Towards an explainable Argumentation-based Dialogue pipeline for Conversational Recommender Systems

Marco Grazioso[1,2,*,†], Martina Di Bratto[1,2,†], Azzurra Mancini[1,†] and Valentina Russo[1,†]

[1] *Logogramma S.R.L., Naples, Italy*
[2] *Interdepartmental Research Center Urban/Eco, University of Naples Federico II, Naples, Italy*

Abstract

Conversational Recommender Systems (CoRS) have become popular for offering personalized recommendations through interactive dialogue. However, with the advent of Large Language Models (LLMs), reasoning and linguistic aspects of dialogue have been sidelined, leaving room for the risk of system hallucinations or nonsensical responses. A new approach, called Argumentative Conversational Recommender Systems (A-CoRS), is considered to address this issue. It offers an explainable and domain-independent method that incorporates cognitive pragmatics to analyze the reasoning behind system dialogue's moves. Starting from this approach, in this preliminary work, we propose a pipeline that aims to ensure explainability while leveraging the powerful capabilities of LLMs. This compositional approach will be crucial for the generative part of the linguistic output, ensuring naturalness and discourse coherence.

Keywords

Conversational Recommender system, Retrieval Augmented Generation, Argumentation-based Dialogue, Large Language Model

1. Introduction

This work centers on an hybrid approach for modelling recommendations in an argumentation-based dialogue (ABD) scenario. [1] Conversational recommender systems (CoRS) can be framed in the field of formal argumentation and more specifically, refer to the argumentation-based dialogue. It considers the problems arising from dialogues involving different agents and whose information are shared and distributed among them. Previous research on ABD highlights the importance of both agent-related and dialogical aspects, as argumentation involves dynamic exchanges of information that vary with turns and participants. Despite the long-standing study of argumentation in dialogues, there is still no unified theoretical framework for managing it [1]. One prominent application for the evaluation of ABD model is in CoRS, as it offers an ideal setting to examine the theoretical model that underpins its computational implementation. The recommendation task is, indeed, particularly suitable for this study due to its inner dialogical structure and the defined goal it encompasses. This task revolves around a clear dialogical pattern that involves two distinct phases, Exploration and Exploitation (E&E). These phases can be viewed as two intertwined types of dialogues: exploration refers here to the system gathering beliefs; conversely, during the exploitation phase, the system capitalises on the best-known option [2]. However, with the advent of Large Language Models (LLMs), reasoning and linguistic aspects of dialogue have been sidelined. As a matter of fact, it has been observed that they excel in generating human-like text responses in a conversational manner. But a closer look shows all the drawbacks of this model: hallucinations, faulty reasoning, emergent abilities etc. This is because, in addition to being efficient in the rules and statistical regularities of language, a competent language user must be able to use language to do things in the world [3], he or she has to act in the world and not just react to it as machine learning agents still do [4]. The work is structured as follows: In Section 2, we will present a new hybrid approach, namely the Argumentative Conversational Recommender System (A-CoRS) as discussed in Di Bratto et al. [5]. Section 3 will cover the authors' theoretical model for selecting plausible arguments. In Section 4, we will show how the system collects beliefs during its interaction with the user. Finally, in Section 5, we will propose a new hybrid pipeline demonstrating how to leverage the power of LLMs while maintaining explainability. In Section 6, a brief discussion and final conclusions will be presented, along with suggestions for future work.

2. Argumentative Conversational Recommender System (A-CoRS)

In recent years, CoRS have garnered attention for their ability to provide personalized recommendations through natural and interactive dialogue [7, 8]. However, the assessment of argument quality and effectiveness, emphasized by argumentation theory scholars [9, 10], has been underexplored in CoRS. To address this, Di Bratto et al. [5] propose a new methodology for selecting and evaluating the quality of argumentation dialogues within the framework of Argumentative Conversational Recommender Systems (A-CoRS). The authors propose an interdisciplinary approach fully explainable, goal-oriented, and domain-independent. The model uses a linguistics-based approach with an exploration-exploitation mechanism grounded in cognitive pragmatics, allowing for the analysis of the reasoning behind each system dialogue move. Nowadays, on the other hand, LLMs represent the most advanced architecture. These generative agents can create original responses based on user input rather than relying on pre-defined text. They utilize Deep Learning models to predict and generate the next elements in a sequence of words. While these systems require significant training and may produce repetitive or nonsensical responses (i.e., hallucinations), there is substantial room for improvement. Integrating computational argumentation formalism could help overcome issues like the lack

The 1st Workshop on Risks, Opportunities, and Evaluation of Generative Models in Recommender Systems (ROEGEN@RECSYS'24)

[*] Corresponding author.

