Argumentative Relation Classification as Plausibility Ranking

Juri Opitz

Leibniz ScienceCampus "Empirical Linguistics and Computational Language Modeling"
Department for Computational Linguistics
Heidelberg University
69120 Heidelberg
`opitz@cl.uni-heidelberg.de`

Abstract

We formulate argumentative relation classification (support vs. attack) as a text-plausibility ranking task. To this aim, we propose a simple reconstruction trick which enables us to build minimal pairs of plausible and implausible texts by simulating natural contexts in which two argumentative units are likely or unlikely to appear. We show that this method is competitive with previous work albeit it is considerably simpler. In a recently introduced content-based version of the task, where contextual discourse clues are hidden, the approach offers a performance increase of more than 10% macro F1. With respect to the scarce *attack*-class, the method achieves a large increase in precision while the incurred loss in recall is small or even nonexistent.

1 Introduction

Argumentative relation classification (ARC) is dedicated to determining the class of the relation which may hold between two arguments or elementary argumentative units, EAUs[1]. For instance, consider the following *premises* given the *topic* or *conclusion* **(0)** *"Overall, marijuana is detrimental to your health."*:

(1) *Use of marijuana causes chronic bronchitis and airflow obstruction.*

(2) *Cannabis does not need to be smoked to receive its potential health benefits.*

In this case, **(1)** has a *positive stance* towards the conclusion **(0)**; in contrast to **(2)**, which has a negative stance towards the conclusion. Additionally, but not less importantly, we can say that **(2)**
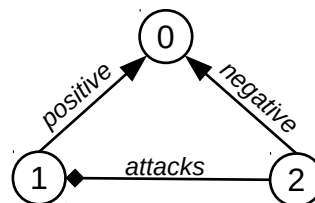

Figure 1: A small argumentation graph containing two general types of relations: premise-topic relations (class: negative/positive) and premise-premise relations (class: supports/attacks).

weakens **(1)** – it casts doubt about its generality by hinting at cannabis application methods which do not involve combustion or inhalation. In this work, we summarize all relations which aim at undermining or weakening another argument or premise ('undercut', 'rebuttal', etc.) as *attack*.[2] The EAUs from our example and their connecting relations are outlined in the graph in Figure 1.

In a rhetorically structured argumentative text[3], **(1)** and **(2)** may appear in configurations such as *On the one hand (1), on the other (2)*; *(1), however, (2)*, etc. Under these circumstances, discourse *context* can predict argumentative relations very well. However, when moving from such 'closed scenario' to a more 'open-world setting', e.g., where EAUs have been mined from heterogeneous documents, we need to determine relations based on their *content*. In this paper, we show that our method works well in both scenarios. In fact, it is in the more general and more difficult content-based setting, where our method provides the most benefits over previous work.

Systems which have learned to predict general argumentative relations have a decisive advantage when compared to systems that have 'only' learned to predict argumentative stances: in an argumenta-

---

[1]Here, we use the term *elementary argumentative units* to denote clauses or small clause-complexes – e.g., **(0)**, **(1)** or **(2)**) – which can be 'instantiated' in an argumentative debate.

[2]For a more 'in-depth' view and discussion of argumentative relations we refer the reader to, e.g., Pollock (1995), Walton (2009) and Besnard and Hunter (2014).

[3]E.g., an argumentative essay.

tive debate, often a debater does not choose to bring forth any argument which supports their stance on the topic. Instead, or additionally, they may choose to select an argument which also attacks the opponent's most recent argument. Therefore, we need not only knowledge about the stances of arguments towards topics, but also about relations to other arguments. Our experiments show that our approach is a step towards this goal.

The remainder of this paper is structured as follows: After discussing related work in Section 2, we propose a simple reconstruction trick which allows us to embed an argumentative source-target pair in a relational discourse context yielding a plausible and implausible text variant (Section 3). In Section 4, we conduct experiments and ablation studies using (i) a standard task setup, where systems are allowed to see EAUs in their document context and (ii) a more difficult 'content-focused' task setup where systems are only allowed to see the spans of the EAU clauses. The code for this paper is available at `https://gitlab.cl.uni-heidelberg.de/opitz/pr4arc`

2   Related work

In this section, we first provide an overview of the data, and the data issues people are confronted with when developing argumentative relation classification (ARC) systems. We proceed with an overview of existing ARC approaches and conclude by touching on other related tasks.

