Argument Relation Classification through Discourse Markers and Adversarial Training

Michele Luca Contalbo[1], Francesco Guerra[1], Matteo Paganelli[1]

[1]University of Modena and Reggio Emilia, Modena, Italy,
{micheleluca.contalbo, francesco.guerra, matteo.paganelli}@unimore.it

Abstract

Argument relation classification (ARC) identifies supportive, contrasting and neutral relations between argumentative units. The current approaches rely on transformer architectures which have proven to be more effective than traditional methods based on hand-crafted linguistic features. In this paper, we introduce DISARM, which advances the state of the art with a training procedure combining multi-task and adversarial learning strategies. By jointly solving the ARC and discourse marker detection tasks and aligning their embedding spaces into a unified latent space, DISARM outperforms the accuracy of existing approaches.

1 Introduction

Argument relation classification (ARC) is a crucial task in argument mining and aims to automatically identify relations between argumentative units to understand whether they support each other, are in opposition, or have no dependency (Toulmin, 2003; Lippi and Torroni, 2016; Stede and Schneider, 2018; Lawrence and Reed, 2019). It can be applied in various domains, such as political debates, legal and juridical cases, business negotiations. In these scenarios, ARC facilitates the understanding and evaluation of complex discussions by identifying logical connections and assessing their argumentative coherence and effectiveness.

In literature, ARC is typically conceived as a classification problem where pairs of argument units are categorized into predefined relation classes. For example, an ARC model is asked to recognize that the sentences in the first row of Table 1 support each other (i.e., *support* relation), the ones in the second row are in conflict (i.e., *attack* relation), and those in the last row have no dependency (i.e., *neutral* relation).

Traditional ARC approaches rely on the extraction of hand-crafted linguistic features, which derive from the identification of specific discourse elements, syntactic elements and lexical structures (Stab and Gurevych, 2014; Peldszus and Stede, 2015; Stab and Gurevych, 2017; Wachsmuth et al., 2018; Gemechu and Reed, 2019), the analysis of the topics these propositions refer to (Lawrence et al., 2014; Nguyen and Litman, 2016; Fromm et al., 2019), or a combination of the two (Lawrence and Reed, 2015). More recent approaches either address multiple argumentative tasks simultaneously through multi-task learning or integrate common-sense knowledge into the model. Examples of the former include the work by Galassi et al. (2021) and Liu et al. (2023). The first investigates the use of residual networks and neural attention mechanisms to simultaneously classify argument components and their relations. The latter addresses argument mining as a multi-hop reading comprehension task where the model is trained to perform classification and generate a reasoning sequence with transformer-based architectures (Vaswani et al., 2017). In terms of integrating common-sense knowledge, notable methods include ARK (Paul et al., 2020) and KE-RoBERTa (Saadat-Yazdi et al., 2023). ARK follows this idea by combining with a cross-attention layer pairs of sentence representations generated by distinct BiLSTM architectures. The representations are then enhanced via external knowledge coming from both ConceptNet (Speer and Havasi, 2012) and WordNet (Miller, 1995).

Table 1: Examples of argumentative units labeled as support, attack or neutral for the ARC task. The underlined words indicate discourse markers.

Sentence pair	Relation
Exercise reduces stress. *Thus, it's good for mental health.*	Support
Social media connects people easily. *However, it often spreads misinformation.*	Attack
The project deadline is approaching. *Meanwhile, the team is preparing a presentation.*	Neutral

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 18949–18954

KE-RoBERTa (Saadat-Yazdi et al., 2023), one of the current state-of-the-art approaches for ARC, dynamically injects common-sense knowledge in a RoBERTa-based model via a generative model called COMET (Hwang et al., 2021).

In this paper, we address the ARC task from a different perspective: instead of explicitly injecting external knowledge, we introduce a special training procedure for fine-tuning a transformer architecture on the ARC task. This is obtained by combining multi-task and adversarial learning strategies that drive the models to learn meaningful sentence representations supporting the ARC task. We implement this idea in DISARM[1] (*DIScourse markers and adversarial Argument Relation Mining*) which extends the standard fine-tuning of a RoBERTa (Liu et al., 2019) transformer architecture with two main improvements. The first is combining the classification of argument relations with discourse marker discovery (DMD). The second is applying an adversarial procedure to align the sentence representations across the two tasks into a single joint latent space. The intuition is that learning to identify discourse markers (the underlined words in Table 1) helps the model capture meaningful sentence representation properties that can be shared with the ARC task (Jernite et al., 2017; Nie et al., 2017; Malmi et al., 2018). DISARM exploits the data provided by the *Discovery* benchmark (Sileo et al., 2019), where pairs of sentences are labeled with the discourse marker connecting them. A pre-processing task is needed to reduce the 174 discourse markers available in the *Discovery* ground truth to the categories of *elaborative*, *inferential* and *contrastive* markers introduced in Fraser (1999). The experimental evaluation shows that DISARM outperforms competing approaches.

