Evaluating Agent Interactions Through Episodic Knowledge Graphs

Selene Báez Santamaría & Piek Vossen & Thomas Baier

Computational Linguistics and Text-Mining Lab (CLTL)
Vrije Universiteit Amsterdam

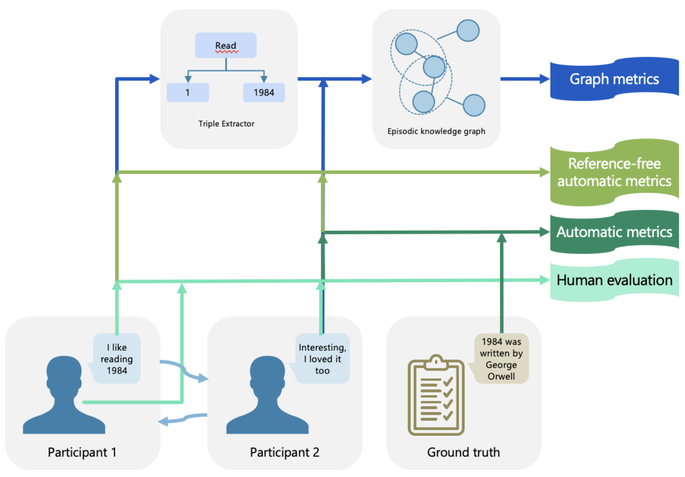
Introduction

We explore the usage of graph representations to measure dialogue quality, as these enable the explicit depiction of connectivity throughout and across conversations.

Dialogue evaluation frameworks

Evaluation frameworks are categorized as:

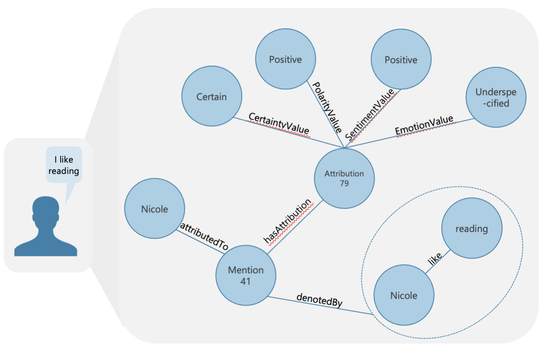
- Human Annotation: Labor intensive, Subjective, Inconsistent across experiments
- Automatic metrics: Require gold data, Punish creativity and personalization
- Reference-free automatic metrics: Reactive, Ignore content relevance



From previous work we observe that processing interactions as simple trancripts is insufficient. We propose representing dialogue as graphs, which can better express connectivity and also serve as an explicit memory. Our evaluation framework is reference-free and considers the interaction between both interlocutors.

Semantic quality in conversations

- Represents content, form, and metadata of multimodal interactions
- RDF enables (partial) semantics and reasoning
- Unique graphs arise from different (sequence) of interactions



Interactions with the same agent result in a series of graphs over time. A cumulative graph can be seen as an **Episodic Knowledge Graph** [1].

Experiments

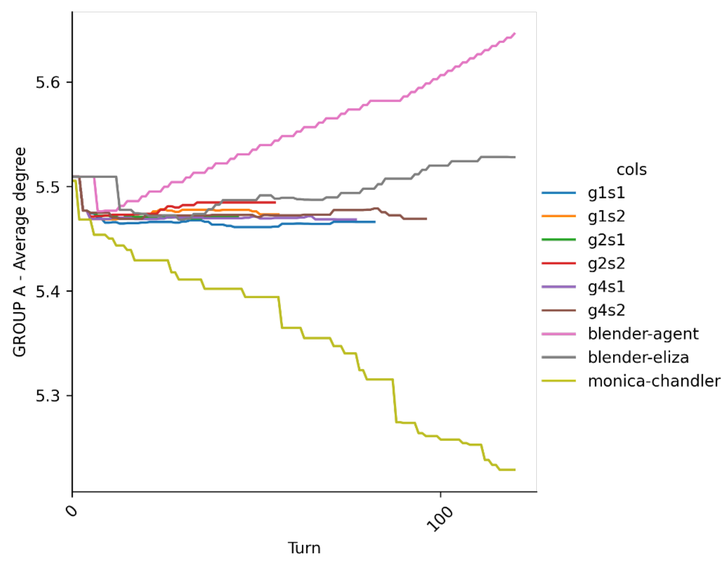
Data: Types of conversations:

- Human-Human: Scripted dialogues between Chandler and Monica in the Friends series.
- Human-Machine: Users interacting with our tailored artificial agent.
- Machine-Machine: Blenderbot [2] chatting with a tailored artificial agent or with Eliza [3].