Argumentative relation data   For general argumentative relations, not many data sets have been developed. One of the largest data sets consists of 402 argumentative student essays and is henceforth denoted by ESSAY (Stab and Gurevych, 2014; Stab and Gurevych, 2017). It has been annotated, i.a., with EAU clauses and more than 3,000 relations which hold among them. By ESSAY-CONTENT, we denote a version of ESSAY from which discourse context is stripped and systems can only access the spans of EAU clauses (Opitz and Frank, 2019b). This setup is more difficult since systems have to learn to model the *content* of two EAUs in order to successfully predict their relation. ESSAY and ESSAY-CONTENT will be more extensively described in Section 4.1, where we also show that our method is efficient across both setups.

Another data set which is annotated with in-depth argumentative annotations is the Microtext corpus covering a variety of political debates in Germany (Peldszus and Stede, 2016). While it has been annotated with a more fine-grained set of relations (e.g., *rebutting attack, undercutting attack, linked support, example support*) it is rather small in size (the recently extended version (Skeppstedt et al., 2018) contains about 700 relation tuples). Similar to ESSAY-CONTENT, a variant of the Microtext corpus exists where argumentative units are detached from discourse context (Wachsmuth et al., 2018). We believe that systems that have learned to predict argumentative relations based on the *content* of argumentative units have advantages over systems which focus too much on contextual discourse clues. For example, content-focused systems can better be expected to solve large-scale cross-document tasks where EAUs are mined from many heterogeneous documents. Our reconstruction trick provides one step towards this goal: it exploits *potential* discourse configurations without depending on seeing the *true* discourse context.

A key reason for the data scarcity of annotated general argumentative relations is that creating high-quality data for 'premise-premise' relations is a challenging task. Perhaps, it is more challenging than creating data for argumentative stance detection since topics or conclusions are often 'a-priori' well understood (e.g., *Cannabis should be legalized*) and always occur as the stance-relation target. In that sense, it may be easier and quicker to tell if an argument supports a conclusion compared to deciding whether an argument supports another argument.

ARC systems   A linear SVM classifier that is trained on a diverse set of features provides competitive performance on ESSAY (Stab and Gurevych, 2017). A subsequent joint global graph optimization step, similarly to (Peldszus and Stede, 2015; Hou and Jochim, 2017), yields no further improvement for classifying the relations in this data. The SVM classifier incorporates features extracted from the EAU spans as well as their context (e.g., leading or trailing words). On ESSAY-CONTENT, where systems only see the EAU clause spans, the performance of the SVM suffers a loss of more than 10 pp. macro F1 (Opitz and Frank, 2019b) – an analysis indicates that the SVM focuses immoderately on features extracted from the EAU context and tends to neglect their actual content. This underpins the need for argumentative relation classification systems with deeper understanding of argumentation, i.e., systems that base their prediction on the actual

content of two EAUs – the method we present in this paper aims at this.

The first neural approach for ARC (Cocarascu and Toni, 2017) proposes a neural network with a Siamese structure (Koch et al., 2015; Mueller and Thyagarajan, 2016; Cocarascu and Toni, 2017). By means of a shared weight space it projects source and target EAU to a joint distributional vector space. Finally, it classifies the vector offset using a softmax-function. The authors conduct experiments on a data set which comprises texts about movies, technology and politics.

A similar model has been adopted recently where (symbolic) knowledge from large background knowledge-graphs is injected into the Siamese model by concatenating highly abstracted multi-hop knowledge paths to the source-target offset (Kobbe et al., 2019). Although there are consistent gains observed by including the knowledge, the gains appear to be relatively small. In this aspect, we believe that incorporating knowledge of the right form could make it possible to further enhance the system we propose in this paper. However, as of now, it is an active topic of discussion *whether* (symbolic) background knowledge may help in automatic argumentation and, even more so, *which* (form of) knowledge would be needed.

Computational argument mining and analysis Argumentation is ubiquitous and argumentative structures can be recovered from a broad spectrum of texts. For example, they can be recovered from online dialogue (Swanson et al., 2015; Budzynska et al., 2014) and scientific research articles (Lauscher et al., 2018a; Lauscher et al., 2018b), where, e.g., researchers may directly or indirectly convey arguments for why some method is better than another. By now, there exists a substantial body of research publications covering a variety of argument analysis topics. For a general overview, we refer the reader to Lippi and Torroni (2016) and Peldszus and Stede (2013).

Plausibility ranking Another task that can be addressed as a text plausibility ranking task is the resolution of difficult pronouns in the *Winograd Schema Challenge* (Levesque et al., 2012; Opitz and Frank, 2018). To resolve shell nouns and abstract anaphora (e.g., 'I like *that*'.) Marasović et al. (2017) utilize syntactic patterns to gather plausible candidate resolutions from a background corpus in order to extend the scarce training data.

