Argument and counter-argument generation: a critical survey

Xiaoou Wang[1], Elena Cabrio[1], and Serena Villata[1]

Université Côte d'Azur, CNRS, Inria, I3S, France
{xiaoou.wang,elena.cabrio,serena.villata}@univ-cotedazur.fr

Abstract. Argument Generation (AG) is becoming an increasingly active research topic in Natural Language Processing (NLP), and a large variety of terms has been used to highlight different aspects and methods of AG such as argument construction, argument retrieval, argument synthesis and argument summarization, producing a vast literature. This article aims to draw a comprehensive picture of the literature concerning argument generation and counter-argument generation (CAG). Despite the increasing interest on this topic, no attempt has been made yet to critically review the diverse and rich literature in AG and CAG. By confronting works from the relevant subareas of NLP, we provide a holistic vision that is essential for future works aiming to produce understandable, convincing and ethically sound arguments and counter-arguments.

Keywords: Argument generation · Counter-argument generation · Argument retrieval · Argument mining

1 Introduction

Argument Mining (AM) is a research area which aims at identifying and classifying argumentative structures from text. The increasing interest in the literature for this area, due to its applications in tacking substantial societal challenges as propaganda detection, fact checking and *explainable Artificial Intelligence*, resulted in the publication of several surveys [13, 30]. The research area of AM has now been expanded to the generation of natural language arguments. To this date, *Argument Generation (AG)* is still considered as a hard task and no standard methods exist. To the best of our knowledge, no survey has been published on this subject. A recent paper by Lauscher *et al.* [29] discussed the role of knowledge in the general context of argumentation including AM, argument assessment, argument reasoning and AG, without focusing on the state of the art of this latter domain as well as the main trends and challenges faced by most researchers. However, researches in AG are clearly on the rise and a huge variety of methods have been explored. Also, multiple research directions have been sketched, from the perspective of generating argumentative components (e.g., claim, premise and enthymeme) as well as the employment of rhetorical strategies [45] and users' beliefs [1] to guide the argument generation. Applications

of AG are diverse and numerous, of which the most relevant are *writing assistance* [48], *legal decision making* [6], *collective decision making* [12] and *Counter Narrative Generation* [43] to fight online hate speech.

In this paper, we aim to lay out a comprehensive picture of the studies on AG and *Counter-argument Generation (CAG)*, where counter-arguments are in essence arguments against other arguments. Due to the large variety of topics and research communities covered by AG and CAG, studies published in these two fields often fail to cite each other. It is important to underline that AG is a complex task including multiple subtasks and it is essential to have a holistic view of the ongoing works in all the relevant subareas in order to design reliable end-to-end argumentative systems. With the idea of federating relevant communities in mind, we propose the current survey with the following contributions:

1. We draw a historical view of the development of studies in AG and CAG, providing a detailed outline of the main results and trends in various subareas of AG and CAG, along with a summary of the main datasets for these tasks.
2. We discuss the main issues and some open challenges in AG and CAG.
3. We point out 4 most promising research directions in AG and CAG.

2 Data to text argument generation

Studies on argument generation started around 1990s in the spirit of recommender systems. Considerable research has been devoted to developing computational models for automatically generating and presenting *evaluative arguments*. The general idea of these studies was to design computer systems serving as advisors to support humans in similar communicative settings. These studies were mainly concerned with producing short texts from structured data such as knowledge graphs representing domain knowledge as well as users. We call this family of approaches *Data to Text Generation*.

The general principles of data to text generation were formalized by Carenini [14] and applied to their *Generator of Evaluative Arguments* recommending houses to a client. The generation process first involves a deep generation phase which is agnostic of the target language since it consists mainly in the selection of knowledge chunks based on the comparison of a *User Model* and a *Domain Model* (e.g., the profile of a buyer and the profile of a house), and the selection of the argumentative strategy. The second phase, *content realization*, involves the actual text generation requiring specific grammatical knowledge of the target language such as verbal inflections and logical connectors.