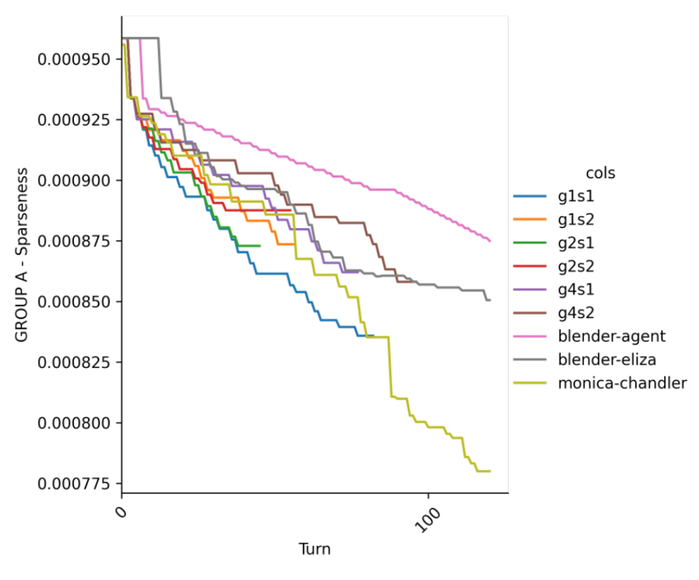
Method: We build graphs for these interactions and compute 64 graph metrics, computed on every turn to see the gradual changes in the graph. We collect human judgments based on the DSTC9 questionnaire [4]. Finally, we compute correlations between the human annotations and the graph metrics.

Results

Average node degree correlates with fluency. A steep curve down signals a fluent conversation. Human-Human conversations are the most fluent; while Machine-Machine are the least fluent.

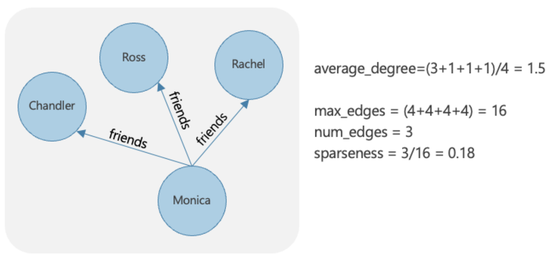


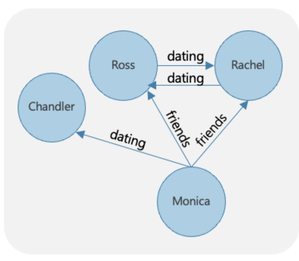
Graph sparseness correlates with the overall human rating. Successful conversations expand initially dense graphs. Similar to the previous, Human-Human are the most successful conversations; however Machine-Machine conversations diverge.

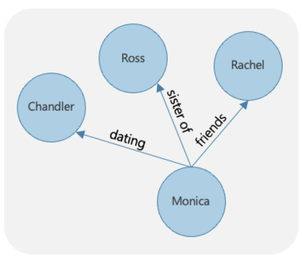


Findings

Connectivity metrics such as average node degree are able to measure changes in structure, while sparseness also incorporates semantics.







Take away

Our results suggest that graph connectivity metrics can approximate both fluency and overall human ratings. A fluid interaction is characterized by the continuous introduction of new entities. However, for a successful conversation, these entities should be sufficiently semantically related a) to the agent's knowledge, and b) to previous interactions.

References

Selene Baez Santamaria, Thomas Baier, Taewoon Kim, Lea Krause, Jaap Kruijt, and Piek Vossen.
EMISSOR: A platform for capturing multimodal interactions as episodic memories and interpretations with situated scenario-based ontological references.
In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, June 2021.

Jungseob Lee, Midan Shim, Suhyune Son, Yujin Kim, Chanjun Park, and Heuiseok Lim.
Empirical study on blenderbot 2.0 errors analysis in terms of model, data and user-centric approach.
arXiv preprint arXiv:2201.03239, 2022.

Joseph Weizenbaum.
Eliza—a computer program for the study of natural language communication between man and machine.
Communications of the ACM, 9(1):36–45, 1966.

Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Rastogi, et al.
Overview of the ninth dialog system technology challenge: Dstc9.
arXiv preprint arXiv:2011.06486, 2020.

Acknowledgements

This research was funded by the Vrije Universiteit Amsterdam and the Netherlands Organisation for Scientific Research (NWO) via the *Spinoza* grant (SPI 63-260) awarded to Piek Vossen, and the *Hybrid Intelligence Centre* via the Zwaartekracht grant (024.004.022)

VU VRIJE UNIVERSITEIT AMSTERDAM Hybrid Intelligence