3 Context reconstruction and model

In this section, we first propose a simple reconstruction trick which allows us to build minimal pairs of plausible and implausible argumentative texts. Then, we describe a Siamese neural sequence ranking model which addresses the task of ranking texts according to their plausibility.

Constructing plausible and implausible argumentative discourse contexts Consider two EAU clauses a_2 (source) and a_1 (target) where we need to decide whether a_2 *supports* a_1 or a_2 *attacks* a_1. In the absence of contextual discourse clues[4], a system must learn to predict this relation by considering the semantic content of a_1 and a_2. We approach this task by offering two alternative context reconstructions and asking our model in what context a_1 and a_2 are more likely to appear. More precisely, our reconstruction trick is as follows:

(a) a_1. Additionally, a_2 .

(b) a_1. Admittedly, a_2.

where **(a)** signals that two argumentative units likely stand in a *support*-relation and **(b)** signals the opposite ('attack'). In our experiments (Section 4), we also examine other possible discourse connectors for our reconstruction (e.g., moreover/however). From here, we ask our model which of the two reconstructions leads to a more plausible 'reading': **(a)** or **(b)**? E.g., consider the *cannabis*-example from Section 1; applying our reconstruction trick yields the following *implausible-plausible* minimal pair (r^-, r^+):

(3a) [$_{r^-}$ *Use of marijuana causes chronic bronchitis and airflow obstruction.* ***Additionally,*** *cannabis does not need to be smoked to receive its potential health benefits.*]

(3b) [$_{r^+}$ *Use of marijuana causes chronic bronchitis and airflow obstruction.* ***Admittedly,*** *cannabis does not need to be smoked to receive its potential health benefits.*]

Clearly, **(3b)** constitutes a more plausible reconstruction compared with **(3a)**. Exactly this is what we desire our model to learn: assessing the fine-grained differences between two texts which differ

---

[4]To name just one situation: consider a cross-document relation classification setup where a_1 stems from a different document than a_2. Any specific textual discourse context would not only be more or less unimportant, but also bears the potential to confuse the system.

score **score**

Plausibility Prediction

Reading Encoder

Use of marijuana causes chronic bronchitis (...).
Additionally, cannabis does not need to be smoked (...).

Use of marijuana causes chronic bronchitis (...).
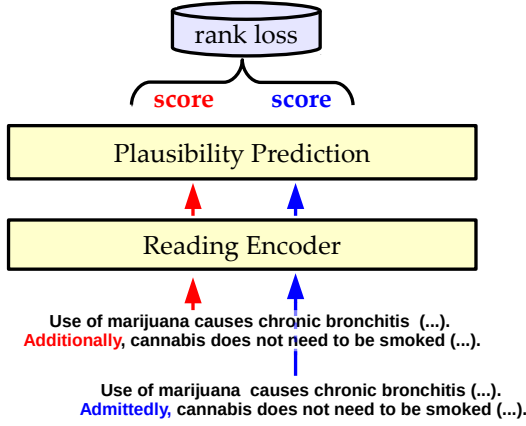Admittedly, cannabis does not need to be smoked (...).

Figure 2: Siamese model outline. Two competing reading-reconstructions are fed through a Siamese encoder (Reading Encoder). The vectors are then mapped by means of a (Siamese) linear combination and a selu-activation onto two corresponding plausibility scores (Plausibility Prediction). By reducing the ranking loss, we force the model to assign higher scores to more plausible readings.

in only one phrase. This phrase, however, determines whether the text in its entirety is *implausible* or *plausible*.

3.1 Loss and model

Ranking loss We argue that a ranking approach (*which reading is more plausible?*) is more suitable for addressing our problem compared with a classification approach (*plausible vs. implausible*). The reason is that ranking allows for a more relaxed and graded notion of textual plausibility: we want the model to *prefer* one variant and not to *choose* one variant. This is accomplished by reducing the margin ranking loss on the training data $\{(r_i^+, r_i^-)\}_{i=1}^n$:

$$\mathcal{L}_\theta = \frac{1}{n}\sum_{i=1}^n \left[1 - score_\theta(r_i^+) + score_\theta(r_i^-) \right], \quad (1)$$

where $score(\cdot)$ is a plausibility prediction model parameterized by θ. The plausibility-prediction model which we use is described in detail in the following paragraphs. Since $\mathcal{L}_\theta$ is differentiable with respect to the model's parameters θ, we can learn them with gradient descent.

Model overview We desire the $score(\cdot)$ function to return a number $p \in \mathbb{R}$ reflecting the plausibility

of a text sequence made up of words $w_1, ... w_n$. In our case, this function is instantiated with (i) a Siamese reading encoder (*Reading Encoder*, Figure 2) and a Siamese plausibility prediction layer for producing a plausibility score for any given text (*Plausibility Prediction*, Figure 2). Now, we will describe these two components more closely.