Data to text generation is cumbersome since it involves a lot of manual work to build the knowledge bases and the actual knowledge acquisition process has to be restarted whenever a new domain is being tackled. Around the beginning of the 2010s, a shift took place in argument generation: first, the design of debating systems started to draw the attention of researchers (a prominent event was the Project Debater[1] of IBM, started in 2012); secondly, inspired by techniques in *Natural Language Generation (NLG)*, researchers adopt a *Text to Text*

[1] https://www.research.ibm.com/artificial-intelligence/project-debater/

Generation approach, which can either be further divided into several subtasks or generate full arguments in an end-to-end fashion.

3 Text to text argument generation

This section provides a complete outline of the main trends in AG and CAG using the text to text approach, summarized in Table 1.

3.1 Generation of argument components

A claim forms the basis of an argument, being the assertion that the argument aims to establish. Therefore, claim construction may be viewed as a first step in argument generation. It should be noted that *Claim Generation (CG)* is different from *Claim Retrieval* which consists in employing argument mining to identify existing claims in a corpus. To retrieve arguments, Levy et al. [32] have developed the task of *Context-dependent Claim Detection* whose objective is to identify supporting and attacking claims related to a topic from a Wikipedia corpus. The tasks of *Evidence Retrieval* [38] and *Claim Stance Classification* [7] are also related topics for the AG task. The goal here is to retrieve pro and con arguments for a given query. In the following sections, however, we will focus mainly on claim generation.

In its simplest form, *Claim Generation* takes a debate topic as input and the output is a concise assertion with a clear stance on this topic. To automatically generate new claims, Bilu and Slonim [10] used traditional linguistic features for predicting the suitability of candidate claims. Concretely, the authors drew insights from the fact that a predicate on a certain topic can be used to other topics under certain constraints. For instance, the predicate "is a violation of free speech" can be applied both to "banning violent video games" and "Internet censorship". Their framework employs two stages: first, given a topic, word2vec embeddings are used to select top k similar predicates from a Predicate Lexicon; second, the top-k predicates are combined with new topics and a logistic regression classifier is used to predict if the new claim is valid or not, using features such as n-grams. Gretz *et al.* [21] expanded this framework by leveraging GPT-2 to generate claims on topics and showed the potential of large language models in this task. Furthermore, Alshomary *et al.* [1] studied how to encode specific beliefs into generated claims.

Contrastive Claim Generation (CAG) is motivated by the observation that negation has an important function in argumentation. Bilu *et al.* [9] proposed a rule-based system to augment a set of claims by automatically suggesting a meaningful negation, which means that an opposite claim must be grammatically correct, semantically clear and logically valid. The authors concluded from this study that explicit negation is not always possible. To better tackle this issue, Hidey and McKeown [24] used a sequence to sequence model to encode the original claim with an attention mechanism. They used a sequence of words and a sequence of edits as encoder representations and found that the latter is more

effective. Another line of research, initiated by Chen *et al.* [16], is related to CAG. The authors proposed to use autoencoders for the task of *Bias Flipping* (i.e., switch the left or right bias of an article). An encoder conditioned on the source bias is used to encode the input text, while a decoder conditioned on the target bias decodes the encoder representation into a new text.

Bar-Haim *et al.* [7] introduced the task of *Premise Target Identification (PTI)* which identifies the target in a premise. Based on this task, Alshomary *et al.* [5] initiated the task of *Conclusion Target Inference (CTI)*, inspired by the observation that conclusions are not often explicitly formulated. They used a BIO sequence labeling to detect the boundary of the target of premises, then a ranking model [47] to select the premise target that is the most representative of the conclusion target. The authors also explored the use of a triplet neural network to select the most similar conclusion target to a premise target from a knowledge base containing all the conclusion targets. A hybrid approach, however, yielded the best results.

The last subtask of AG is called *Enthymeme Reconstruction (ER)*, where an enthymeme is an implicit premise that clarifies how a conclusion is inferred from the given premises. Boltužić and Šnajder [11] studied how to identify such enthymemes given the other components. Similarly, Habernal *et al.* [23] present the task of identifying the correct enthymeme from two options. More recently, a large dataset [15] studying abductive reasoning in narrative text was created to enable the use of neural models in this line of research.

