Argument and Counter-argument Generation: a Critical Survey

Xiaoou Wang, Elena Cabrio, Serena Villata
xiaoou.wang@inria.fr

Université Côte d'Azur, CNRS, Inria, I3S, France

June 21, 2023

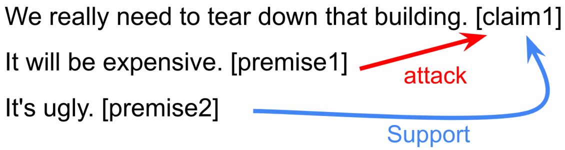
Contents

Context

What is an argument?

Definition of an argument

A constellation of propositions related to a claim (also called standpoint) which is the proposition that the argument seeks to establish.



- The fundamental components are **claims and premises**;
- Components can **attack/support** each other.

- Argument mining (AM) is a research field in NLP which aims at automatically **extracting and identifying argumentative structures** from natural language text.
- Argument Generation (AG) refers to **the generation of arguments in natural language**.
- AG has now become an expansion of AM with numerous **socially beneficial** applications such as:
  - Legal decision making [1];
  - Collective decision making [2];
  - Counter Narrative Generation to fight online hate speech [3].
  - Writing assistance [4].

Motivation
What are the available resources?

To the best of our knowledge, no survey has been published on AG and CG. The existing resources are:

- A brief chapter in [5] summarizing several relevant works up **to 2018**;
- A survey on the role of **knowledge** in AM, argument assessment, argument reasoning and AG [6].

Objectives

Why proposing a survey on (Counter-)Argument Generation?

In the meantime, a huge variety of methods have been explored in AG, under various names such as argument construction, argument retrieval, argument synthesis and argument summarization.

We propose a holistic view of AG and CAG which

- Illustrates the **historical landscape** of developments in AG and CAG research;
- Provides a detailed outline of the **main results** and especially, **various tasks and subtasks** in AG and CAG;
- Synthesizes **the key datasets**;
- Discusses **the main issues and some open challenges** in AG and CAG.

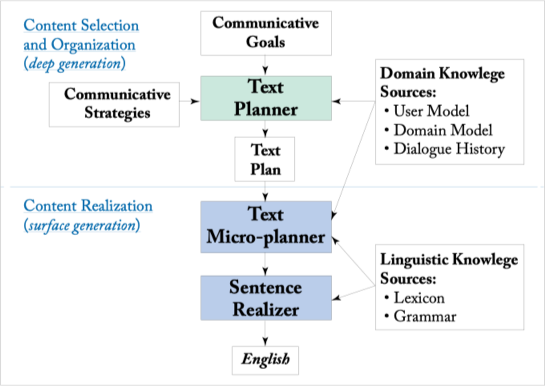
Data to text argument generation

Around 1990s, in the spirit of recommender systems

Formalized by Carenini [7] and applied to their *Generator of Evaluative Arguments* recommending houses to a client.

1. Deep generation phase;
    - Agnostic of the target language
    - Selects 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)
    - Selects argumentative strategies
2. Content realization phase
    - Requires **specific grammatical knowledge of the target language** such as verbal inflections and logical connectors

Architecture of typical data to text generation systems [7]



Data to text argument generation

Towards text to text generation

Main drawbacks:

- **Manual work** to build the knowledge base;
- 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 AG:

- **Debating systems** started to emerge (Project Debater, IBM[1])
- **Natural Language Generation** started to be used in AG.

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

Main research areas in AG and CAG:

- Generation of argument components
  - Claim Generation (CG)
  - Contrastive Claim Generation (CCG)
  - Bias Flipping
  - Premise Target Identification (PTI) and Conclusion Target Inference (CTI)
  - Enthymeme Reconstruction (ER)
- Generation of full arguments
  - Rule-based argument generation
  - Summarization-based approach
  - Other research directions
  - Counter-argument generation

Claim Generation is different from **Claim Retrieval**:

- Input: a debate topic (Internet censorship)
- Output: an assertion with a clear stance (Internet censorship is a violation of free speech)

Claim Generation

Representative works

Representative works in CG:

- Bilu and Slonim [8]
  1. A predicate on a certain topic can be used to other topics;
  2. Given a topic, word2vec embeddings are used to **select top** k **similar predicates** from a Predicate Lexicon;
  3. 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.
- Gretz et al. [9] showed the potential of GPT-2 in CG.
- Alshomary et al. [10]
  - **Encode users' beliefs** into CG by leveraging bag-of-words representations of users' stances on various topics;
  - Combine learned beliefs with an argumentative Language Model.