[†] These authors contributed equally.

mgrazioso@logogramm.com (M. Grazioso);
mdibratto@logogramma.com (M. D. Bratto);
amancini@logogramma.com (A. Mancini); vrusso@logogramma.com (V. Russo)
0000-0002-4056-544X (M. Grazioso)

[1] The presentation video of this work is available at the following link: https://youtu.be/sfgbdAgQcd0

Theoretical Model	Description	Computational Score	Description
Credibility	a measure of the number and values of all supporting data, contrasted with all conflicting data, down to external and internal sources.	**Authority score**	the authority score identifies the node with a fundamental role in the graph since a solid number of hub nodes support its validity.
Importance	a measure of the epistemic connectivity of the datum, i.e., the number and values of the data that the agent will have to revise, should they revise that single one.	**Hub score**	The hub score identifies nodes that, in the graph, support many authoritative nodes through their outgoing relationships.
Relevance	a measure of the pragmatic utility of the datum, i.e., the number and values of the (pursued) goals that depends on that datum.	**Etropy**	the entropy identifies which relevant and less certain data needs a feedback to collect more certain knowledge and improve the possibilities for the dialogue goal to be achieved.
(Un-)likeability	a measure of the motivational appeal of the datum, i.e., the number and values of the (pursued) goals that are directly fulfilled by that datum.	**Hard Evidence**	the system beliefs involvement in the selection of the feature explicates the user appeal toward that kind of data

Table 1
Cognitive properties described in [6] and then mapped on computational scores on a graph database as in [5]

of explainability [11]. Many other scholars have already attempted to provide effective solutions [12, 13, 14]. Nevertheless, these approaches still aim to enhance the apparent reasoning capabilities of LLMs by using logical tools such as graphs or trees to generate adequate prompts. This work presents a model that combines rule-based systems, NLP, and generative models using a linguistically motivated approach. It first leverages logical reasoning, with the results serving as prompts for a large language model, thereby ensuring explainability, natural linguistic output, and coherent discourse.

3. Theoretical model

A theoretical model is needed to identify and select relevant data as system beliefs. The Data-oriented Belief Revision (DBR) model, introduced by Paglieri and Castelfranchi [6], evaluates the reliability and strength of data, and incorporates Toulmin's argumentation model [15], as it aligns with the belief-changing process. The model identifies four key properties of data based on cognitive reasons: i) **Credibility**: a measure of the number and values of all supporting data; ii)**Importance**: a measure of the epistemic connectivity of the datum; iii)**Relevance**: a measure of the pragmatic utility of the datum; iv) **(Un-)Likeability**: a measure of the motivational appeal of the datum. The selection of a theoretical model for argument selection hinges on the measurability of certain features. The dialogue management module utilizes a graph database [16] containing common knowledge from Linked Open Data. Using the HITS algorithm [17], the system analyzes the graph to assign authority and hub scores to nodes, helping to prioritize the disambiguation of data within dialogues. The plausibility of new information, determined by its connectivity within the user's knowledge, influences its acceptance as a belief. Therefore, the system selects data based on numerical measures on the graph, which are mapped to the cognitive properties of the DBR model, as shown in the table 1.