3.2 Generation of full arguments

Rule-based argument generation. Sato *et al.* [40] presented the first end-to-end rule-based retrieval system to generate argument scripts in the first round of a debate. A user first selects a motion and a stance which agrees or disagrees with the motion. A *Motion Analysis component* then extracts the target of the motion and its stance. The *Value Selection component* selects the 5 most relevant talking points. Then the *Sentence Retrieval Component* retrieves sentences relevant to each value from the corpus, and finally, the *Sentence Rephrasing component* arranges the retrieved sentences to build the final argument.

Summarization-based approach. Due to the complexity to maintain the components in systems like [40], some studies proposed to use a neural summarization approach. Instead of producing single-sided arguments, summarization-based approaches also generate arguments representing both stances, which is particularly useful for controversial topics. From the perspective of argument generation, Alshomary *et al.* [2] argued that the objective of argument summarization is to extract snippets containing the main claim and the supporting reason of an argument. This task is called *Argument Snippet Generation (ASG)*. The authors addressed two goals of ASG: representativeness based on how much the core information of an argument is kept, and argumentativeness. They modified the LexRank algorithm [19] to account for the representativeness and improved argumentativeness of sentences by using lexicons of discourse and

claim markers. One limitation of this approach is redundancy: since the summarization is based on top-ranked arguments retrieved by an argument search engine, there is no guarantee that the snippets represent different aspects. To tackle this redundancy issue, Alshomary *et al.* [3] adapted an approach from comparative summarization which was designed to answer questions like "What is different between the coverage in NYTimes and BBC?". The authors defined an argument snippet as contrastive if it highlights the uniqueness of an input argument compared to other arguments returned by an argument search engine. They extended the graph-based approach of [2], which ranks sentences based on their centrality and argumentativeness, by encoding an extra term to account for the sentence's similarity to other arguments. Their results showed a clear improvement, with a tradeoff between representativeness and contrastiveness.

Other research directions in AG. One of the emerging research areas in full argument generation is *Audience-oriented Argument Generation*. Alshomary *et al.* [1] implemented audience-based features in AG to enhance the persuasiveness of the generated arguments. They trained a BERT-based classifier to identify morals such as care, fairness and loyalty in arguments and used the Project Debater's API to generate arguments based on morals on 6 topics. In *Rhetoric-based Argument Generation*, Wachsmuth *et al.* [45] created a benchmark dataset with manually synthesized arguments that follow rhetorical strategies, containing 260 argumentative texts on 10 topic-stance pairs. Based on this dataset, EI baff *et al.* [18] proposed a computational model to generate arguments according to a specific rhetorical strategy (Logos vs. Pathos) by imitating the process of selecting, arranging, and phrasing *Argumentative Discourse Units* (ADUs). Concretely, their approach viewed AG as a Language Modeling Task by considering ADUs as words and arguments as sentences. They first identified different ADU types using clustering then learned to select unit types matching the given strategy. The selected units are then arranged according to their argumentative roles (Thesis, Con, Pro). Finally, the argument is phrased by predicting the best set of semantically related ADUs for the arranged structure using supervised regression. Finally, the dialogue aspect of AG is getting more and more attention from researchers. Graph-based [36], rule-based [20] and retrieval-based neural generative systems [31] have all been explored, with more or less success and very different metrics for evaluation.

Counter-argument generation. Besides Sato *et al.* [40]'s value-based AG system, Wachsmuth *et al.* [46] designed another rule-based system to retrieve a counter-argument by identifying opposing conclusions to a given claim in the debate pool *idebate.org*. Concretely, their system detects similar conclusions with dissimilar premises and consider such arguments as counter-arguments. However, neural CAG is by far the most investigated approach because of the inherent overhead of maintaining rule-based systems. Hua and Wang [26] tackled this task in two steps: evidence retrieval and text generation. The authors first retrieved relevant Wikipedia articles using sentences in the original argument and then re-ranked the articles' paragraphs using TF-IDF similarity to the argument. The top-ranked sentences, concatenated with the input argument, were