Reading encoder First, we use a contextual language model[5] to infer a sequence of word embeddings: $e_1, ..., e_n$, which correspond to words $w_1, ... w_n$. Here, we hope that already the contextual language model provides statistical information indicating whether a specific word sequence may be considered as rather plausible or rather implausible ('inductive bias'). The sequence of word embeddings $e_1, ..., e_n$ is further multiplied by a sequence of positive indicator coefficient embeddings: $e_1 \cdot c_1, ..., e_n \cdot c_n$.[6] This allows the model to learn to better distinguish between the source, target and the connector text (we learn three corresponding indicator embeddings). The resulting sequence is further processed by (ii) a Bi-LSTM (Hochreiter and Schmidhuber, 1997) to construct hidden states $H = h_1, ..., h_n$ (we concatenate hidden states of forward and backward read) and (iii) a four-headed scaled dot-product self-attention mechanism (Vaswani et al., 2017), where in our case we use $H = Q = K = V$:

$$Heads(Q, K, V) = [head_1; ...; head_4]W^O$$
$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$
$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V,$$

where $W_{(\cdot)}^{(\cdot)}$ are parameters of the model. Finally, we compute a weighted average of the final sequence of hidden states to construct a vectorized reading representation v (Felbo et al., 2017):

$$e_t = Heads(\cdot)_t W^A \qquad a_t = \frac{exp(e_t)}{\sum_{i=1}^T exp(e_i)}$$
$$v = \sum_{i=1}^T a_i Heads(\cdot)_i,$$

where $Heads(\cdot)_t$ is the vector corresponding to time step t computed by the previous scaled dot-

---

[5] We use BERT (Devlin et al., 2019) to infer the contextual embeddings. In our ablation experiments, we also present results based on ELMo embeddings (Peters et al., 2018)

[6] Similar to Opitz and Frank (2019a).

abbreviation	'support'	'attack'
A/A	Additionally,	Admittedly,
A/D	I agree,	I disagree,
M/H	Moreover,	However,
Y/N	Yes,	No,

Table 1: Argumentative discourse connector sentence adverbials and the argumentative relation class which they are likely to signal.

product attention step and v is a final vectorized representation of the input reading.

Plausibility prediction At plausibility prediction time, the vector representation v, which we obtained by the previous step, is mapped to a single score by means of a linear combination with a weight vector. Lastly, a selu-function (Klambauer et al., 2017) produces the desired plausibility-score:

$$p = selu(v^T w). \qquad (2)$$

This score, computed once for each of the two competing reconstructions, allows a comparison with respect to their (predicted) plausibility. For our ARC experiments, where we desire a final classification, we predict the argumentative relation class by inspecting the discourse connector of the reconstruction which obtains a higher plausibility score. E.g., if $score(EAU_1, additionally, EAU_2) \geq score(EAU_1, admittedly, EAU_2)$ we predict the argumentative 'support' relation – otherwise we predict the 'attack' relation.

4 Experiments

We begin this section by describing the experimental setup used to evaluate our neural plausibility ranker. Next, we present our main results and finally perform several analyses and study the effects of ablating model components.

4.1 Setup

Discourse links To construct plausible and implausible texts, we experiment with eight different discourse connectors which have the potential to 'signal' argumentative relation types. They make up, in total, four minimal pairs (Table 1).

Data We use the student essay corpus v02 (Stab and Gurevych, 2017) in two versions: ESSAY and ESSAY-CONTENT. What is common to both is that they contain data from the same 402 argumentative essays written by students about a variety of topics. The essays have been annotated with, i.a. spans of argumentative units and their relations with each other (support vs. attack). Since only the argumentative clauses have been annotated, we can clean EAUs from their discourse context, which yields ESSAY-CONTENT. For example, consider EAU_1. *To add on this, EAU_2*. While in ESSAY, a system is allowed to see EAU-surrounding tokens (*to add on this*), in ESSAY-CONTENT, systems are allowed to see only the spans of the EAUs to predict their relation (i.e., EAU_1, EAU_2). In the easy case, *to add on this* may be enough to predict a support relation with high confidence and accuracy without even seeing the content of the EAUs – in the hard case, however, a system must learn to assess the actual content of the premises. In ESSAY-CONTENT, the performance of the feature-based SVM described by Stab and Gurevych (2017) drops by more than 23% macro F1 compared to the standard setup (ESSAY) where shallow discourse context is accessible (Opitz and Frank, 2019b).