Negation has an important function in argumentation.

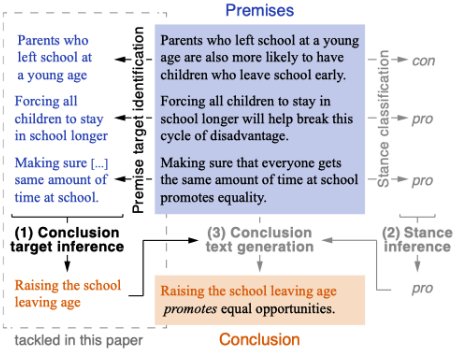
- Explicit negation is not always possible [11].
- Hidey and McKeown [12] used a seq2seq model to encode the original claim with an attention mechanism:
    - A sequence of **words** or a sequence of **edits** were used as decoder input.
        - For edits, "DELETE-N tokens" specifies n previous words to delete.
        - Hillary Clinton for president 2020 -> Hillary Clinton DELETE-2 Bernie Sanders for president 2020 DELETE-1
    - The sequence of edits representation is more effective.
- The task of *Bias Flipping* [13] (i.e., switch the left or right bias of an article) is similar to CCG.

Conclusion Generation is sometimes necessary because **conclusions often remain implicit**.

- **Premise Target Identification** [14] identifies the target in a premise.
- Based on this task, Alshomary et al. [15] initiated the task of **Conclusion Target Inference** identifying the final target in a conclusion.
- An explicit conclusion is generated once the target is identified.

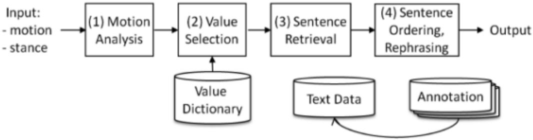
Enthymeme Reconstruction clarifies how a conclusion is inferred from the given premises.

Illustration of a model of generating an argument's conclusion from its premises [15]



Sato et al. [16] presented the **first end-to-end rule-based** retrieval system to generate arguments:

- At least **4 distinct components** need maintainance;
- The value dictionary containing talking points (economy, health, etc.) is **hand-made**.



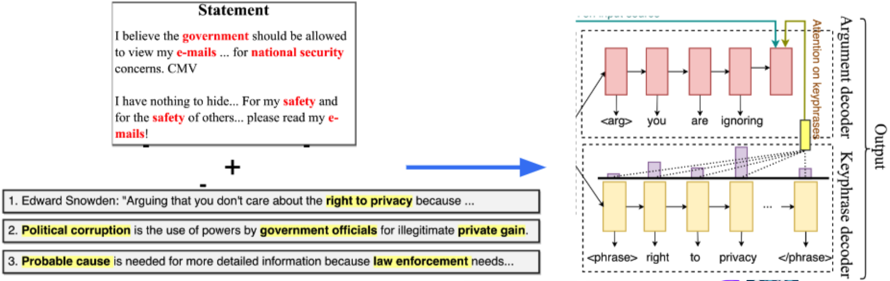
Some studies propose to use a neural summarization approach which:

- Generates arguments representing **both stances** (particularly useful for controversial topics);
- Formulates the summary by stating **the main claim and the supporting reason**. This task is called **Argument Snippet Generation (ASG)** [17];
- Draws inspiration from comparative summarization (What is different between the coverage in NYTimes and BBC) to **avoid redundancy**.