2 The DISARM approach

DISARM consists of two main components (see Figure 1), i.e., a RoBERTa-based encoder and a series of classification heads, which during training alternatively process data associated with ARC and DMD tasks.

2.1 Encoder

Consider a dataset for the target ARC task, i.e., $T(s_1^{\mathsf{ARC}}, s_2^{\mathsf{ARC}}, y^{\mathsf{ARC}})$, composed of argumentative units with associated relation categories, and an equal sized dataset extracted from *Discovery*, i.e., $S(s_1^{\mathsf{DMD}}, s_2^{\mathsf{DMD}}, y^{\mathsf{DMD}})$, made up of sentences and the category of the discourse marker that connects them. We format each input as

$$x^k = \texttt{<s>}\ s_1^k\ \texttt{</s><s>}\ s_2^k\ \texttt{</s>} \quad \text{with } k \in \{\mathsf{ARC}, \mathsf{DMD}\}$$

and feed it into a `roberta-base` encoder f_e. It generates a set of embeddings $h_l^k = f_e(x^k) = (h_{l,1}^k, ..., h_{l,n}^k)$ with n being the number of tokens inside x^k and l indicating the output of the last encoder block. Building on previous work (Jawahar et al., 2019), which show that shallow transformer blocks capture superficial linguistic features, whereas deeper ones encode more complex semantic information, we average the embeddings of the last layer h_l^k with those of the first h_i^k in order to capture both syntactic and semantic features.

$$h^k = avg([h_i^k, h_l^k]) \tag{1}$$

We apply cross-attention to further emphasize the comparison of the two sentences.

$$K = W_1 e_1(h^k),\ V = W_2 e_1(h^k),\ Q = W_3 e_2(h^k)$$

$$\tilde{h}^k = \mathtt{avg}(\mathtt{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V) \tag{2}$$

where e_j extracts the embeddings of the j-th sentence. Then, the resulting embeddings are averaged to yield the final sentence representation $\tilde{h}^k$.

2.2 Classification heads

Three classification heads process the encoder output:

- **Head**$^{\mathsf{ARC}}$, which classifies the samples of T in *support* / *attack* / *neutral*;
- **Head**$^{\mathsf{DMD}}$, which classifies the samples of S in *elaborative* / *inferential* / *contrastive*;
- **Head**$^{\mathtt{domain}}$ which classifies each embedding into its own original dataset (i.e., S or T).

While the first two heads encourage knowledge sharing between the ARC and DMD tasks, the third aligns the two embedding spaces. To implement the latter we exploit a **Gradient Reversal Layer** (GRL) (Ganin and Lempitsky, 2015). By multiplying the gradient by a negative scalar $-\lambda$ during backpropagation, GRL forces the model to learn invariant features between the ARC and DMD tasks, pushing the embeddings into a joint latent space.

[1]The code is available at https://github.com/softlab-unimore/disarm

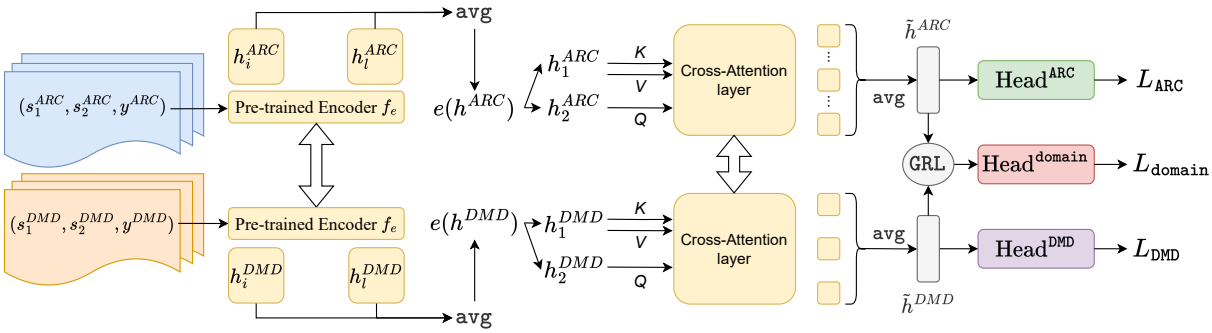

Figure 1: Overview of the DISARM architecture.