Baselines We display the results of a competitive feature based SVM. It requires, i.a., syntactic parsing, constituency-tree sentiment annotation (Socher et al., 2013) and discourse parsing (Lin et al., 2014) as pre-processing steps (Stab and Gurevych, 2017; Opitz and Frank, 2019b). In contrast, our method does not depend on any pre-processing.

Model instantiation For each possible minimal pair, we instantiate a different model based on the pre-trained BERT model (the BERT model remains fixed during optimization). More specifically, we infer the word embeddings and average over the last four layers to produce a sequence of vectors with 1024 dimensions. Forward and backward LSTM have 256 neurons each. For development purposes we split off 1149 examples from the training data. The rank loss (Eq. 1) is minimized by performing stochastic gradient descent with Adam (Kingma and Ba, 2014)[7]. After each epoch, the model is evaluated on the development data. Finally, we select the parameters from the epoch with maximum F1 score on the development data.

In our tables, each model is denoted *ArgRanker$_{dcs}$* where *dcs* indicates which pair of discourse connectors was used for reconstruc-

---

[7]The learning rate is set to 0.001, the mini-batch size to 64 and the maximum number of epochs to 25.

System	ESSAY	ESSAY-CONTENT
majority baseline	47.8	47.8
SVM with features	68.0	57.3
ArgRanker$_{A/A}$	67.2$^{\pm1.0}$	58.6$^{\pm1.4}$
ArgRanker$_{A/D}$	69.2$^{\pm2.4}$	**59.2**$^{\pm0.7}$
ArgRanker$_{M/H}$	68.8$^{\pm1.7}$	**63.8**$^{\pm2.1}$
ArgRanker$_{Y/N}$	67.3$^{\pm0.8}$	58.3$^{\pm1.8}$
ArgRanker$_{vote}$	**70.9**$^{\pm0.7}$	**60.7**$^{\pm1.7}$

Table 2: Macro F1 results. underlined: best result; **bold**: improves against SVM withstanding standard deviation.

	Attack		Support	
System	Precision	Recall	Precision	Recall
majority	0.00	0.00	8.0	100.0
SVM with features	20.0	22.9	93.0	91.8
ArgRanker$_{A/A}$	**31.0**$^{\pm5.9}$	18.2$^{\pm2.1}$	92.9$^{\pm0.1}$	**96.2**$^{\pm1.1}$
ArgRanker$_{A/D}$	**28.0**$^{\pm4.4}$	22.8$^{\pm3.2}$	93.1$^{\pm0.2}$	**94.5**$^{\pm1.6}$
ArgRanker$_{M/H}$	**39.9**$^{\pm9.4}$	**30.0**$^{\pm6.8}$	**93.8**$^{\pm0.6}$	**95.5**$^{\pm2.4}$
ArgRanker$_{Y/N}$	25.6$^{\pm5.8}$	23.7$^{\pm5.1}$	93.1$^{\pm0.3}$	93.0$^{\pm3.3}$
ArgRanker$_{vote}$	**51.4**$^{\pm7.3}$	17.7$^{\pm3.4}$	93.0$^{\pm0.2}$	**98.4**$^{\pm0.7}$

Table 3: Precision and recall scores for each class on ESSAY-CONTENT. underlined: best result; **bold**: improves against SVM withstanding standard deviation.

tion. *ArgRanker$_{vote}$* denotes a model where we aggregate the predictions over the four different minimal-pair single models ('ensemble model'). All results are averaged over five runs.

4.2 Results

Macro F1 results Table 2 lists the macro F1 results[8] of our experiments.

On ESSAY, our method is competitive with the SVM that relies on extensive pre-processing. On ESSAY-CONTENT, where models are forced to learn to assess the content of EAUs, our method outperforms the feature-based SVM across all configurations. The best performance on this data is provided by ArgRanker$_{M/H}$, which is trained on *Moreover-However* reconstructions (+6.5 pp. macro F1, relative improvement: 11%). Our ensemble model ArgRanker$_{vote}$, which aggregates the predictions of the individual ArgRankers in a simple vote, achieves an improvement of +3.4 pp. macro F1 (relative improvement: 6%).

