features extracted from an external knowledge base that represent relevant information about knowledge relation paths connecting concepts and entities mentioned in the two argumentative units (cf. Section 3 and relation features derived from KB, Figure 2). The vector vK(A, B) that encodes such knowledge features is concatenated to the argument relation vector r(A, B) to yield the extended vector representation r′ (A, B) of the argumentative relation: r′ (A, B) = r(A, B) ⊕ v_K(A, B), where x ⊕ y denotes concatenation of vectors x, y. This final relation representation is further processed by a fully connected feed-forward layer (FF, Figure 2) with two output units and softmax-activations for providing the support and attack probabilities.

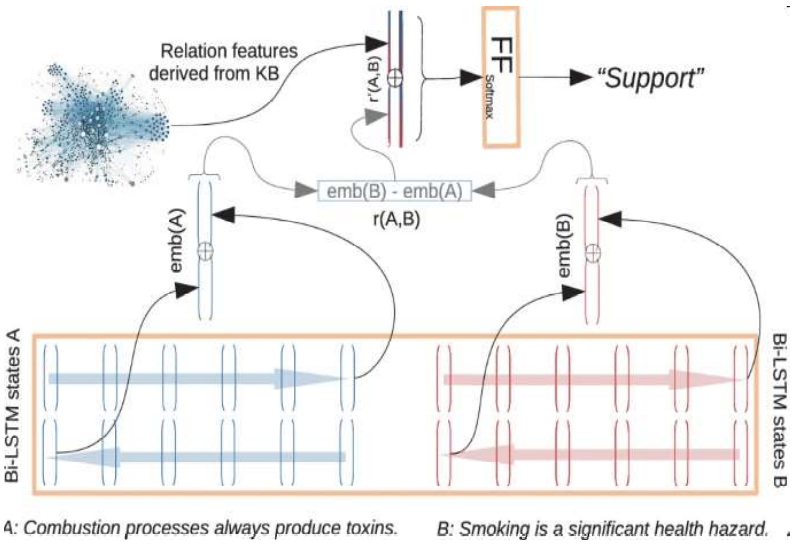
Relation features derived from KB
r'(A,B)
FF
Softmax
"Support"
emb(B) - emb(A)
r(A,B)
emb(A)
emb(B)
Bi-LSTM states A
Bi-LSTM states B
A: Combustion processes always produce toxins.
B: Smoking is a significant health hazard.

Figure 2. Architecture of the Siamese neural argumentative relation classifier. After embedding the argumentative units, their relation is defined as the vector offset between the unit representations in argument space. This representation can be enriched with a feature vector derived from background knowledge sources (e.g., ConceptNet)

	Debatepedia	*Microtexts*	*Student essays (Essays)*
Total number of relations	*14,441*	*308*	*1,473*
Number of attack relations	*7,184*	*84*	*161*
Number of support relations	*7,257*	*224*	*1,312*

Table 1. Data statistics for the different experimental datasets

[6] *we use pre-trained 300d Glove vectors [20].*

5. Experiments

We conduct experiments on three argumentative data sets from different domains, which will be described in the following section. Because we want the models to focus on the background knowledge involved in the argumentation, we consider only the argumentative units without their context and position. This increases the difficulty of the task as models are prevented from exploiting contextual and positional features.

5.1. Data

Student Essays (Essays)

The student essays consist of 90 persuasive essays in the English language. The essays were selected from essayforum7 and annotated by [25]. The corpus contains 1473 annotated argumentative relations: 1312 were labeled as support and the remaining 161 were labeled as attack relations. We apply the same split between training and test data as [26] and [16]. For our purpose, we make use of pairs of attacking and supporting argumentative units and dismiss all other information about the position and context and the annotated argumentative components and stances.