Other research directions in AG:

- **Audience-oriented Argument Generation**. To enhance the persuasiveness of the generated arguments, Alshomary et al. [10] trained a BERT-based classifier to identify morals such as care and fairness and used the Project Debater's API to generate arguments based on morals.
- El Baff et al. [18] proposed a computational model to generate arguments according to **a specific rhetorical strategy** (Logos vs. Pathos).
- Finally, the **dialogue** aspect of AG is getting more and more attention from researchers.

In terms of CAG, Hua and Wang [19] proceeded in two steps by using a seq2seq neural network: evidence retrieval and text generation. Especially, the decoding phase has a **distinct talking points generation step**.

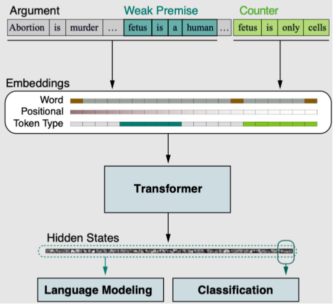


Text to text generation

Counter-argument generation by attacking weak premises

Alshomary et al. [20] proposed to attack an argument by challenging the validity of one of its premises:

- **Rank weak premises** by using the learn-to-rank framework [21].
- Combine next-token prediction and counter-argument classification to generate counter-arguments.



Text to text generation

Key datasets

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

Task	Datasets	Source	Size
CG	[24]	Crowd annotation	30k arguments, 71 topics
	[48]	Wikipedia articles	2.3k claims, 58 topics
Belief-based CG	[1]	debate.org	51k claims, 27k topics
CCG	[26]	Reddit	1,083,520 pairs of contrastive claims
Bias Flipping	[18]	Biased headlines from allsides.com	6458 claim-like headlines
CG or PTI	[7]	Wikipedia articles	2,394 claims, 55 topics
CTI	[61]	idebate.org	2,259 arguments, 676 topics
ER	[25]	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	[59]	Written by experts based on pools of ADUs representing pros and cons	130 logos-oriented and 130 pathos-oriented arguments, 10 topics
	[28]	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

Most common metrics used in CG and CAG are BLEU and METEOR which **don't capture essential qualities of an argument** [22]:

- *Cogency* (when premises are relevant to the argument's conclusion and sufficient to draw the conclusion);
- *Reasonableness* (when an argument resolves an issue in a sufficient way that is acceptable to the target audience);
- *Effectiveness*, etc.

Chen et al. [13] found that for a successful flipping (CAG), **the overlapping of generated and ground-truth headlines is very low**, making overlap-based metrics unreliable.

- Human evaluation shows that the quality of fully-generated counterarguments is yet **lower than that of a simple concatenation of evidence passages** [19, 23].
- The simple criteria of **understandability** of an argument is far from being reached.
  - In [13], out of 200 generated headlines, only 73 were understandable.
  - The rule-based system of Sato et al. [16] has the same drawback (50 out of 86 sentences are judged as non-understandable).

Challenges and open research directions

Understandability

Output of the system of Hua and Wang [19]

Input: Putin is trying to re-form a "Soviet Union" with his past actions in Georgia and current actions in Ukraine

Human: So far Putin has only really targeted areas where the local population is majority russian (or historically majority russian) and sympathetic to russia. That's not to say that he wouldn't go further, but I don't think there is yet any compelling reason to believe he would.

Retrieval: In 1968, the Soviet Union and Warsaw pact allies invaded Czechoslovakia to halt the Prague spring reforms. It was this significant lack of flexibility that was exploited by the US to undermine the strength of the Soviet Union and thus foster its reform. In the summer of 1991, the new union treaty, which would have turned the Soviet Union into a much looser union, was agreed upon by eight republics.

Our Model: There are two reasons you are so far beyond. There is no reason to see the military army. You can infer what they assume it, so they tend to protect up. There. If we assume it were bad, they can not be controlled. So we're talking going to ignore the wars of the world. The fact of the matter is not bad. I would also assume it. However, the government can not be reminded of the world.