encoded and fed to the decoder producing first some keyphrases, then the counterarguments per se by attending to the keyphrases at the same time. Hua *et al.* [25]'s model further improves the previous method by extracting (instead of generating) keyphrases from the input statement. Also, it ranks evidence passages by their keyphrase overlap with the input statement and also their sentiment toward the input statement to encourage counter-evidence. More recently, Alshomary *et al.* [4] proposed to attack an argument by challenging the validity of one of its premises on the CMV dataset [28]. Concretely, the task is divided into two subtasks: Weak-Premise Ranking using the learn-to-rank framework [35] and Premise Attack Generation. For the generation part, they used OpenAI's GPT [37] as a pretrained language model and a joint-learning approach combining next-token prediction and counter-argument classification (given two concatenated segments, decide whether the second is a counter-argument to the first). Their approach did not outperform the baseline of [27], however, a manual evaluation in terms of content richness, correctness and grammaticality showed that their approach yielded better results.

Table 1. Datasets in AG and CAG classified by subareas.

Task	Datasets	Source	Size
CG	[22]	Crowd annotation	30k arguments, 71 topics
	[38]	Wikipedia articles	2.3k claims, 58 topics
Belief-based CG	[1]	debate.org	51k claims, 27k topics
CCG	[24]	Reddit	1,083,520 pairs of contrastive claims
Bias Flipping	[16]	Biased headlines from allsides.com	6458 claim-like headlines
CG or PTI	[7]	Wikipedia articles	2,394 claims, 55 topics
CTI	[47]	idebate.org	2,259 arguments, 676 topics
ER	[23]	Comments section of the New York Times	2k arguments with two enthymemes of which one is correct
	[8]	Extended from a collection of five sentence stories	7,2k argument-hypothesis pairs
ASG	[2]	args.me	83 arguments along with two-sentence snippets
AG and CAG	[45]	Written by experts based on pools of ADUs representing pros and cons	130 logos-oriented and 130 pathos-oriented arguments, 10 topics
	[26]	Change My View (CMV) channel of Reddit	26,525 arguments, 305,475 counter-arguments
	[4]	CMV	111.9k triples of argument, weak premise and counter-argument

4 Challenges and open research directions

Despite the rich literature produced in AG and CAG, these two fields are still rapidly evolving and many challenges remain to be addressed. In this section, we highlight some of the most important challenges faced in AG and CAG.

Evaluation. Most automatic evaluation metrics used in CG and CAG are some commonly adopted metrics in machine translation and summarization such as BLEU, ROUGE and METEOR. Although automatic metrics are necessary for large-scale evaluation, the above-mentioned metrics are not specifically designed for argumentation and do not capture the essential qualities of an argument such as *cogency* (when an argument contains acceptable premises that are relevant to the argument's conclusion and that are sufficient to draw the conclusion), and *reasonableness* (when an argument contributes to the resolution of the given issue in a sufficient way that is acceptable to the target audience) [44]. In [26, 25], despite some encouraging results using BLEU and ROUGE, for both studies, human evaluation shows that the quality of fully-generated counterarguments is yet lower than that of a simple concatenation of evidence passages in terms of topical relevance and counterness. In fact, the simple criteria of understandability of an argument is far from being reached. In [16], out of 200 generated headlines, only 73 were understandable. The rule-based system of Sato *et al.* [40] has the same drawback (50 out of 86 sentences are judged as non-understandable). In addition, Chen *et al.* [16] found that for a successful flipping (CAG), the overlapping of generated and ground-truth headlines is very low, making overlap-based metrics unreliable. As for manual evaluation, a huge variety of author-dependent metrics is defined in the literature, making the cross-study comparability difficult. Studies concerning automatic argument quality evaluation [41, 34] are arising and should be integrated to AG and CAG.