4. Conversational AI and Belief Graph

The Conversational AI represents the implementation of the proposed theoretical model [5]. The model extracts relevant sub-graphs from the knowledge database, analyzes them in the context of user information, and evaluates potential dialogue moves. It combines the strengths of long-term planning from rule-based AI with the generalization and fuzzy decision-making of probabilistic methods. When the user answers, the graph structure is updated to represent the system belief graph according to the feedback and a new set of base target items, consistent with the new beliefs, is extracted together with their features. The details of this processing are exemplified in Figure 1. Di Bratto et al. [5] demonstrated, through human evaluation of simulated dialogues based on their model, that the perceived relevance and plausibility of selected arguments were rated as effective for the acceptability of the recommendation. Furthermore, in a more recent study [18], the authors asked to human users to evaluated their real interaction with the complete dialogue system. Results show that, while the complete dialogue system was seen as likable, easy to use, and controllable, it was also considered unmotivating and slow. This work seeks to refine the linguistic output to enhance user perception by leveraging Retrieval-Augmented Generation (RAG)[19].

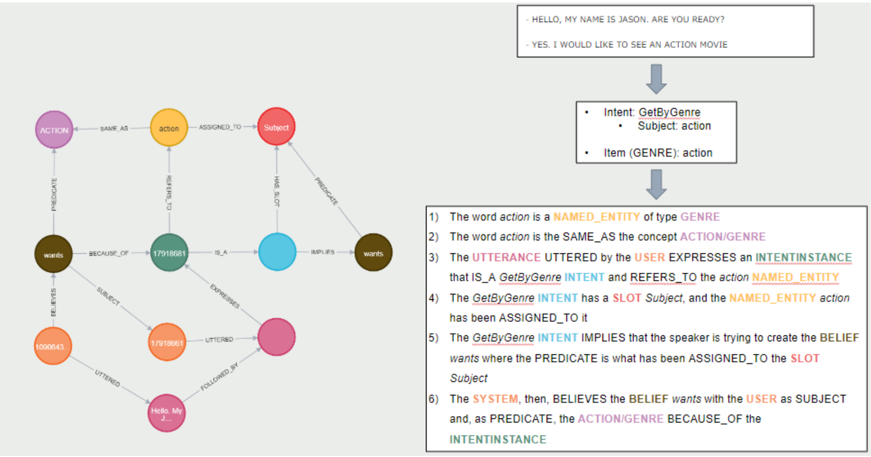
5. RAG configuration for Argument-augmented recommendation

LLMs have shown impressive capabilities in solving several tasks in zero/few-shot configuration, reaching or improving state-of-the-art models' performances [21, 22]. In particular, LLMs are able to combine different textual sources in order to answer questions, summarise contents [23], or paraphrasing texts [24]. However, Large Language Models (LLMs) face limitations in accessing domain-specific, real-time, or proprietary information, which can lead to inaccuracies or hallucinations. To overcome this, Retrieval-Augmented Generation (RAG) integrates external knowledge during the generation process, improving the accuracy and relevance of the content produced by LLMs [25]. In particular, the goal of RAG-empowered LLMs is to improve their capabilities for societal benefits, but research shows they can be manipulated [26, 27], leading to unreliable decisions and privacy issues. To prevent harm, trustworthy RA-LLMs must have robustness, fairness, explainability, and privacy protection. In this work, we propose an architecture based on the concept of RAG [19] which has shown to mitigate issues like hallucinations by ensuring explainability, [28] and it demonstrate to still keep good performances when using fine-tuned small LLMs [29].

The architecture proposed in the Figure 2 leverages these capabilities by combining the collected belief data with the recommended movies domain data to generate a human-like recommendation text providing a robust argumentation. Moreover, since obtained texts are based on input gathered from the knowledge and belief graph, the system internal process is always observable and, then, explainable.

The architecture is composed of the following elements:

- a **Dialogue system** in charge of conversing with users while collecting beliefs.
- a **Recommendation Engine** in charge of reasons upon knowledge graph and belief data to obtain a list of recommended items.
- an **Argument Generator** in charge of aggregating the domain knowledge about the recommended



- HELLO, MY NAME IS JASON. ARE YOU READY?

