Conversational Agent Debating Capability

Mohamad Faraj Makkawi
IMT Atlantique
Rennes, France
mohamad.makkawi@imt-atlantique.net

Hassan Khan
IMT Atlantique
Rennes, France
hassan.khan@imt-atlantique.net

Abstract—

Index Terms—**AI, Debate Systems, Argumentation, Conversational Agents, Argumentation Service Platform with Integrated Components (ASPIC).**

I. Introduction

Various conversational agents are developed to argue, and various methodologies are applied in the development of arguments. Rule-based agents use pre-defined templates, and they provide structured responses, which are inflexible. Retrieval-based agents learn arguments from a database, and this enables them to respond rapidly, although the responses may be irrelevant contextually [3]. Hybrid argumentation agents use retrieval with reasoning mechanisms to produce more flexible arguments, while hierarchical persuasion agents refine their arguments through the utilization of user feedback, thereby improving interaction [5]. Explainable debate agents give explicit justification to their arguments, transparency enhanced and trust.

This work investigates the argumentation capabilities of dialogue agents with a special emphasis on how symbolic artificial intelligence techniques, in the form of ASPIC and defeasible logic programming, improve argumentation by using systematic reasoning and knowledge representation. We highlight the strengths of symbolic techniques in terms of explainability and abstraction manipulation [11], along with discussing problems such as managing uncertainty, very large datasets, and real-time adaptation limits [11]. By going through current architectures, we try to evaluate the potential and limitations of symbolic AI in creating successful debate-capable conversational agents.

II. Debate Definition and What Makes a Good Debate

A debate is a structured conversation through the presentation of arguments by opposing sides with a view to expressing their positions, and justifications, and enabling the scrutiny of different perspectives [1] [2]. It can also be viewed as an argumentative game where the interactions are governed by predefined rules, and players engage in structured moves [6]. A good debate consists of empirical evidence-based arguments that are well-supported, clear explanations that spell out reasoning using simple language, and an awareness of counterarguments to back persuasive attempts. Moreover, logical coherence between premises and the categorization of argument quality are elements that make the responses effective as a whole [6].

To evaluate debate performance effectively, the judges are to evaluate organized argumentation based on clear organization, logical consistency of premises, and refutation order of counter-arguments. The quality of evidence is most important and must be based on well-researched facts, credible sources, and resolution specificity. Teams must demonstrate prioritization of argumentation and insight into the most significant "point of clash" [8]. Recognize fabricated evidence, ad hominem, or rule violations. Qualitative estimation of argument strength takes precedence over quantitative counting [7] [8].

III. Conversational Agent Debating Techniques and Applications

- **Argument graphs:** A visual representations of inter-relations between arguments, which can be formally represented as a directed graph G=(V,E), where V is the vertices (arguments) and E is the directed edges representing support and attack relations. The visualization through such graphical representations can make complicated interrelations between arguments easy to understand, making them understandable by audience members in a more readable way. Though argument graphs efficiently promote understanding and involvement, their persuasiveness relies on the quality and strength of the arguments. Therefore, they get a positive assessment for their ability to provide clarity in persuasive conversations [4] [6].
- **ASPIC framework:** A formal format for argument presentation, with the arguments mathematically definable as $A = (C, R, \lhd, \text{attack})$, whereby C represents claim, R reasons, and $\lhd$ are relations of attack and support. With the formal setup, there comes unambiguous demonstration logically, together with enabling structures for constructing trees of argument and thus fortifying the argument for persuasion based on the reason of the audience. Evaluative perspective shows that ASPIC's strong logical structure is well-suited to enable sound arguments, and thus it is the best option for promoting persuasion [6].
- **Multi-Attribute Argumentation Framework:** Ranks arguments on multiple criteria, such criteria being mea-

surable through the formula

$$S(A) = \sum_{i=1}^{n} w_i \cdot v_i$$

where w_i are weights for each criterion and v_i are the respective scores for each criterion. The multi-aspect analysis process allows for sensitive rankings that are capable of communicating directly with the values of the audience, optimally maximizing the persuasive power of the framework. In general, this approach is graded as superb for its ability to be able to handle heterogeneous preferences, enhancing arguments across various stakeholder groups [6].

IV. COMPARATIVE ANALYSIS AND EVALUATION OF DEBATE AGENTS

This section provides a comparative analysis of conversational agents for debate applications, with an overview of agent types in Table I and their evaluation criteria in Table II.