Argumentation strategies other than rhetoric. Argumentation strategies such as hypothetical reasoning, reasoning by cases, premise-to-goal arguments such as inference to the best explanation [33] have been the focus of the earliest studies in AG [49]. Current studies are mainly focused on the computational aspects and concentrate less on these aspects, which are however important to produce convincing arguments according to different audience, in the same vein of the modeling of users' beliefs in AG and CAG systems.

Other challenges. Although the main goal of argumentation is to convince instead of proving the truthfulness of a thesis, the truthfulness issue must be considered to fight against online disinformation. In the case of retrieval-based systems, the reliability of the retrieved claims and evidences must be checked. Recent studies have started investigating the automatic evaluation of fairness in argument retrieval [17], the automated fact-checking of claims [39] and the automatic detection of insufficiently supported arguments [42]. These dimensions are particularly relevant to prevent the spread of disinformation, especially in view of the increasing use of large language models such as GPT which, trained on datasets such as CMV [28], are prone to inject bias and unreliable information in the generated texts. Last but not least, when the argument is not sufficiently elaborated, clarification questions should be triggered to request additional in-

formation to have a meaningful dialogue with the end user. This line of research has already been introduced in question answering, but a deeper investigation is required in AG and CAG.

Notes on ethical issues. As for many other NLP methods, AG has the potential of being misused, as it allows to automatically generate a variety of potentially false assertions regarding a topic of interest. Also, as discussed above, current methods in AG and CAG inherit the biases and truthfulness issues of the underlying language models. While ethical issues must be considered when AG systems are deployed at a large scale, two points are worth noting: *i)* the main objective of AG and CAG is to generate coherent and understandable (counter-)arguments based on a given input, which still remains the biggest challenge to be resolved; *ii)* AG and CAG systems allow for arguments to be generated on both stances towards a topic, thus if one side on a topic is misrepresented, it would be easily uncovered and this can contribute to the discovery of the inherent bias pertaining to large generative language models.

5 Conclusions

Studies on AG and and CAG are clearly on the rise, with multiple subareas and research directions. In this work, we draw a comprehensive outline of the subareas of AG and CAG as well as the biggest challenges in these two research fields. Our comparative examination of the existing literature highlights four promising lines of future research: *i)* the integration of users' beliefs and preferences in AG, which is reminiscent of early studies on AG in recommender systems where the user's profile play a role; *ii)* the development of intelligent argument dialogue systems, since arguments must be exchanged in a continuous fashion to reach a consensus; *iii)* the design of novel evaluation metrics concerning the quality of automatically generated arguments, and *iv)* the integration of fact-checking into AG to produce consistent, verified and sound arguments. All these challenges call for more innovative and reliable methods which would eventually allow for applications of AG and CAG in a even larger diversity of scenarios.

Acknowledgements This work has been partially supported by the ANR project ATTENTION (ANR21-CE23-0037) and the French government through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

References

1. Alshomary, M., Chen, W.F., Gurcke, T., Wachsmuth, H.: Belief-based Generation of Argumentative Claims (Jan 2021)
2. Alshomary, M., Düsterhus, N., Wachsmuth, H.: Extractive Snippet Generation for Arguments. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1969–1972. ACM, Virtual Event China (Jul 2020)