Microtexts. *This corpus consists of 112 short argumentative texts [19]. The corpus was created in German and has been translated to English. We use only the English version. The corpus is annotated with argumentation graphs where the nodes are argumentative units and the edges are argumentative functions. We again collect pairs of attacking and supporting argumentative units. Therefore, we consider only direct connections between two argumentative units that are labeled as support or rebut. We deliberately ignore the undercut function as an undercut is an attack on the argumentative relation between two argumentative units. This way, we extract 308 argumentative relations whereof 224 are support and 84 are attack relations. To achieve a proper split between training and testing data, we use all the Microtexts about public broadcasting fees on demand, school uniforms, increase weight of BA thesis in final grade and charge tuition fees for testing and all the others for training.*

Debatepedia. *This is a website where users can contribute to debates on some specific topic[8]. Most debates consist of a title, a topic that is formulated as a polar question (e.g. Should the legal age for drinking alcohol be lowered?), subtopics and arguments that are either in favor or against the topic. We crawled the Debatepedia website and extracted all arguments with a valid URL. In many arguments, the argument's claim is highlighted, so we used this feature to identify the claims, and removed the arguments that did not have any highlighted text. This resulted in 573 debates. We generate the pairs of argument units by pairing the topic of the debate to the claim. If the argument is in favor of the topic, then its claim supports the topic, else it attacks the topic. This way, we generate a large corpus containing 14441 pairs of argument units whereof 7257 are in support and 7184 are in attack relations. We arbitrarily chose 114 (20%) out of the 573 debates for testing and use the rest for training[9].*

5.2. Knowledge Graphs

DBpedia.[10] This knowledge graph contains information from Wikipedia[11] in a structured way. The English version contains more than 4 million entities classified in an ontology. For our work with DBpedia, we included the following datasets in English version in addition to the DBpedia Ontology (Version 2016-10): article categories, category labels, instance types, labels, mapping-based objects and SKOS categories. To achieve less meaningless paths, we excluded all the resources whose URI starts with Category:Lists_of, List_of,

[7] https://essayforum.com/

[8] http://www.debatepedia.org/

[9] For information about accessing the data, see http://explain.cl.uni-heidelberg.de/.

[10] https://wiki.dbpedia.org/

[11] https://www.wikipedia.org/

Glossary_of, Category:Glossaries_of, Images_of, Category: Indexes_of, Category: Outlines_of, Category:Draft-Class, Category:Wikipedia as well as the resource owl:Thing. For linking tokens in the argumentative units to entities in DBpedia, we use DBpedia Spotlight[12] with a minimum confidence of 0.3 and support of 1.

ConceptNet.[13] *ConceptNet is a crowd-sourced resource of commonsense knowledge created by the Open Mind Common Sense (OMCS) project [23], to which were later added expert?created resources [24]. It has been built in response to the difficulties of automatic acquisition of commonsense knowledge. The current version, ConceptNet 5.6, comprises 37 relations, some of which are commonly used in other resources like WordNet (e.g. IsA, PartOf) while most others are more specific to capturing commonsense information and as such are particular to ConceptNet (e.g. HasPrerequisite or MotivatedByGoal). We use the English version of ConceptNet 5.6 which consists of 1.9 million concepts and 1.1 million links to other databases like DBpedia for instance. We deleted all self-loops as they don't contain any valuable information. Linking of tokens to ConceptNet is done in a straightforward way: We split the argumentative unit into maximum-length sequences of words that can be mapped to concepts. If a concept consists only of stop words or has a degree of less then three, it is dismissed[14]. This way, unconnected and only weakly connected concepts are avoided. If a concept consists of a single word, we use Stanford CoreNLP ([14]) to find out whether this is an adjective, noun or verb, in order to link it to the appropriate concept in ConceptNet, if possible.*

5.3. Baselines

In this paper, we focus on local argumentative relation classification, thus our work is not directly comparable to prior work which proposes global, i.e., contextually aware classifiers for this task [26, 16, 18]. More specifically, we are interested in a classification setup that is agnostic of the contextual surface features such as discourse markers and position in discourse, and that restricts classification to the analysis of two argumentative units combined with the background knowledge that connects them.