Table 2: Descriptive statistics for ARC and DMD data. The last column reports the frequency of *support*, *attack* and *neutral* for ARC datasets and that of *elaborative*, *inferential* and *contrastive* classes for *Discovery*.

	Task	Train	Dev	Test	Target freq (%)
SE	ARC	3,070	1,142	1,100	90/10/-
DB	ARC	6,486	2,163	2,162	50/50/-
M-ARG	ARC	3,283	410	411	9/3/88
Discovery	DMD	1.56M	87K	87K	32/29/39

2.3 Loss function

The model uses **cross-entropy** to calculate the losses $L_{\texttt{ARC}}$, $L_{\texttt{DMD}}$, $L_{\texttt{domain}}$ of the ARC, DMD and domain classifiers respectively. The total loss is given by their scaled sum:

$$L = L_{\texttt{ARC}} + \beta L_{\texttt{DMD}} + \gamma L_{\texttt{domain}} \quad \beta, \gamma \in [0, 1] \qquad (3)$$

3 Experimental evaluation

3.1 Datasets and Competing Approaches

We consider three datasets, typically used in the literature to evaluate the ARC task: *Student Essay* (SE) (Stab and Gurevych, 2017), *Debatepedia* (DB) (Paul et al., 2020) and *M-ARG* (Mestre et al., 2021). Table 2 reports some descriptive statistics. While in SE and DB the sentence pairs are labeled with two classes (i.e., *support* or *attack*), in M-ARG there are 3 classes (i.e., *support*, *attack*, and *neutral*). In general, the *support* class is the most frequent one and the DB dataset is twice the size of the other benchmarks. We selected two state-of-the-art approaches as representative competitors for DISARM: ARK (Paul et al., 2020) and KE-RoBERTa (Saadat-Yazdi et al., 2023).

3.2 Experimental setup and execution

DISARM was fine-tuned on the ARC benchmarks for 30 epochs, with a batch size of 64, AdamW optimizer with weight decay $1e-2$, learning rate $1e-5$ and $\lambda = 1e-2$. To manage class imbalance we used a class weighting of $1:10$, $1:1$ and $9.375:30:1$ for the two and three classes of SE, DB and M-ARG respectively. The results are averaged across six different runs. We performed a grid search in the interval $[0, 1]$ with a step size of 0.2 on the validation set to determine the optimal weights γ and β for the domain adaptation and discourse marker detection losses.

3.3 Effectiveness

The analysis of the results shown in Table 3 allows us to answer two main questions, concerning: (1) the effectiveness of our approach compared to state-of-the-art methods, and (2) whether the good performance derives from the knowledge of discourse markers or our special training.

Comparison with state-of-the-art. The last row of Table 3 shows that DISARM outperforms the reference approaches listed in the first two rows. The last column in the same table shows that the average improvement in F1 score across the three datasets compared to KE-RoBERTa is 1.22.

Discourse markers vs special training. An ablation study allowed us to understand the reasons for such an improvement. RoBERTa+ is a simplified version of DISARM obtained by removing both multi-task and adversarial learning processes (i.e., both $Head^{\texttt{DMD}}$ and $Head^{\texttt{domain}}$ are removed). Table 3 shows that RoBERTa+ achieves surprisingly good results on two of the three datasets tested (a similar conclusion was obtained in Ruiz-Dolz et al., 2021). The results are close to KE-RoBERTa that, we recall, relies on external common-sense knowl-

Table 3: Accuracy (F1 score). **Bold** values indicate the best results, underlined values the second-best ones. Δ values indicate the average accuracy difference wrt KE-RoBERTa. The results for ARK and KE-RoBERTa are taken from Saadat-Yazdi et al. (2023). Values for RoBERTa+, RoBERTa+ INJ, DISARM (MTL) and DISARM represent the average of 6 seeds, with standard deviations in brackets.

	SE	DB	M-ARG	Δ
ARK	60.00	64.00	-	
KE-RoBERTa	70.00	75.00	49.00	
RoBERTa+	65.15 (2.1)	74.7 (0.6)	50.37 (3.7)	-1.26
RoBERTa+ INJ	65.83 (1.7)	74.97 (0.8)	49.35 (2.7)	-1.28
DISARM (MTL)	69.74 (1.8)	76.14 (0.7)	50.88 (2.5)	+0.92
DISARM	**70.1** (1.6)	**76.22** (0.7)	**51.34** (3.2)	**+1.22**