3. Alshomary, M., Rieskamp, J., Wachsmuth, H.: Generating Contrastive Snippets for Argument Search. In: Toni, F., Polberg, S., Booth, R., Caminada, M., Kido, H. (eds.) Frontiers in Artificial Intelligence and Applications. IOS Press (Sep 2022)
4. Alshomary, M., Syed, S., Dhar, A., Potthast, M., Wachsmuth, H.: Counter-Argument Generation by Attacking Weak Premises. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 1816–1827. Association for Computational Linguistics, Online (Aug 2021)
5. Alshomary, M., Syed, S., Potthast, M., Wachsmuth, H.: Target Inference in Argument Conclusion Generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4334–4345. Association for Computational Linguistics, Online (Jul 2020)
6. Ashley, K.D., Walker, V.R.: Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law. pp. 176–180 (2013)
7. Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A., Slonim, N.: Stance classification of context-dependent claims. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 251–261 (2017)
8. Bhagavatula, C., Bras, R.L., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, S.W.t., Choi, Y.: Abductive commonsense reasoning. arXiv preprint arXiv:1908.05739 (2019)
9. Bilu, Y., Hershcovich, D., Slonim, N.: Automatic Claim Negation: Why, How and When. In: Proceedings of the 2nd Workshop on Argumentation Mining. pp. 84–93. Association for Computational Linguistics, Denver, CO (Jun 2015)
10. Bilu, Y., Slonim, N.: Claim synthesis via predicate recycling. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 525–530 (2016)
11. Boltužić, F., Šnajder, J.: Fill the gap! analyzing implicit premises between claims from online debates. In: Proceedings of the Third Workshop on Argument Mining (ArgMining2016). pp. 124–133 (2016)
12. Bose, T., Reina, A., Marshall, J.A.: Collective decision-making. Current opinion in behavioral sciences **16**, 30–34 (2017)
13. Cabrio, E., Villata, S.: Five Years of Argument Mining: A Data-driven Analysis. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. pp. 5427–5433. International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden (Jul 2018)
14. Carenini, G.: GEA: A Complete, Modular System for Generating Evaluative Arguments. In: Goos, G., Hartmanis, J., van Leeuwen, J., Alexandrov, V.N., Dongarra, J.J., Juliano, B.A., Renner, R.S., Tan, C.J.K. (eds.) Computational Science — ICCS 2001, vol. 2073, pp. 959–968. Springer Berlin Heidelberg, Berlin, Heidelberg (2001)
15. Chakrabarty, T., Trivedi, A., Muresan, S.: Implicit Premise Generation with Discourse-aware Commonsense Knowledge Models (Sep 2021)
16. Chen, W.F., Wachsmuth, H., Al-Khatib, K., Stein, B.: Learning to Flip the Bias of News Headlines. In: Proceedings of the 11th International Conference on Natural Language Generation. pp. 79–88. Association for Computational Linguistics, Tilburg University, The Netherlands (Nov 2018)
17. Cherumanal, S.P., Spina, D., Scholer, F., Croft, W.B.: Evaluating Fairness in Argument Retrieval. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. pp. 3363–3367 (Oct 2021)

18. El Baff, R., Wachsmuth, H., Al Khatib, K., Stede, M., Stein, B.: Computational argumentation synthesis as a language modeling task. In: Proceedings of the 12th International Conference on Natural Language Generation. pp. 54–64 (2019)
19. Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research **22**, 457–479 (2004)
20. Farag, Y., Brand, C., Amidei, J., Piwek, P., Stafford, T., Stoyanchev, S., Vlachos, A.: Opening up Minds with Argumentative Dialogues. Findings of EMNLP (Empirical Methods in Natural Language Processing) pp. In–Press (2022)
21. Gretz, S., Bilu, Y., Cohen-Karlik, E., Slonim, N.: The workweek is the best time to start a family–A Study of GPT-2 Based Claim Generation. arXiv preprint arXiv:2010.06185 (2020)
22. Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., Slonim, N.: A large-scale dataset for argument quality ranking: Construction and analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 7805–7813 (2020)
23. Habernal, I., Wachsmuth, H., Gurevych, I., Stein, B.: The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. arXiv preprint arXiv:1708.01425 (2017)
24. Hidey, C., McKeown, K.: Fixed That for You: Generating Contrastive Claims with Semantic Edits. 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). pp. 1756–1767. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)
25. Hua, X., Hu, Z., Wang, L.: Argument generation with retrieval, planning, and realization. arXiv preprint arXiv:1906.03717 (2019)
26. Hua, X., Wang, L.: Neural Argument Generation Augmented with Externally Retrieved Evidence (May 2018)
27. Hua, X., Wang, L.: Sentence-Level Content Planning and Style Specification for Neural Text Generation (Sep 2019)
28. Jo, Y., Bang, S., Manzoor, E., Hovy, E., Reed, C.: Detecting attackable sentences in arguments. arXiv preprint arXiv:2010.02660 (2020)
29. Lauscher, A., Wachsmuth, H., Gurevych, I., Glavaš, G.: Scientia Potentia Est—On the Role of Knowledge in Computational Argumentation. Transactions of the Association for Computational Linguistics **10**, 1392–1422 (Dec 2022)
30. Lawrence, J., Reed, C.: Argument Mining: A Survey. Computational Linguistics **45**(4), 765–818 (Jan 2020)
31. Le, D.T., Nguyen, C.T., Nguyen, K.A.: Dave the debater: A retrieval-based and generative argumentative dialogue agent. In: Proceedings of the 5th Workshop on Argument Mining. pp. 121–130. Association for Computational Linguistics, Brussels, Belgium (Nov 2018)
32. Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E., Slonim, N.: Context Dependent Claim Detection. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 1489–1500. Dublin City University and Association for Computational Linguistics, Dublin, Ireland (Aug 2014)
33. Lipton, P.: Inference to the best explanation. A Companion to the Philosophy of Science pp. 184–193 (2017)
34. Marro, S., Cabrio, E., Villata, S.: Graph Embeddings for Argumentation Quality Assessment. In: EMNLP 2022-Conference on Empirical Methods in Natural Language Processing (2022)