Nevertheless, in order to compare to knowledge-lean paradigms of related work, we replicate features used in the most related previous work [26, 15]. To this end, we train a linear classifier with the replicated (linguistic) features, which we denote as Ling. As Ling features we use the sentiment of both argumentative units as features, as described in [15]. We simplified the negation features of [15] and use Stanford CoreNLP ([14]) to only recognize whether there is some negation in an argumentative unit. From [26] we adopted the structural features which contain token and punctuation statistics and two features indicating whether a modal verb occurs. Additionally, we use each pair of words, one from each argumentative unit, as a binary feature. We only included pairs that do not contain a stopword and occurred in at least one percent of all the training instances.

5.4. NN Model Optimization and Configurations

Optimization. *We split the data into a training and a test set as described in section 5.1. For development purposes, we once randomly split off 200 examples from the training data of Debatepedia and Essays and 100 examples from the smaller Microtexts data. Let the training data be defined as* $D = \{(x_i, y_i)\}_{i=1}^{N}$*, where* x_i *consists of a source and target argument unit and* $y_i \in \{0, 1\}^2$ *is the one-hot vector corresponding to the two relation classes: (support, attack). Let, for any datum indicated by i,* $p_{i,s}$ *be the support-probability assigned by our model and* $p_{i,\ a}$ *the attack-probability. Using stochastic mini batch gradient descent (batch size: 32) with Adam [12], we minimize the categorical cross entropy loss over the training data, H, computed as in Equation 1:*

[10] https://wiki.dbpedia.org/

[11] https://www.wikipedia.org/

[12] https://www.dbpedia-spotlight.org/

[13] http://conceptnet.io/

[14] We use the default stopwordlist from https://www.ranks.nl/stopwords including can.

$$H = -\frac{1}{N}\sum_{i=1}^{N}(y_{i,s} \cdot \log p_{i,s} + y_{i,a} \cdot \log p_{i,a}), \tag{1}$$

where $y_{i,s}$ = 1 if observation i is classified as support and 0 otherwise (and similarly $y_{i,a}$ = 1 if observation i is classified as attack and 0 otherwise). We optimize all parameters of the model except the word embeddings.

Configurations. *Building on our basic Siamese model (NN), we inject (i), the graph features derived from ConceptNet (NN+CN); (ii), the same features but derived from DBpedia (NN+DB) and (iii), a concatenation of both (NN+DB+CN). For comparison purposes, we also run experiments using only the feature vector derived from the knowledge base. This is achieved by basing the classification only on this feature vector (obtained from DBpedia (DB), ConceptNet (CN) or DBpedia+ConceptNet (DB+CN)), ignoring and leaving out the embedded relation. Instead of concatenating knowledge features to our Siamese relation classification model, we also perform experiments where we concatenate the linguistic feature vector to the argument relation embedding (NN+Ling). Our full-feature argumentative relation classification model is NN+Ling+CN+DB.*

5.5. Results

Table 2 presents the F1 scores that our evaluated models obtain on all three datasets. The main observation is that overall, the knowledge base enhanced model NN+Ling+CN+DB achieves the best results. Second, the baselines Ling, random and majority are outper?formed by all configurations of the neural Siamese model NN on all three data set.

The performance of our basic Siamese model (NN), for almost all evaluation metrics and data sets, is situated between Ling and all NNs which are augmented with knowledge. NN outperforming Ling indicates that the neural model is able to capture surface features not explicitly modeled by Ling. However the combination NN+Ling does achieve better results than NN suggesting that the two types of features are complementary.