TABLE I
OVERVIEW OF DIFFERENT CONVERSATIONAL AGENT TYPES FOR DEBATE APPLICATIONS

Agent Type	Approach	Use Cases	Strengths	Weaknesses
Rule-Based	Predefined templates	Compliance systems [12]	Predictable outcomes	Rigid structure
Retrieval-Based	DB queries + RAG	Customer support chatbots [1]	Dynamic data handling	Retrieval dependency
Hybrid	Rule-retrieval integration	Legal argumentation [3]	Context-aware reasoning	Complex setup
Hierarchical	Feedback-driven adaptation [5]	Policy negotiation	Personalized arguments	Feedback latency
Explainable	Justification chains [13]	Diagnostics, Legal systems	Transparent explanations	Depth limitations

V. SYMBOLIC METHODS: EXPLAINABILITY AND VALIDITY IN LOGICAL REASONING

Symbolic methods emphasize knowledge representation using structured, legible symbols founded on sound logical reasoning to reason about problems. They leverage formal logic, rule-based systems, and knowledge graphs. Their greatest strength is explainable reasoning, most critical in the event that there might be a necessity of justifying an argument or choice [11]. To evaluate reasoning in symbolic methods, we ensure that each step logically follows earlier premises without contradiction, usually through formal entailment tests or truth tables of propositional logic. In richer systems, soundness is achieved by analyzing streams of reasoning, comparing them with symbolic ground truths (for example, mathematical theorems or logical derivations), and cross-verifying against recognized rules [20]. This explainability enhances user trust in the system and its usability [11].

TABLE II
EVALUATION METRICS FOR CONVERSATIONAL AGENTS IN DEBATE APPLICATIONS

Evaluation criteria	Rule-Based	Retrieval	Hybrid [19]	Hierarchical [18]	Explainable [14] [17]
Accuracy	High in narrow domains [14]	Moderate (depends on data quality) [15]	Balanced across tasks	Varies by hierarchy level	Moderate (prioritizes transparency over optimization)
Response Time [16]	Fast (deterministic)	Moderate (retrieval latency)	Slow (multi-layer processing)	Slowest (sequential decision-making)	Moderate (additional explainability overhead)
Cost Efficiency	Low (static rules)	Moderate (retrieval infrastructure)	High (compute-intensive)	High (maintenance complexity)	Moderate (explanation generation costs)
Scalability	Poor (rule complexity)	High (modular data updates)	Moderate (resource-intensive)	Best (layered abstraction)	Limited (interpretability constraints)
Stability	High (deterministic)	Moderate (data drift sensitivity)	High (adaptive tuning)	High (isolated failures)	Moderate (explanation consistency)
User Satisfaction	Low (inflexible)	Moderate (context-aware)	High (balanced UX)	Moderate (learning curve)	Highest (trust through clarity)

VI. LIMITATIONS OF CONVERSATIONAL DEBATING AGENTS

- **Uncertainty and Big Data Management:** AI Debate Systems may need to handle vast amounts of data from multiple sources, often containing noisy or inconsistent information [11].
- **Knowledge Acquisition Bottleneck:** Symbolic AI programs typically require hand-coded rules and knowledge, leading to high human effort and cost in translating real-world problems into system inputs [11].
- **Real-time Dynamic Assessments:** AI debate systems face challenges in reacting dynamically to emerging arguments or rapidly changing contexts in real-time scenarios [11].
- **Logical Inconsistencies:** Internal logical inconsistencies present a major hurdle in designing secure AI systems, such as debate frameworks. This issue highlights the need to integrate constrained logical verification or consistency inference tools into the AI architecture [10].

VII. CONCLUSION

ACKNOWLEDGMENT

REFERENCES

[1] Rakshit, Geetanjali, Kevin K. Bowden, Lena Reed, Amita Misra, and Marilyn Walker. "Debbie, the debate bot of the future." In *Advanced Social Interaction with Agents: 8th International Workshop on Spoken Dialog Systems*, pp. 45-52. Springer International Publishing, 2019.

[2] Tan, Chenhao, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. "Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions." In *Proceedings of the 25th International Conference on World Wide Web*, pp. 613-624, 2016.

[3] Kulatska, Iryna. "ArgueBot: Enabling debates through a hybrid retrieval-generation-based chatbot." Master's thesis, University of Twente, 2019.

