A Brief Survey of Offline Explainability Metrics for Conversational Recommender Systems

J. May and K. Poudel

Dept. of Computer Science, Middle Tennessee State University, Murfreesboro, Tennessee, USA
Jam2ft@mtmail.mtsu.edu, khem.poudel@mtsu.edu

Conversational recommendation systems (CRS) are being embedded into search engines, streaming services, and commercial products. Explanations in conversational recommendation systems can increase user trust and satisfaction and help system designers fine-tune and improve the system. [1-2] Evaluating the quality of explanations for a recommendation system is a current challenge due to its relative newness as an attribute to optimize for, a small number of conversational recommendation datasets, and also due to a lack of enough annotated data. [1,3-4] To facilitate the improvement of conversational recommendation systems explain-ability we survey and propose the creation of an automated metric to score how well a conversational recommender system explains itself for multiple levels of stakeholders, which evaluates the explain-ability of recommendation output and for each turn dialog output.

We will compare 4 CRS datasets: E-Redial [5], OpenDialKG [6], INSPIRED [7], and ReDial [8] in how each dataset was created, stored, the size of the dataset, and storage type. This information will be used to determine what data formats our new metric, AutoRecExp, needs to support. AutoRecEcp is meant to be model-independent and thus be compatible with as many storage formats as possible, and is modeled after ExpScore [9].

Dialogue will be split into system and user entries, and then system outputs will be subdivided into clarifying chit-chat, and justification groupings. From there the dialogue will be examined in explainable terms. System output will be compared to the user's query to determine how well the system explained its response in relevance, length, grammatical correctness, popularity, and subjectivity [1-4,9-10]. A quality recommendation has multiple facets. AutoRecExp will be an aggregate of other measurements, but will additionally output a score for each of the above requirements. The output of the AutoRecExp readout will be customizable so that data erroneous for a given task can be omitted, and the importance of other facets can be highlighted.

Other methods focus on a deep learning approach. These methods have their own benefits but run counter to some tenets of explainable artificial intelligence (XAI). Deep learning evaluative methods may be accurate but are much more difficult to explain due to their complexity. The complexity affects the transparency of the metric.

In this work, we have examined the benefits of XAI and how explain-ability can be incorporated into and benefit CRSs. We propose a purely algorithmic metric, AutoRecExp, which examines the output of a CRS and rates how explainable its output is. Some issues to address for automated metrics are that recommendation systems are generally not designed for high levels of transparent output. Consumers don't want that level of data, and companies don't want to expose system functions. To this end, explain-ability needs to conform to what companies are willing, or can be reliably compelled to give up without stalling and legal maneuvering. Global explain-ability will be a future goal due to the small chance that companies will consent to high explain-ability. Future works can focus on processing different data types such as images, videos, and music, and improving the speed and correlation of automated metrics with human evaluation.

REFERENCES

[1] Z. Fayyaz, M. Ebrahimian, D. Nawara, A. Ibrahim, and R. Kashef, "Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities," *Applied Sciences*, vol. 10, no. 21, p. 7748, Jan. 2020, doi: 10.3390/app10217748.

[2] X. Chen, Y. Zhang, and J.-R. Wen, "Measuring 'Why' in Recommender Systems: a Comprehensive Survey on the Evaluation of Explainable Recommendation." arXiv, Feb. 13, 2022. doi: 10.48550/arXiv.2202.06466.

[3] A. Vultureanu-Albişi and C. Bădică, "Recommender Systems: An Explainable AI Perspective," in *2021 International Conference on INnovations in Intelligent Systems and Applications (INISTA)*, Aug. 2021, pp. 1–6. doi: 10.1109/INISTA52262.2021.9548125.

[4] K. Zhou, Y. Zhou, W. X. Zhao, X. Wang, and J.-R. Wen, "Towards Topic-Guided Conversational Recommender System." arXiv, Nov. 02, 2020. doi: 10.48550/arXiv.2010.04125.

[5] S. Guo, S. Zhang, W. Sun, P. Ren, Z. Chen, and Z. Ren, "Towards Explainable Conversational Recommender Systems," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2023, pp. 2786–2795. doi: 10.1145/3539618.3591884.

[6] S. Moon, P. Shah, A. Kumar, and R. Subba, "OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 845–854. doi: 10.18653/v1/P19-1081.

[7] S. A. Hayati, D. Kang, Q. Zhu, W. Shi, and Z. Yu, "INSPIRED: Toward Sociable Recommendation Dialog Systems," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 8142–8152. doi: 10.18653/v1/2020.emnlp-main.654.

[8] R. Li, S. Kahou, H. Schulz, V. Michalski, L. Charlin, and C. Pal, "Towards Deep Conversational Recommendations." arXiv, Mar. 04, 2019. doi: 10.48550/arXiv.1812.07617.