More detailed results Table 3 indicates that our method offers other advantages besides raw macro F1 gains. The very rare *attack*-class is detected with a much greater precision compared with the SVM. The difference can range from an improvement of 5.6 pp. (ArgRanker$_{Y/N}$, relative improvement: 28%) up to a maximum improvement of 31.4 pp. (ArgRanker$_{vote}$, relative improvement: 157%). With such a large increase in precision, one might expect a drop in recall – however, this is only the case to a very small extent. The greatest drop in recall is incurred by ArgRanker$_{vote}$ (-5.2 pp.) and thus can be said to lie in the shadow of its precision gains (+31.4 pp.). Moreover, when
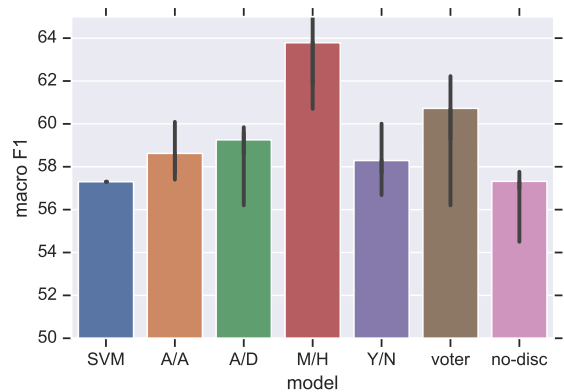

Figure 3: Scores for different models using BERT embeddings and SVM (left column) on ESSAY-CONTENT. Reconstruction with *Moreover/however* offers the largest improvement. Non-linguistically motivated connectors result in reduced performance ('+'/'-': *no-disc*, right column).

we use the discourse connector minimal pairs A/D and M/H, our model outperforms the SVM in the *attack*-class both in precision and recall. Most notably, when we instantiate our reconstructions with *Moreover/However*, we see a large gain in precision (+19.9 pp., relative improvement: 99.5%) but also an observable gain in recall (+7.1 pp., relative improvement: 31.0%).

With regard to the majority class (*support*), we make two observations: (i) precision-wise, all of our models outperform or are on par with the SVM; (ii) recall-wise, all of our models outperform the SVM. The greatest gain in recall for *support* is achieved by ArgRanker$_{vote}$ (+6.6 pp.).

4.3 Ablation experiments and analysis

Linguistically motivated discourse reconstruction What is the outcome of instantiating the dis-

---

[8]Macro F1 in our case is defined as the unweighted mean over the F1 scores for our two classes.

System	model configuration			
	basic	ELMo	-coeff.	-att.
majority	47.8	-	-	-
SVM (Stab and Gurevych, 2017; Opitz and Frank, 2019b)	57.3	-	-	-
ArgRanker$_{A/A}$	$58.6^{\pm1.4}$	$55.7^{\pm1.6}$	$57.4^{\pm2.1}$	$58.3^{\pm1.2}$
ArgRanker$_{A/D}$	$59.2^{\pm0.7}$	$60.2^{\pm2.2}$	$59.6^{\pm2.1}$	$56.2^{\pm1.9}$
ArgRanker$_{M/H}$	$63.8^{\pm2.1}$	$59.4^{\pm2.5}$	$61.1^{\pm1.0}$	$60.7^{\pm1.9}$
ArgRanker$_{Y/N}$	$58.3^{\pm1.8}$	$57.6^{\pm1.7}$	$57.7^{\pm1.2}$	$58.2^{\pm3.0}$
ArgRanker$_{vote}$	$60.7^{\pm1.7}$	$59.5^{\pm1.9}$	$60.2^{\pm2.2}$	56.2
ArgRanker$_{-discourse}$	$57.3^{\pm0.4}$	$63.6^{\pm1.3}$	$57.3^{\pm2.8}$	54.5

Table 4: Ablation experiments: Macro F1 results on ESSAY-CONTENT. *ArgRanker$_{-discourse}$*: a system where we replace the natural discourse connectors with 'linguistically meaningless' placeholders (i.e., support: '+', attack: '-' instead of, e.g., support: 'Moreover', attack: 'However'). *ELMo*: we use ELMo instead of BERT; *-coeff.*: we abstain from learning source-target specific coefficients; *-att.*: we ablate the self-attention and use the last states of the Bi-LSTM (concatenation of each read) for prediction.

course reconstructions with 'meaningless' connectors? I.e., instead of instantiating the attack/support context with linguistically motivated connectors, such as, e.g., *I agree/I disagree*, we instantiate the contexts with the meaningless tokens '+' and '-'. On one hand, this means that the new discourse configuration is still discriminative (either supporting or attacking). On the other hand, however, the discriminating reconstruction is not any more linguistically motivated. Thus, we hypothesize that the linguistically motivated reconstructions better 'trigger' the contextual BERT model into giving a useful inductive bias about whether a certain reading is plausible or not.