With respect to knowledge enhanced models, both NN+CN and NN+DB outper?form NN in terms of macro-F1, indicating that they manage to successfully use external knowledge. However, our experiments show no benefit from bringing together features from both ConceptNet and DBpedia on top of the NN system, a result that requires more investigation. Nevertheless, when ConceptNet and DBpedia features are brought together on top of NN+Ling features, the system achieves the best results. Training a linear classifier solely with the

	F1 scores								
	Debatepedia			Microtexts			Essays		
	support	attack	macro	support	attack	macro	support	attack	macro
random	$50.2^{\pm1}$	$50.1^{\pm1}$	$50.2^{\pm1}$	$73.0^{\pm5}$	$27.8^{\pm11}$	$50.4^{\pm8}$	$89.2^{\pm1}$	$10.5^{\pm4}$	$49.8^{\pm3}$
majority	**66.3**	0.0	33.2	82.1	0.0	41.1	**94.9**	0.0	47.5
Ling	61.4	49.8	55.6	73.3	42.9	58.1	**94.9**	0.0	47.5
DB	43.7	56.8	50.2	81.1	0.0	40.5	94.8	0.0	47.4
CN	45.6	55.1	50.3	65.9	31.8	48.9	**94.9**	0.0	47.5
DB+CN	46.4	55.3	50.8	82.1	0.0	41.1	**94.9**	0.0	47.5
NN+Ling	58.1	55.7	56.9	77.7	35.2	66.7	92.7	20.7	56.7
NN	58.6	57.6	58.1	74.2	46.5	60.3	78.7	17.1	47.9
NN+DB	56.8	**59.7**	58.2	77.4	46.2	61.8	84.1	19.5	51.8
NN+CN	60.3	56.8	**58.6**	**83.5**	41.4	62.4	86.5	20.2	53.3
NN+DB+CN	58.6	57.6	58.1	81.2	38.7	59.9	88.0	16.3	52.1
NN+Ling+CN+DB	58.6	56.2	57.4	82.5	**51.4**	**67.0**	91.2	**25.7**	**58.7**

Table 2. Results over different systems and data sets

	vs. NN baseline								
	Debatepedia		Microtexts		Essays		Total		
	Δ sup.	Δ att.	Δ sup.	Δ att.	Δ sup.	Δ att.	Δ sup.	Δ att.	Δ att. + Δ sup.
Ling	**153**	-231	0	-2	**95**	-12	**248**	-245	3
NN+DB	-47	**22**	3	**-1**	26	**-1**	-18	**20**	2
NN+CN	63	-78	**10**	-4	39	-2	107	-84	**23**
NN+DB+CN	13	-43	8	-4	49	-5	70	-52	18

Table 3. Number of cases which were labeled incorrectly by the NN baseline but correctly by another model minus the number of cases which were labeled correctly by the NN baseline but incorrectly by another model. Worst and best values are highlighted.

from both ConceptNet and DBpedia on top of the NN system, a result that requires more investigation. Nevertheless, when ConceptNet and DBpedia features are brought together on top of NN+Ling features, the system achieves the best results. Training a linear classifier solely with the background knowledge features achieves lower results than the Ling baseline, and also lower than all other configurations on top of NN. This indicates that the knowledge features are only useful when in conjunction with text based features.

With respect to the two targeted argumentative relation classes, attack relations are more challenging to capture in the Microtexts and Essays datasets, because of the very low frequency in the data (see Table 1). It is interesting to notice that on our biggest and most balanced dataset (Debatepedia), NN+DB provides more accurate detection of attack relations than of support relations, and that overall the settings that use DBpedia achieve better results at detecting the attack relation, than the settings that do not use DBpedia. This might be because DBpedia does not capture lexical knowledge, therefore attacking concepts lie further away in the graph than they do in ConceptNet. This is a very interesting insight and worth more investigation in the future.