[9] B. Wen, Y. Feng, Y. Zhang, and C. Shah, "ExpScore: Learning Metrics for Recommendation Explanation," in *Proceedings of the ACM Web Conference 2022*, in WWW '22. New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 3740–3744. doi: 10.1145/3485447.3512269.

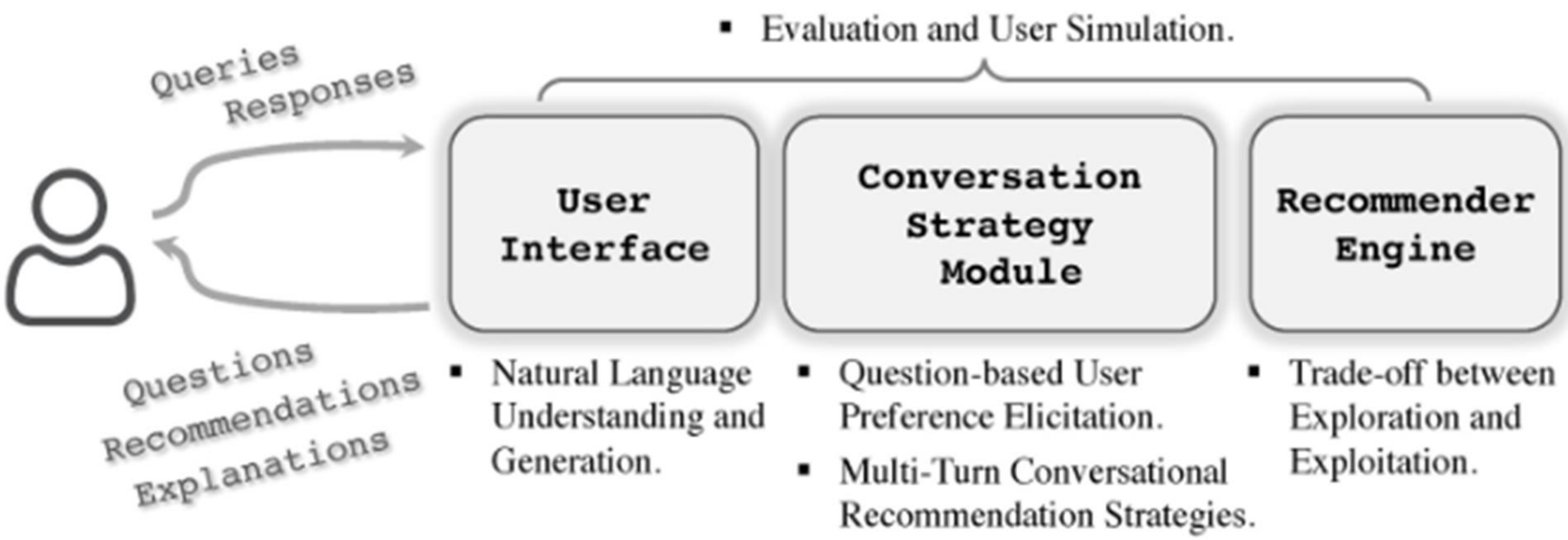
[10] Z. Chen *et al.*, "Towards Explainable Conversational Recommendation," presented at the Twenty-Ninth International Joint Conference on Artificial Intelligence, Jul. 2020, pp. 2994–3000. doi: 10.24963/ijcai.2020/414.

A Brief Survey of Offline Explainability Metrics for Conversational Recommender Systems

Joseph May, Khem Poudel

Middle Tennessee State University

Department of Computer Science



- Conversational Recommender Systems
- Modified search process
- System acts as a recommender

Background

Current Metrics

- BLEU
- ROGUE
- METEOR
- Perplexity
- N-gram
- Pretraining Deep Learning Regimes
  - BERT
  - ChatGPT

- Word Overlap
- Language / Contextual understanding
- Tools / Metrics are not yet refined for evaluating recommendations in a conversational setting

RECOEXP and Recommendation Factors

Factors of Explainability	Definition	Calculation
Relevance	If the recommendation is relevant to the query	Semantic similarity between explanation and item review (cosine similarity)
Length	How long the explanation is	Number of words after stop words are removed
Readability	How easy the recommendation is to read	Flesch-Kincaid readability test
Word Importance	The importance of words in the recommendation	Inverted term frequency
Repetition	How many duplicate segments are in a sentence	Number of repeated words once stop word have been removed
Subjectivity	If the recommendation includes personal opinions and emotion	TextBlob
Polarity	Confidence level that the recommendation is positive or negative	TextBlob
Grammatical Correctness	Misspelled words and incorrect usage of language	Python Language Tool
Feature Appearance	If an explanation captures item features	Counting how many words the recommendation includes of an item's features.