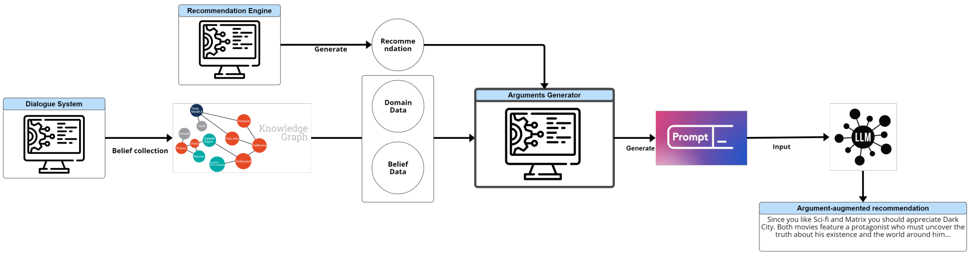
- YES. I WOULD LIKE TO SEE AN ACTION MOVIE

- Intent: GetByGenre
    - Subject: action
- Item (GENRE): action

1) The word *action* is a NAMED_ENTITY of type GENRE
2) The word *action* is the SAME_AS the concept ACTION/GENRE
3) The UTTERANCE UTTERED by the USER EXPRESSES an INTENTINSTANCE that IS_A *GetByGenre* INTENT and REFERS_TO the *action* NAMED_ENTITY
4) The *GetByGenre* INTENT has a SLOT *Subject*, and the NAMED_ENTITY *action* has been ASSIGNED_TO it
5) The *GetByGenre* INTENT IMPLIES that the speaker is trying to create the BELIEF *wants* where the PREDICATE is what has been ASSIGNED_TO the SLOT *Subject*
6) The SYSTEM, then, BELIEVES the BELIEF *wants* with the USER as SUBJECT and, as PREDICATE, the ACTION/GENRE BECAUSE_OF the INTENTINSTANCE

Figure 1: An example of belief graph structure to collect user preferences in the movie recommendation domain [20, 5]

.



Since you like Sci-fi and Matrix you should appreciate Dark City. Both movies feature a protagonist who must uncover the truth about his existence and the world around him...

Figure 2: An overview of the presented pipeline.

items with the belief data obtaining a coherent prompt.

- an **LLM** which takes as input the generated prompt and provide an argument-augmented recommendation.

```
Given that:
Marco likes the <Genre> genre
Marco likes the movie <Movie 1>
The recommended movies are: <Movie 2>,
<Movie 3>
Plots:
- <Movie 1>: <plot 1>
- <Movie 2>: <plot 2>
- <Movie 3>: <plot 3>
Recommend the suggested movies by
providing an explanation based only
on the facts and plots described above.
```

Listing 1: Prompt structure

6. Discussion and conclusion

In this work, we presented an argumentation-based dialogue model applied to A-CoRS. We chose a theoretical model for data selection as beliefs where cognitive properties of arguments are mapped onto numerical measures computed over a knowledge graph. This approach guides the conversation and selects the most plausible and effective arguments to request feedback on, thereby providing valuable recommendations based on user preferences. The system's collected beliefs have been stored in the database as a graph, to be then retrieved in the following recommendation steps.Finally, a RAG system has been implemented to integrate recommended items, knowledge data, and belief data collected during the interaction, thereby generating a prompt to feed an LLM and provide an argument-augmented recommendation.

The implementation of a RAG-like system helps mitigate hallucination issues, as the model is only required to use the information included in the prompt. Although empirical tests show that large models provide good recommendations, RAG systems have also demonstrated strong performance with smaller, fine-tuned models, making them suitable for low-resource scenarios [29]. Additionally, discourse coherence and explainability are ensured because the reasoning process occurs outside the LLM, allowing us to observe the internal state of the belief graph, the domain sources used for the recommendation, and the data provided as input to the LLM.

This work represents early-stage research that needs further investigation and a rigorous test bed to provide statistically relevant data for evaluating recommendation accuracy, LLM issue occurrences, user acceptance rates, and user experience through satisfaction questionnaires.

Acknowledgments

This work was supported by the European Union - Next Generation EU under the Italian National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.3, CUP E83C22004640001, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART")

References

[1] H. Prakken, Historical overview of formal argumentation, volume 1, College Publications, 2018.

[2] C. Gao, W. Lei, X. He, M. de Rijke, T.-S. Chua, Advances and challenges in conversational recommender systems: A survey, AI Open 2 (2021) 100–126.

[3] J. L. Austin, How to do things with words, volume 88, Oxford university press, 1975.