Seq2seq + evd: 1. The reason is that the government isn't the power of the war of Russia, at the time of the war of Russia. In Russia, there is no way to continue to control the military.

Seq2seq + KP: I don't think it's a bit. I don't think you're forgetting. I don't want to see the Syrian American dream in the US. I don't see the Syrian American dream of American citizens. I don't see the Syrian American dream of American citizens. I don't see the Syrian American military in the world. I don't think it wouldn't be true.

Argumentation strategies were the focus of early studies [49]:

- Hypothetical reasoning;
- Reasoning by cases;
- Inference to the best explanation.

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.

Challenges and open research directions

Ethical issues

Although the main goal of argumentation is to convince, the truthfulness issue must be considered in certain contexts.
However:

- GPT-like models have **bias** [24] and produce **hallucinations** [25].
- Training data such as ChangeMyView are collected from Reddit, leading to **unverified claims and premises**.

Solutions:

- **Automatic evaluation of fairness** in argument retrieval [26];
- **Automated fact-checking of claims** [27], **automatic detection of insufficiently supported arguments** [28], etc.

Takeaway messages

Promising research directions

Studies on AG and CAG are clearly on the rise, with multiple subareas and research directions.

Four lines of research are particularly promising:

- Integration of **users' beliefs and preferences** in AG;
- Development of intelligent argument **dialogue systems**;
- Design of **novel evaluation metrics** concerning the quality of automatically generated arguments;
- Integration of **fact-checking** into AG to produce consistent, verified and sound arguments.

Takeaway messages

The great challenge

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.

Thank You for Your Attention!

Bibliography I

[1] Kevin D. Ashley and Vern R. Walker. "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*. 2013, pp. 176–180.

[2] Thomas Bose, Andreagiovanni Reina, and James AR Marshall. "Collective Decision-Making". In: *Current opinion in behavioral sciences* 16 (2017), pp. 30–34.

[3] Serra Sinem Tekiroglu, Yi-Ling Chung, and Marco Guerini. "Generating Counter Narratives against Online Hate Speech: Data and Strategies". In: *arXiv preprint arXiv:2004.04216* (2020). arXiv: 2004.04216.

Bibliography II

[4] Bronwyn Woods et al. "Formative Essay Feedback Using Predictive Scoring Models". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax NS Canada: ACM, Aug. 2017, pp. 2071–2080. ISBN: 978-1-4503-4887-4. DOI: `10.1145/3097983.3098160`. (Visited on 01/12/2023).

[5] Manfred Stede and Jodi Schneider. "Argumentation Mining". In: *Synthesis Lectures on Human Language Technologies* 11.2 (2018), pp. 1–191.

[6] Anne Lauscher et al. "Scientia Potentia Est—On the Role of Knowledge in Computational Argumentation". In: *Transactions of the Association for Computational Linguistics* 10 (Dec. 2022), pp. 1392–1422. ISSN: 2307-387X. DOI: `10.1162/tacl_a_00525`. (Visited on 01/17/2023).

Bibliography III

[7] Giuseppe Carenini. "GEA: A Complete, Modular System for Generating Evaluative Arguments". In: *Computational Science — ICCS 2001*. Ed. by G. Goos et al. Vol. 2073. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 959–968. ISBN: 978-3-540-42232-7 978-3-540-45545-5. DOI: 10.1007/3-540-45545-0_108. (Visited on 01/16/2023).

[8] Yonatan Bilu and Noam Slonim. "Claim Synthesis via Predicate Recycling". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2016, pp. 525–530.

[9] Shai Gretz et al. "The Workweek Is the Best Time to Start a Family–A Study of GPT-2 Based Claim Generation". In: *arXiv preprint arXiv:2010.06185* (2020). arXiv: 2010.06185.

Bibliography IV

[10] Milad Alshomary et al. *Belief-Based Generation of Argumentative Claims*. Jan. 2021. arXiv: 2101.09765 [cs]. (Visited on 01/24/2023).