Comparative Analysis of the Neural Models. To give deeper insights into the performances of our knowledge enhanced models, we present a deeper comparison between them and the NN and Ling predictions. The results over all three data sets are displayed in Table 3. In total, NN+CN provides most corrections of otherwise falsely classified cases (+23 over all data sets; -15 on Debatepedia, +6 on Microtext and +37 on Essays). A correction of a false-positive attack label (+107 in total) appears to be more likely than a correction of a false-positive support label,

argumentative unit A (source)	argumentative unit B (target)	y	Δ
prohibition has kept marijuana out of children's hands	prohibition does more harm than good	ATT	0.66
using technology or advanced facilities do not make food lose its nutrition and quality	investing much time in cooking food will guarantee nutrition as well as quality of food for their family	ATT	0.15
they will have a bad result in school	even people who are not interested in online game can still be negatively affected by using computer too much	SUP	0.84
Education and training are fundamental rights which the state , the society must provide	Tuition fees should not generally be charged by universities	SUP	0.38

Table 4. Examples from Microtext and Essays which were assigned a significantly higher probability for the correct label by the knowledge-augmented model (NN+CN) compared to our neural baseline model (NN).

in fact, for the attack label, the knowledge augmented model makes more mis-corrections than corrections (-84 in total, with the strongest such effect on Debatepedia). This means that the knowledge helps the model in determining support relations more than in determining attack relations. Overall, the knowledge-enhanced models, especially NN+CN, tend to have a better overall correction ratio compared to Ling.

Examples. *To understand where the injection of background knowledge helps the most, we investigated the AU pairs which were falsely classified by NN but correctly classified by NN+CN.We rank these cases according to the margin pNN+CN (c) - pNN (c), where p(c) is the probability of the correct class. Four cases with large margins are displayed in Table 4. In the first example, there is only one explicit link in the form of a shared word (prohibition). The attack-relation has its foundation in the fact that A probably views prohibition (of marijuana) rather positively. His belief is based on the premise that children are protected by prohibition – the protection of children from drugs is widely considered as something highly desirable. On the other hand, B views prohibition more negatively and thus B can consider itself attacked by A. The baseline NN mislabeled the relation as a support relation, assigning the attack relation a low probability. The knowledge augmented model, in contrast, predicted the correct label very confidently. All four examples have in common that there are no shallow markers which somehow could predict the outcome. For proper resolution of these examples, knowledge about the world needs to be applied in conjunction with knowledge about syntax (e.g., by removing the negation from the fourth example, the support relation transforms into a attack relation).*

6. Conclusion

In this paper, we have investigated the use of background knowledge for argumentative relation classification. We introduced a Siamese neural network system that uses word embeddings and can be enriched with specifically designed feature vectors. We designed features that exploit knowledge graphs such as ConceptNet and DBpedia and evaluate their usefulness. Experimental results on three different corpora show that knowledge based features capture aspects that are complementary to the surface features, and can substantially improve the classification results.

Our presented study is a first step towards a knowledge-rich argument analysis and opens new research directions into investigating and exploiting knowledge graphs for argumentation understanding. We plan to explore more sophisticated ways to make use of background knowledge for argumentation structure reconstruction and for explaining arguments.

References

[1] Afantenos, S., Peldszus, A., Stede, M. (2018). Comparing decoding mechanisms for parsing argumentative structures. *Argument and Computation*, 9, 177–192.

[2] Becker, M., Staniek, M., Nastase, V., Frank, A. (2017). Enriching Argumentative Texts with Implicit Knowledge. In F. Frasinca, A. Ittoo, L. M. Nguyen, & E. Metais (Eds.), Applications of Natural Language to Data Bases (NLDB) - Natural Language Processing and Information Systems, Lecture Notes in Computer Science. Springer. Retrieved from http://www.cl.uni-heidelberg.de/~mbecker/pdf/enriching-argumentative-texts.pdf.

[3] Botschen, T., Sorokin, D., Gurevych, I. (2018). Frame- and Entity-Based Knowledge for Common-Sense Argumentative Reasoning. In Proceedings of the 5th Workshop on Argument Mining (pp. 90–96). Retrieved from http://aclweb.org/anthology/W18-5211.