[4] J. Pearl, D. Mackenzie, The book of why: the new science of cause and effect, Basic books, 2018.

[5] M. Di Bratto, A. Origlia, M. Di Maro, S. Mennella, Linguistics-based dialogue simulations to evaluate argumentative conversational recommender systems, User Modeling and User-Adapted Interaction (2024) 1–31.

[6] F. Paglieri, C. Castelfranchi, Revising beliefs through arguments: Bridging the gap between argumentation and belief revision in mas, in: Argumentation in Multi-Agent Systems: First International Workshop, ArgMAS 2004, New York, NY, USA, July 19, 2004, Revised Selected and Invited Papers 1, Springer, 2005, pp. 78–94.

[7] D. Jannach, A. Manzoor, W. Cai, L. Chen, A survey on conversational recommender systems, ACM Computing Surveys (CSUR) 54 (2021) 1–36.

[8] D. Pramod, P. Bafna, Conversational recommender systems techniques, tools, acceptance, and adoption: A state of the art review, Expert Systems with Applications (2022) 117539.

[9] F. Macagno, Argument relevance and structure. assessing and developing students' uses of evidence, International Journal of Educational Research 79 (2016) 180–194.

[10] F. Macagno, Argumentation profiles: A tool for analyzing argumentative strategies, Informal Logic 42 (2022) 83–138.

[11] F. Castagna, N. Kökciyan, I. Sassoon, S. Parsons, E. Sklar, Computational argumentation-based chatbots: a survey, Journal of Artificial Intelligence Research 80 (2024) 1271–1310.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[13] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, K. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, Advances in Neural Information Processing Systems 36 (2024).

[14] M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, M. Podstawski, L. Gianinazzi, J. Gajda, T. Lehmann, H. Niewiadomski, P. Nyczyk, et al., Graph of thoughts: Solving elaborate problems with large language models, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 17682–17690.

[15] S. E. Toulmin, The uses of argument, Cambridge university press, 2003.

[16] J. Webber, A programmatic introduction to neo4j, in: Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity, 2012, pp. 217–218.

[17] J. M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM (JACM) 46 (1999) 604–632.

[18] M. Bratto, M. Maro, A. Origlia, On the use of plausible arguments in explainable conversational ai, 2024, pp. 4054–4058. doi:10.21437/Interspeech.2024-839.

[19] K. Meduri, G. S. Nadella, H. Gonaygunta, M. H. Maturi, F. Fatima, Efficient rag framework for large-scale knowledge bases (2024).

[20] A. Origlia, M. Di Bratto, M. Di Maro, S. Mennella, Developing embodied conversational agents in the unreal engine: The fantasia plugin, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 6950–6951.

[21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[22] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 (2023).

[23] H. Jin, Y. Zhang, D. Meng, J. Wang, J. Tan, A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods, arXiv preprint arXiv:2403.02901 (2024).

[24] V. Yadav, Z. Tang, V. Srinivasan, Pag-llm: Paraphrase and aggregate with large language models for minimizing intent classification errors, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 2569–2573. URL: https://doi.org/10.1145/3626772.3657959. doi:10.1145/3626772.3657959.

[25] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, S. Tang, Graph retrieval-augmented generation: A survey, arXiv preprint arXiv:2408.08921 (2024).

[26] G. Deng, Y. Liu, K. Wang, Y. Li, T. Zhang, Y. Liu, Pandora: Jailbreak gpts by retrieval augmented generation poisoning, arXiv preprint arXiv:2402.08416 (2024).

[27] W. Zou, R. Geng, B. Wang, J. Jia, Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models, arXiv preprint arXiv:2402.07867 (2024).

[28] K. Shuster, S. Poff, M. Chen, D. Kiela, J. Weston, Retrieval augmentation reduces hallucination in conversation, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 3784–3803. URL: https://aclanthology.org/2021.findings-emnlp.320. doi:10.18653/v1/2021.findings-emnlp.320.

[29] T. Zhang, S. G. Patil, N. Jain, S. Shen, M. Zaharia, I. Stoica, J. E. Gonzalez, Raft: Adapting language model to domain specific rag, arXiv preprint arXiv:2403.10131 (2024).