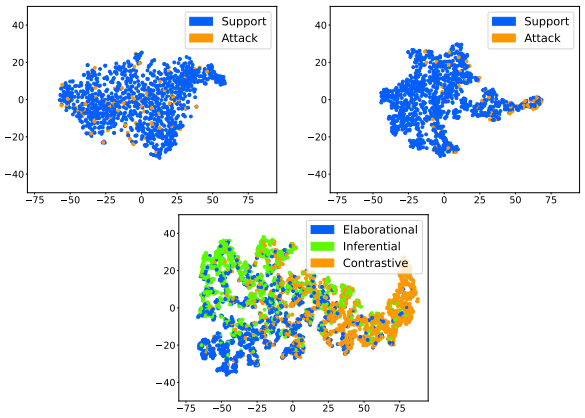

Figure 2: Impact of adversarial training on embedding space. Upper plots show the t-SNE projection of the embedding space produced by RoBERTa+ (top left) and DISARM (top right) on SE. The bottom plot shows the embeddings produced by DISARM on *Discovery*. DISARM aligns specific classes from ARC and DMD closer together in the embedding space, such as the *attack* class and the *contrastive* class.

edge. The marked improvement of DISARM over KE-RoBERTa and RoBERTa+ highlights the fundamental role of multi-task and adversarial learning in enhancing performance.

Building on this research, we performed a second study to understand whether the key contribution comes from utilizing discourse markers or from the specific training that encourages knowledge sharing between the two tasks. To investigate this, we implemented RoBERTa+ & INJ, a variant of RoBERTa+ that explicitly injects, into the input, the discourse markers predicted from another `roberta-base` model trained exclusively on Discovery. The results show a decrease of around 1% compared to KE-RoBERTa and let us conclude that injecting discourse markers into the input text can carry superficial knowledge distracting the model from the content of the analyzed propositions, as reported also in Opitz and Frank (2019). However, we observe that DISARM (MTL) achieves a significant improvement of about 1% on average compared to KE-RoBERTa when trained in a multi-task setting without adversarial learning (i.e., including only $Head^{\text{DMD}}$ and $Head^{\text{ARC}}$). This result suggests that the method used for injecting this knowledge into the model significantly impacts performance. Therefore, we conclude that the preferable approach is to inject this knowledge by solving the DMD task rather than explicitly inserting it into the input. Further refining this conclusion, we find that an even more effective method is to combine multi-task learning with adversarial learning. This combined approach used by DISARM fosters deeper knowledge sharing, directly enhancing the expressiveness of the embedding space generated by the encoder component.

3.4 Impact of adversarial training on the embedding space

We assess the impact of adversarial learning on the embedding space by (1) observing how sentence representations change when the model is trained with and without this technique, and (2) analyzing whether the embedding spaces of the ARC and DMD tasks align into a joint latent space. Regarding the first point, we extract the sentence embeddings generated by both DISARM and RoBERTa+ on the SE dataset (see the upper plots in Figure 2). As expected, the adversarial training produces discriminative sentence embeddings that are clearly separated into the *support* and *attack* classes. Regarding the second point, we compare the embeddings generated by DISARM on both SE and *Discovery* datasets (see the upper right and bottom plots in the Figure). We observe that sentence embeddings associated with the *attack* class in the SE dataset and those related to sentences containing *constrastive* discourse markers in the *Discovery* dataset are mapped into the same space. This demonstrates that adversarial training generates discriminative sentence embeddings by aligning the embedding spaces of the ARC and DMD tasks.

4 Conclusion

We presented DISARM, an argument relation classifier that injects knowledge of discourse markers into a pre-trained RoBERTa model via multi-task and adversarial learning. The experimental evaluation shows that this model outperforms previous state-of-the-art methods and learns discriminative sentence embeddings supporting the task.

Limitations

The experiments do not report cross-domain evaluations where the proposed model is trained on training data from a given domain and tested on a different domain. Therefore, the robustness of the model on out-of-domain data has not been fully evaluated. Furthermore, we observe higher standard deviation on smaller datasets (e.g. M-ARG). Such reduced dimensionality makes training more unstable (as also discussed in Devlin et al., 2019).

The proposed approach solves an argument relation classification task by integrating the knowledge of discourse markers that are extracted from the Discovery dataset. Given the low dimensionality of the ARC datasets, we integrated a subset of the data from Discovery. This allows us to better align the two tasks, avoiding domain balancing problems. Therefore DISARM does not make extensive use of Discovery data. We plan to address these limitations in future works.