From Table 4 and Figure 3, we see that, indeed, our model functions better when provided with linguistically motivated reconstructions instead of the non-linguistically motivated reconstruction (Figure 3: columns *A/A, A/D, M/H, Y/N* vs. bottom row in Table 4 and right column in Figure 3). This holds true across all model configurations and all linguistically motivated discourse connector pairs.[9]

More specifically, we find that the *Moreover/However* reconstruction appears to offer the most useful inductive bias (middle column, Figure 3). Our *ArgRanker* based on this reconstruction outperforms all other configurations by more than 4 pp. macro F1 (compared with *Agree/Disagree*) and more than 6 pp. macro F1 compared with the

---

[9] An exception constitutes the model based on ELMo embeddings, which appears to work better when provided with the non-linguistically motivated connector pair.

non-linguistically motivated reconstruction. One reason could be located in the fact that BERT was trained, i.a., on the Wikipedia corpus: we compute a simple word frequency statistic over this corpus and see that the terms *Moreover* and *However* appear more frequently in this corpus (e.g., *however*: appr. 29,900,000 occurrences) than, e.g., *Admittedly* (appr. 17,000 occurrences). Also, by manually inspecting a small amount of occurrences in Wikipedia, we find that *moreover* and *however* tend to occur in more 'argumentative' contexts, or, at least, connect two discourse units in a contrasting (*however*) or supporting (*moreover*) way. On the other hand, e.g., *I agree* tends to occur in less argumentative contexts, such as in *I agree to the terms of service*. We believe that contextual language models trained on interactive discourse texts (e.g., online discussion platforms) instead of encyclopedic texts would greatly help to provide our model with better embeddings in the situations where we want to compose plausible and implausible texts by means of more 'interactive' connectors (*I agree/I disagree*; *Yes/No*; etc.).

BERT vs. ELMo In our first experiment, we replace the BERT embeddings with ELMo embeddings – we want to 'probe' which of the two embedding generators is better suited to rank argumentative texts according to their plausibility. First, we see that ELMo embeddings provide better performance than the feature based baseline, with one exception: ArgRanker$_{A/A}$, where we reconstruct contexts by inserting *Additionally* and *Admittedly*

(a) Target vs. source coefficients   (b) Connector vs. source coefficients   (c) Connector vs. target coefficients
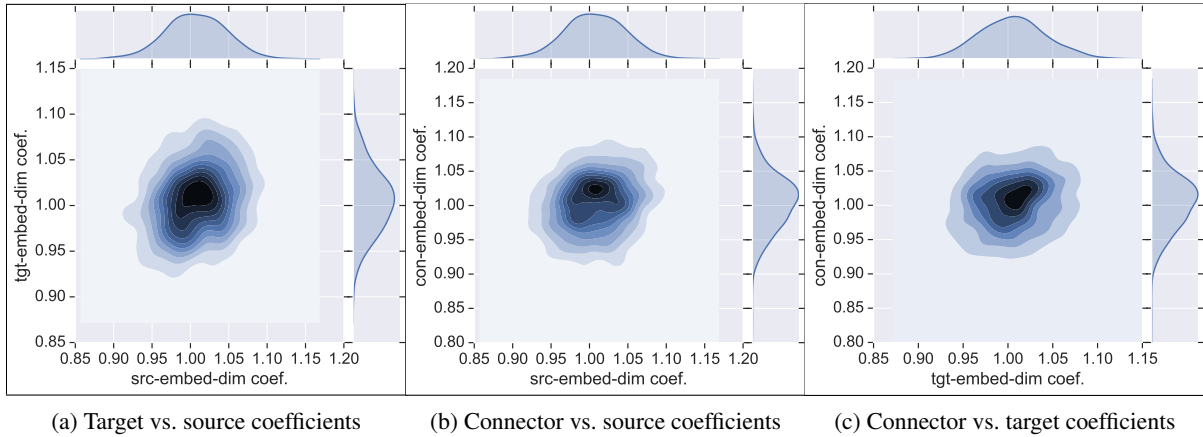
Figure 4: Investigation of the three contextual coefficient-embeddings which our model has learned. The coefficients are initialized with ones and assume, during training, a Gaussian-like distribution. By all appearances, the model uses some coefficients to 'deflate' the impact of an embedding dimension in, e.g., the text corresponding to the source EAU, while 'inflating' the impact of an embedding dimension in, e.g., the text corresponding to the target EAU (Figure 4a, regions on the upper left).

(Table 4, **ELMo**). Second, the ELMo embeddings in most cases fall short in comparison to BERT embeddings – again, however, with one exception: ArgRanker$_{-discourse}$, which does not use the linguistically motivated reconstruction.