35. Pasumarthi, R.K., Bruch, S., Wang, X., Li, C., Bendersky, M., Najork, M., Pfeifer, J., Golbandi, N., Anil, R., Wolf, S.: Tf-ranking: Scalable tensorflow library for learning-to-rank. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2970–2978 (2019)
36. Prakken, H.: A persuasive chatbot using a crowd-sourced argument graph and concerns. Computational Models of Argument **326**, 9 (2020)
37. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving Language Understanding by Generative Pre-Training p. 12 (2018)
38. Rinott, R., Dankin, L., Alzate, C., Khapra, M.M., Aharoni, E., Slonim, N.: Show me your evidence-an automatic method for context dependent evidence detection. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 440–450 (2015)
39. Sathe, A., Ather, S., Le, T.M., Perry, N., Park, J.: Automated Fact-Checking of Claims from Wikipedia. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 6874–6882. European Language Resources Association, Marseille, France (May 2020)
40. Sato, M., Yanai, K., Miyoshi, T., Yanase, T., Iwayama, M., Sun, Q., Niwa, Y.: End-to-end Argument Generation System in Debating. In: Proceedings of ACL-IJCNLP 2015 System Demonstrations. pp. 109–114. Association for Computational Linguistics and The Asian Federation of Natural Language Processing, Beijing, China (Jul 2015)
41. Saveleva, E., Petukhova, V., Mosbach, M., Klakow, D.: Graph-based argument quality assessment. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). pp. 1268–1280 (2021)
42. Stab, C., Gurevych, I.: Recognizing insufficiently supported arguments in argumentative essays. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 980–990 (2017)
43. Tekiroglu, S.S., Chung, Y.L., Guerini, M.: Generating counter narratives against online hate speech: Data and strategies. arXiv preprint arXiv:2004.04216 (2020)
44. Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T.A., Hirst, G., Stein, B.: Computational Argumentation Quality Assessment in Natural Language. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 176–187. Association for Computational Linguistics, Valencia, Spain (Apr 2017)
45. Wachsmuth, H., Stede, M., El Baff, R., Al Khatib, K., Skeppstedt, M., Stein, B.: Argumentation synthesis following rhetorical strategies. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 3753–3765 (2018)
46. Wachsmuth, H., Syed, S., Stein, B.: Retrieval of the Best Counterargument without Prior Topic Knowledge. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 241–251. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)
47. Wang, L., Ling, W.: Neural network-based abstract generation for opinions and arguments. arXiv preprint arXiv:1606.02785 (2016)
48. Woods, B., Adamson, D., Miel, S., Mayfield, E.: Formative Essay Feedback Using Predictive Scoring Models. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 2071–2080. ACM, Halifax NS Canada (Aug 2017)
49. Zukerman, I., McConachy, R., George, S.: Using argumentation strategies in automated argument generation. In: INLG'2000 Proceedings of the First International Conference on Natural Language Generation. pp. 55–62 (2000)