Risks

The primary risk with using DISARM is the potential for misuse of the prediction model. Users could leverage DISARM to highlight misleading patterns in the relation between argumentative units, exploiting claims to further their own views.
Another issue is data scarcity and underrepresentation of certain social groups and languages. These biases could be amplified during model training, leading to distorted predictions. As a prototype, DISARM should be integrated into a broader framework that includes other argument mining tasks and systems to mitigate harmful predictions.
Finally, it is unclear how specific discourse cues impact the model's performance. In this paper, we have shown how discourse markers can be leveraged for the ARC task. Yet, we did not investigate the ambiguity of these markers and their potential for adversarial attacks. Addressing these gaps is a goal for future research.

Use of AI assistants

In the process of writing this paper, we used AI assistants to help in translating text from other languages to English, as well as generating initial drafts for some of the paragraphs. The AI-generated content was used exclusively as a starting point, with significant additional work done by the authors. Finally, during the development of DISARM, we used AI assistants to aid in the debugging of our code.

Acknowledgments

This work was partially funded by the RESIST0 project (PR-FESR Emilia-Romagna 2021-2027) through a grant to the AIRI research center at the University of Modena and Reggio Emilia.

References

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. Association for Computational Linguistics.

Bruce Fraser. 1999. What are discourse markers? *Journal of Pragmatics*, 31(7):931–952. Pragmatics: The Loaded Discipline?

Michael Fromm, Evgeniy Faerman, and Thomas Seidl. 2019. TACAM: topic and context aware argument mining. In *WI*, pages 99–106. ACM.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2021. Multi-task attentive residual networks for argument mining. *CoRR*, abs/2102.12227.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1180–1189. PMLR.

Debela Gemechu and Chris Reed. 2019. Decompositional argument mining: A general purpose approach for argument graph construction. In *ACL (1)*, pages 516–526. Association for Computational Linguistics.

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. In *AAAI*, pages 6384–6392. AAAI Press.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657. Association for Computational Linguistics.

Yacine Jernite, Samuel R. Bowman, and David A. Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *CoRR*, abs/1705.00557.

John Lawrence and Chris Reed. 2015. Combining argument mining techniques. In *ArgMining@HLT-NAACL*, pages 127–136. The Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Comput. Linguistics*, 45(4):765–818.

John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. 2014. Mining arguments from 19th century philosophical texts using topic based modelling. In *ArgMining@ACL*, pages 79–87. The Association for Computer Linguistics.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.*, 16(2):10:1–10:25.

Boyang Liu, Viktor Schlegel, Riza Batista-Navarro, and Sophia Ananiadou. 2023. Argument mining as a multi-hop generative machine reading comprehension task. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10846–10858. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Eric Malmi, Daniele Pighin, Sebastian Krause, and Mikhail Kozhevnikov. 2018. Automatic prediction of discourse connectives. In *LREC*. European Language Resources Association (ELRA).

Rafael Mestre, Razvan Milicin, Stuart Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *ArgMining@EMNLP*, pages 78–88. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Huy Nguyen and Diane J. Litman. 2016. Context-aware argumentative relation mining. In *ACL (1)*. The Association for Computer Linguistics.

Allen Nie, Erin D. Bennett, and Noah D. Goodman. 2017. Dissent: Sentence representation learning from explicit discourse relations. *CoRR*, abs/1710.04334.

Juri Opitz and Anette Frank. 2019. Dissecting content and context in argumentative relation analysis. In *ArgMining@ACL*, pages 25–34. Association for Computational Linguistics.

Debjit Paul, Juri Opitz, Maria Becker, Jonathan Kobbe, Graeme Hirst, and Anette Frank. 2020. Argumentative relation classification with background knowledge. In *COMMA*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 319–330. IOS Press.

Andreas Peldszus and Manfred Stede. 2015. Towards detecting counter-considerations in text. In *ArgMining@HLT-NAACL*, pages 104–109. The Association for Computational Linguistics.

Ramon Ruiz-Dolz, José Alemany, Stella Heras Barberá, and Ana García-Fornes. 2021. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intell. Syst.*, 36(6):62–70.

Ameer Saadat-Yazdi, Jeff Z. Pan, and Nadin Kökciyan. 2023. Uncovering implicit inferences for improved relational argument mining. In *EACL*, pages 2476–2487. Association for Computational Linguistics.

Damien Sileo, Tim Van de Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. In *NAACL-HLT (1)*, pages 3477–3486. Association for Computational Linguistics.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686. European Language Resources Association (ELRA).

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *EMNLP*, pages 46–56. ACL.

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

Manfred Stede and Jodi Schneider. 2018. *Argumentation Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Stephen E. Toulmin. 2003. *The Uses of Argument*, 2 edition. Cambridge University Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *ACL (1)*, pages 241–251. Association for Computational Linguistics.