[4] Bowman, S. R., Angeli, G., Potts, C., Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (EMNLP).

[5] Chen, Z., Song, W., Liu, L. (2018). TRANSRW at SemEval-2018 Task 12: Transforming Semantic Representations for Argument Reasoning Comprehension. In Proceedings of The 12th International Workshop on Semantic Evaluation (pp. 1142–1145). doi:10.18653/v1/

S18-1194.

[6] Choi, H. S., Lee, H. (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* (pp. 773–777). doi:10.18653/v1/S18-1122.

[7] Habernal, I., Wachsmuth, H., Gurevych, I., Stein, B. (2018). SemEval-2018 Task 12: The Argument Reasoning Comprehension Task. *In Proceedings of The 12th International Workshop on Semantic Evaluation* (pp. 763–772).

[8] Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735–1780.

[9] Hou, Y., Jochim, C. (2017). Argument Relation Classification Using a Joint Inference Model. In *Proceedings of the 4th Workshop on Argument Mining* (pp. 60–66).

[10] Hulpus, I., Prangnawarat, N., Hayes, C. (2015). Path-based semantic relatedness on linked data and its use to word and entity disambiguation. In *International Semantic Web Conference* (pp. 442–457). Springer.

[11] Kim, T., Choi, J., Lee, S. (2018). SNU_IDS at SemEval-2018 Task 12: Sentence Encoder with Contextualized Vectors for Argument Reasoning Comprehension. In *Proceedings of The 12th International Workshop on Semantic Evaluation* (pp. 1083–1088).

[12] Kingma, D. P., Ba, J. (2014). Adam: A Method for Stochastic Optimization. *CoRR, abs/1412.6980*. arXiv:1412.6980.

[13] Lin, Z., Kan, M.-Y., Ng, H. T. (2009). Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 343–351).

[14] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations* (pp. 55–60).

[15] Menini, S., Tonelli, S. (2016). Agreement and Disagreement: Comparison of Points of View in the Political Domain. In *COLING* (pp. 2461–2470).

[16] Nguyen, H. N., Litman, D. J. (2016). Context-aware Argumentative Relation Mining. In *ACL* (pp. 1127–1137).

[17] Peldszus, A., Stede, M. (2013). From Argument Diagrams to Argumentation Mining in Texts: A Survey. *Int. J. Cogn. Inform. Nat. Intell., 7*(1), 1–31.

[18] Peldszus, A., Stede, M. (2015). Joint prediction in MST-style discourse parsing for argumentation mining. In *EMNLP* (pp. 938–948).

[19] Peldszus, A., Stede, M. (2016). An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon 2015 / Vol. 2* (pp. 801–815). College Publications.

[20] Pennington, J., Socher, R., Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).

[21] Persing, I., Ng, V. (2016). End-to-End Argumentation Mining in Student Essays. In *HLT-NAACL* (pp. 1384–1394).

[22] Potash, P., Bhattacharya, R., Rumshisky, A. (2017). Length, Interchangeability, and

External Knowledge: Observations from Predicting Argument Convincingness. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 342–351).

[23] Singh, P. (2002). The Open Mind Common Sense Project. Retrieved from http://zoo.cs.yale.edu/classes/cs671/12f/12f-papers/singh-omcs-project.pdf.

[24] Speer, R., Havasi, C. (2012). Representing General Relational Knowledge in ConceptNet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)* (pp. 3679–3686). *European Language Resources Association* (ELRA).

[25] Stab, C., Gurevych, I. (2014). Annotating Argument Components and Relations in Persuasive Essays. In *COLING* (pp. 1501–1510).

[26] Stab, C., Gurevych, I. (2014). Identifying Argumentative Discourse Structures in Persuasive Essays. In *EMNLP* (pp. 46–56).

[27] Stab, C., Gurevych, I. (2017). Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics, 43*, 619–659.