Indicator embedding coefficients Now, we want to investigate if learning coefficients to better distinguish between source and target has helped our model. Recall, that the three coefficient indicator embeddings correspond to *source/target EAU span* and the *discourse connector span* and allow the model to highlight certain word embedding indices differently with respect to these three spans. For most connector pairs, learning the coefficients helps and their ablation leads to a performance drop (Table 4, **-coeff**; e.g. ArgRanker$_{M/H}$: -2.7 pp. macro F1).

Finally, we plot the learnt coefficients of the three different indicator embeddings against each other to analyze their appearance after training. Figure 4 displays all values from all discourse connector parameterizations $\cdot/\cdot$ of ArgRanker$_{\cdot/\cdot}$. More specifically, we are interested in the following question: *Have we learned that certain contextual word embedding indices are important to inflate (deflate) with respect to the source or the target?* From inspecting Figure 4, we see that this appears to be the case. For example, there is a set of embedding indices where coefficients are used to magnify the corresponding values in the target EAU and deflate them in the source EAU (Figure 4a, top left region) – while for another set of embedding indices

the opposite is true (Figure 4a, bottom right). Furthermore, the learnt coefficients have assumed a normal-like distribution after training (distribution plots on the sides of Figures 4a, 4b, 4c).

Self-attention Finally, we want to investigate the effect of ablating the self-attention mechanisms from our model. More precisely, we predict the plausibility scores based on a concatenation of the last state of forward and backward read of the Bi-LSTM. Throughout all different discourse reconstruction strategies, we see drops in performance (Table 4, **-att**). However, while we see observable drops in some cases (ArgRanker$_{vote}$: -4.5 pp. macro F1), they are comparatively small in other cases (ArgRanker$_{Y/N}$: -0.1 pp.).

5 Conclusion

We have treated argumentative relation classification in a new light, as a task where we learn to rank candidate texts according to their plausibility. To this aim, we have proposed a simple reconstruction trick which allows us to embed source and target argumentative units into plausible and implausible argumentative discourse contexts. In order to learn to rank such texts according to their plausibility, we have adapted a neural Siamese ranking model. Our experiments on an established data set have shown that the approach is competitive with previous work albeit it does not require pre-processing. In the 'content-based' setting – which is more difficult because models cannot base their decisions on shallow clues in the discourse context – the method

outperforms previous work by a considerable margin. In particular with respect to the scarce class *attack* we observed substantial improvements in precision.

Acknowledgments

References

Philippe Besnard and Anthony Hunter. 2014. Constructing argument graphs with deductive arguments: a tutorial. *Argument & Computation*, 5(1):5–30.

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. Towards argument mining from dialogue. In *COMMA*, pages 185–196.

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.

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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark, September. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. Self-normalizing neural networks. *CoRR*, abs/1706.02515.

Jonathan Kobbe, Juri Opitz, Maria Becker, Ioana Hulpus, Heiner Stuckenschmidt, and Anette Frank. 2019. Exploiting Background Knowledge for Argumentative Relation Classification. In Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski, editors, *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *OpenAccess Series in Informatics (OASIcs)*, pages 8:1–8:14, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2.

Anne Lauscher, Goran Glavaš, and Kai Eckert. 2018a. Arguminsci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 22–28.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018b. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium, November. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.

Ana Marasović, Leo Born, Juri Opitz, and Anette Frank. 2017. A mention-ranking model for abstract anaphora resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 221–232, Copenhagen, Denmark, September. Association for Computational Linguistics.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Juri Opitz and Anette Frank. 2018. Addressing the Winograd schema challenge as a sequence ranking task. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 41–52, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Juri Opitz and Anette Frank. 2019a. An argument-marker model for syntax-agnostic proto-role labeling. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 224–234, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Juri Opitz and Anette Frank. 2019b. Dissecting content and context in argumentative relation analysis. In *Proceedings of the 6th Workshop on Argument Mining*, pages 25–34, Florence, Italy, August. Association for Computational Linguistics.

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

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948.

Andreas Peldszus and Manfred Stede. 2016. An annotated corpus of argumentative microtexts. In D. Mohammed and M. Lewinski, editors, *Argumentation and Reasoned Action - Proc. of the 1st European Conference on Argumentation, Lisbon, 2015*. College Publications, London.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

John L. Pollock. 1995. *Cognitive Carpentry: A Blueprint for How to Build a Person*. MIT Press, Cambridge, MA, USA.

Maria Skeppstedt, Andreas Peldszus, and Manfred Stede. 2018. More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 155–163, Brussels, Belgium, November. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

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

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

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, pages 217–226.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al-Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3753–3765, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Douglas Walton. 2009. Objections, rebuttals and refutations. In *Argument Cultures: Proceedings of the 8th OSSA Conference*, pages 1–10, Windsor, Ontario.