[4] Chalaguine, Lisa A., and Anthony Hunter. "A persuasive chatbot using a crowdsourced argument graph and concerns." *Computational Models of Argument* 326 (2020): 9.

[5] Sakai, Kazuki, Ryuichiro Higashinaka, Yuichiro Yoshikawa, Hiroshi Ishiguro, and Junji Tomita. "Hierarchical argumentation structure for persuasive argumentative dialogue generation." *IEICE TRANSACTIONS on Information and Systems* 103, no. 2 (2020): 424-434.

[6] Engelmann, Débora, Juliana Damasio, Alison R. Panisson, Viviana Mascardi, and Rafael H. Bordini. "Argumentation as a method for explainable AI: A systematic literature review." In *17th IEEE Iberian Conference on Information Systems and Technologies (CISTI)*, 2022, pp. 1-6.

[7] Debating for Everyone. "How to Judge a Debate." Available: https://www.debatingforeveryone.com/resources/how-debating-works/how-to-judge-a-debate. [Accessed: Apr. 3, 2025].

[8] Shuster, Kate, and Meany, John. "Judging Debates: The Middle School Public Debate Program Judge Certification Manual." Available: https://esu.fcny.org/esu/programs/middle_school_debate/educators/lesson_plans_teaching_materials/MSPDP_judging_manual:en-us.pdf. [Accessed: Apr. 3, 2025].

[9] Vassiliades, A., Bassiliades, N., and Patkos, T. "Argumentation and explainable artificial intelligence: A survey," *Knowledge Engineering Review*, vol. 36, e5, 1-35, 2021. doi:10.1017/S0269888921000011. [Online]. Available: https://doi.org/10.1017/S0269888921000011. [Accessed: Mar. 23, 2025].

[10] Ilkou, E., and Koutraki, M. "Symbolic vs Sub-symbolic AI Methods: Friends or Enemies?" In *Proceedings of the CIKM 2020 Workshops*, CEUR Workshop Proceedings, ISSN 1613-0073, October 19-20, 2020, Galway, Ireland. [Online]. Available: http://ceur-ws.org. [Accessed: Mar. 23, 2025].

[11] Kasif, Simon. "A Trilogy of AI Safety Frameworks: Paths from Facts and Knowledge Gaps to Reliable Predictions and New Knowledge." *Department of Biomedical Engineering, Program in Bioinformatics, Department of Computer Science, Boston University*.

[12] Monte-Alto, H. H. L. C., Possebom, A. T., Morveli-Espinoza, M. M. M., and Tacla, C. A. "A rule-based argumentation framework for distributed contextual reasoning in dynamic environments." *DYNA*, vol. 88, no. 217, pp. 120-130, 2021. doi:10.15446/dyna.v88n217.90858.

[13] Ali, B., Pawar, S., Palshikar, G. K., & Singh, R. (2022). Constructing a dataset of support and attack relations in legal arguments in court judgements using linguistic rules. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pp. 491-500. Marseille, 20-25 June 2022. European Language Resources Association (ELRA). Retrieved from: https://www.elra.info.

[14] IBM. "AI Agent Evaluation." Available: https://www.ibm.com/think/topics/ai-agent-evaluation. [Accessed: Apr. 3, 2025].

[15] Smythos. "AI Agent Performance Measurement." Available: https://smythos.com/ai-agents/agent-architectures/ai-agent-performance-measurement/. [Accessed: Apr. 3, 2025].

[16] Galileo. "AI Agent Metrics." Available: https://www.galileo.ai/blog/ai-agent-metrics. [Accessed: Apr. 3, 2025].

[17] SuperAnnotate. "AI Agent Evaluation." Available: https://www.superannotate.com/blog/ai-agent-evaluation. [Accessed: Apr. 3, 2025].

[18] Galileo. "Evaluating AI Agent Performance: Benchmarks for Real-World Tasks." Available: https://www.galileo.ai/blog/evaluating-ai-agent-performance-benchmarks-real-world-tasks. [Accessed: Apr. 3, 2025].

[19] Aisera. "AI Agent Evaluation." Available: https://aisera.com/blog/ai-agent-evaluation/. [Accessed: Apr. 3, 2025].

[20] Lee, Jinu, and Hockenmaier, Julia. "Evaluating Step-by-step Reasoning Traces: A Survey." Available: https://arxiv.org/html/2502.12289v1. [Accessed: Apr. 3, 2025].