[11] Yonatan Bilu, Daniel Hershcovich, and Noam Slonim. "Automatic Claim Negation: Why, How and When". In: *Proceedings of the 2nd Workshop on Argumentation Mining*. Denver, CO: Association for Computational Linguistics, June 2015, pp. 84–93. DOI: 10.3115/v1/W15-0511. (Visited on 01/23/2023).

Bibliography V

[12] Christopher Hidey and Kathy McKeown. "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 1756–1767. DOI: 10.18653/v1/N19-1174. (Visited on 12/01/2022).

[13] Wei-Fan Chen et al. "Learning to Flip the Bias of News Headlines". In: *Proceedings of the 11th International Conference on Natural Language Generation*. Tilburg University, The Netherlands: Association for Computational Linguistics, Nov. 2018, pp. 79–88. DOI: 10.18653/v1/W18-6509. (Visited on 01/24/2023).

[14] Roy Bar-Haim et al. "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*. 2017, pp. 251–261.

[15] Milad Alshomary et al. "Target Inference in Argument Conclusion Generation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4334–4345. DOI: 10.18653/v1/2020.acl-main.399. (Visited on 01/17/2023).

Bibliography VII

[16] Misa Sato et al. "End-to-End Argument Generation System in Debating". In: *Proceedings of ACL-IJCNLP 2015 System Demonstrations*. Beijing, China: Association for Computational Linguistics and The Asian Federation of Natural Language Processing, July 2015, pp. 109–114. DOI: 10.3115/v1/P15-4019. (Visited on 01/16/2023).

[17] Milad Alshomary, Nick Düsterhus, and Henning Wachsmuth. "Extractive Snippet Generation for Arguments". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Virtual Event China: ACM, July 2020, pp. 1969–1972. ISBN: 978-1-4503-8016-4. DOI: 10.1145/3397271.3401186. (Visited on 11/14/2022).

[18] Roxanne El Baff et al. “Computational Argumentation Synthesis as a Language Modeling Task”. In: *Proceedings of the 12th International Conference on Natural Language Generation*. 2019, pp. 54–64.

[19] Xinyu Hua and Lu Wang. *Neural Argument Generation Augmented with Externally Retrieved Evidence*. May 2018. arXiv: `1805.10254 [cs]`. (Visited on 01/12/2023).

[20] Milad Alshomary et al. “Counter-Argument Generation by Attacking Weak Premises”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 1816–1827. DOI: `10.18653/v1/2021.findings-acl.159`. (Visited on 01/12/2023).

Bibliography IX

[21] Rama Kumar Pasumarthi et al. "Tf-Ranking: Scalable Tensorflow Library for Learning-to-Rank". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2970–2978.

[22] Henning Wachsmuth et al. "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*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 176–187. (Visited on 06/21/2023).

[23] Xinyu Hua, Zhe Hu, and Lu Wang. "Argument Generation with Retrieval, Planning, and Realization". In: *arXiv preprint arXiv:1906.03717* (2019). arXiv: `1906.03717`.

[24] Stephanie Lin, Jacob Hilton, and Owain Evans. "TruthfulQA: Measuring How Models Mimic Human Falsehoods". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3214–3252. DOI: 10.18653/v1/2022.acl-long.229. (Visited on 01/23/2023).

[25] Hussam Alkaissi and Samy I McFarlane. "Artificial Hallucinations in ChatGPT: Implications in Scientific Writing". In: *Cureus* (Feb. 2023). ISSN: 2168-8184. DOI: 10.7759/cureus.35179. (Visited on 06/07/2023).

[26] Sachin Pathiyan Cherumanal et al. "Evaluating Fairness in Argument Retrieval". In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. Oct. 2021, pp. 3363–3367. DOI: 10.1145/3459637.3482099. arXiv: 2108.10442 [cs]. (Visited on 01/24/2023).

[27] Aalok Sathe et al. "Automated Fact-Checking of Claims from Wikipedia". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 6874–6882. ISBN: 979-10-95546-34-4. (Visited on 01/24/2023).

[28] Christian Stab and Iryna Gurevych. “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*. 2017, pp. 980–990.