Proof of concept model and results

Used BERT and relevance factor to classify explanations in the E-redial dataset

We use the bert uncased model, and crossentropy as our loss function.

We assign conversations cosine similarity scores between what the seeker asked for and what the system recommended, scores >= .87 are classified as good recommendations.

Achieved 75% accuracy

E-redial dataset example:

Table 3: Sample conversation from validation set.

SEEKER:	hi ! i 'm looking for an action filled movie similar to jurassic park
HUMAN:	ok , have you seen jurassic world ?
HRED:	hi ! i love that one ! have you seen the conjuring or the conjuring 2 ?
OURS:	hi what kind of movies do you like ?
SEEKER:	yes i loved that movie as well . are there any good movies without dinosaurs ?
HUMAN:	have you seen king kong ?
HRED:	no , i have n't seen it yet . have you seen the avengers ? it 's about a couple of friends who are their own .
OURS:	i have n't seen that one . i will have to check it out . have you seen jurassic park
SEEKER:	yes ! that was another great giant monster movie :)
HUMAN:	what about jaws & amp ; jaws 2 ? oldies but goodies scary action ? suspense gives me the rumblies
HRED:	i have n't seen that one yet but i heard it was good . i have n't seen it yet . i have n't seen it yet . i have n't seen it yet .
OURS:	i have n't seen that one yet . i will have to check it out . what about walking with dinosaurs ?
SEEKER:	i absolutely loved the first one . one of my favorite movies ever :) honestly i ca n't say i remember much about the second one . jaws 3-d was quite funny and jaws : the revenge was just weird . i do love suspense too ...
HUMAN:	i like suspense but sometimes i can barely handle it ! it gives me anxiety ! lol
HRED:	no , i have n't seen it yet . i will have to check it out . thanks for the chat .
OURS:	i have n't seen that one . i will have to check it out . i hope i helped .
SEEKER:	that 's understandable . as long as it 's not too gory i do n't generally have a problem with those things .
HUMAN:	well , it was great chatting with you ! have a great one !
HRED:	i 'm glad i was able to help you today
OURS:	have you seen immortals ? it 's a war movie but it 's pretty good
SEEKER:	you too ! thanks !

References

[1] Z. Fayyaz, M. Ebrahimian, D. Nawara, A. Ibrahim, and R. Kashef, "Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities," *Applied Sciences*, vol. 10, no. 21, p. 7748, Jan. 2020, doi: 10.3390/app10217748.

[2] X. Chen, Y. Zhang, and J.-R. Wen, "Measuring 'Why' in Recommender Systems: a Comprehensive Survey on the Evaluation of Explainable Recommendation." arXiv, Feb. 13, 2022. doi: 10.48550/arXiv.2202.06466.

[3] A. Vultureanu-Albişi and C. Bădică, "Recommender Systems: An Explainable AI Perspective," in *2021 International Conference on INnovations in Intelligent Systems and Applications (INISTA)*, Aug. 2021, pp. 1–6. doi: 10.1109/INISTA52262.2021.9548125.

[4] K. Zhou, Y. Zhou, W. X. Zhao, X. Wang, and J.-R. Wen, "Towards Topic-Guided Conversational Recommender System." arXiv, Nov. 02, 2020. doi: 10.48550/arXiv.2010.04125.

[5] S. Guo, S. Zhang, W. Sun, P. Ren, Z. Chen, and Z. Ren, "Towards Explainable Conversational Recommender Systems," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2023, pp. 2786–2795. doi: 10.1145/3539618.3591884.

[6] S. Moon, P. Shah, A. Kumar, and R. Subba, "OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 845–854. doi: 10.18653/v1/P19-1081.

References

[7] S. A. Hayati, D. Kang, Q. Zhu, W. Shi, and Z. Yu, “INSPIRED: Toward Sociable Recommendation Dialog Systems,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 8142–8152. doi: 10.18653/v1/2020.emnlp-main.654.

[8] R. Li, S. Kahou, H. Schulz, V. Michalski, L. Charlin, and C. Pal, “Towards Deep Conversational Recommendations.” arXiv, Mar. 04, 2019. doi: 10.48550/arXiv.1812.07617.

[9] B. Wen, Y. Feng, Y. Zhang, and C. Shah, “ExpScore: Learning Metrics for Recommendation Explanation,” in *Proceedings of the ACM Web Conference 2022*, in WWW ’22. New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 3740–3744. doi: 10.1145/3485447.3512269.

[10] Z. Chen *et al.*, “Towards Explainable Conversational Recommendation,” presented at the Twenty-Ninth International Joint Conference on Artificial Intelligence, Jul. 2020, pp. 2994–3000. doi: 10.24963/ijcai.2020/414.