Appendix: The Problem of Coherence in Natural Language Explanations of Recommendations

Jakub Raczyński[a], Mateusz Lango[a,b;*] and Jerzy Stefanowski[a]

[a] Poznan University of Technology, Faculty of Computing and Telecommunications, Poznan, Poland
[b]Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic
ORCiD ID: Mateusz Lango https://orcid.org/0000-0003-2881-5642,
Jerzy Stefanowski https://orcid.org/0000-0002-4949-8271

A Limitations

The proposed explainable recommendation method still has some limitations.

First, the datasets used to train explainable recommenders are constructed on the basis of user reviews, but as noted by Ni at al. [5], a significant portion of the review text may be of little relevance to the user's decision-making process. It may contain general statements ("very good") or discuss the user's personal experiences. Consequently, the explanations generated by the current methods can be interpreted as "If you were to use/buy this item, you would say that ..." which leads to providing subjective opinions as explanations whereas it would be more desirable for a recommender system to be a rational agent that provides reasons why an item is (not) recommended. Little attention has been paid to the desired form of a textual explanation, its type (abductive, counterfactual,...), and possible relations to argumentation mining [4]. This, combined with the additional motivation provided by the current paper (i.e. the lack of coherence in the commonly used datasets), encourages future work on collecting new datasets with reference explanations.

Second, one of the properties of a good explanation is its faithfulness, i.e. the representation of the true reasoning process behind the prediction [1]. The recent work [6] provides some experimental evidence that currently proposed methods for explainable recommendations in natural language (such as NRT or PETER) fail to provide faithfulness, plausibility, and semantic coherence at a sufficient level. Despite the fact that the current work is a step forward toward more faithful and plausible explanations, it still needs further investigation, especially in the direction of semantic coherence of explanations since explanations that seem plausible but are not faithful are misleading the users.

B Frequently Asked Questions (FAQ)

1. *Which kind of explanation is the recommender system expected to provide (e.g. abductive, counterfactual, contrastive, causal)?*
   The type of explanations provided by the system derives from its dependence on the training datasets commonly used in related works. The explanations provided can be classified as abductive - see details in [8] introducing datasets.
2. *Which aspects of the user can the proposed model represent?*
   Each user is represented by an embedding in the neural network that is randomly initialized and learnt together with recommendation and explanation generation tasks. Thus, the system automatically encodes in the user embedding its preference profile (to perform recommendation) and partially the characteristics of its (textual) opinions to improve the explanations.

* Corresponding Author. Email: mlango@cs.put.edu.pl.

C Semantic coherence vs rating-explanation coherence

Recently, the quality of the explanations provided by PETER, Att2Seq, and NRT was further experimentally evaluated by Xie et al. [6]. They found that these methods fail to provide highly faithful (i.e. reflecting the model's decision process for rating prediction) and semantically coherent explanations (their meaning should capture the user's true interest in the product).

It is worth noting that semantic coherence is different from the rating-explanation coherence highlighted in this paper. The former can be assessed by comparing the meaning of the provided explanation with a reference explanation, while the latter requires comparing the meaning of the explanation with the predicted rating score. For example, the explanation "great hotel" for rating 1 (out of 5) is semantically coherent with the reference "great hotel", but not coherent with the rating. Contrary, "great views" for a rating of 5/5 and the same reference would not be semantically coherent but would be rating-explanation coherent.

The issues of factuality, semantic coherence (and more broadly hallucinations) of text generated by neural methods are widely discussed in NLG literature [2], together with some mitigation methods. The semantic coherence with the reference can be measured with e.g. BERTScore [7], but the rating-explanation coherence can not be evaluated with existing measures. The work [6] also does not propose any mitigation methods.

We evaluated the semantic coherence of provided explanations with BERTScore measure [7], as suggested by [6]. The selected semantic coherence metric (BERTScore) reported in Tab 1 indicates similar semantic coherence of explanations generated by CER and PETER+.

Table 1. The evaluation of semantic coherence of explanations provided by CER and PETER+

Dataset	Model	BERTScore		
		Precision	Recall	F1
TripAdvisor	PETER+	0,900	0,882	0,891
	CER	0,902	0,881	0,891
Amazon	PETER+	0,892	0,863	0,877
	CER	0,882	0,862	0,872
Yelp	PETER+	0,894	0,873	0,883
	CER	0,894	0,873	0,883

D Details of preliminary study

In the introduction, we mention that the motivation for our research was a preliminary human evaluation of explanation generated by PETER+[3]. The study was performed on 60 instances (20 examples randomly drawn from each dataset) and all the annotations were performed only by one annotator (the paper's first author). The summary of the results is presented in Table 2.

Table 2. The results of a preliminary manual evaluation of PETER+ explanations.

Type of problem	No. of occurrences
Lack of consistency between explanation and recommendation	15
Failure to justify the opinion	8
Explanation focused on individual experience/assessment	7
Lack of context required to understand the explanation	6
Unnatural truncation of a sentence	6
Repetition of n-grams	4
Occurrence of the UNK token	2
Ungrammatical, incomprehensible bundle of words	2

E Datasets

To verify the utility of the proposed Coherent Explainable Recommender approach, we conducted experiments on three freely available datasets[1] which basic statistics can be found in Tab 3.

Table 3. Basic characteristics of used datasets

	TripAdvisor	Amazon	Yelp
# of users	9765	7506	27147
# of items	628	736	20266
# of explanations	320023	441783	1293247
Avg. # of expl. / user	32.77	58.86	47.64
Avg. # of expl. / item	50.96	60.02	63.81
Avg. # of words / explanation	13.01	14.14	12.32

F Details on experimental setup

The values of hyperparameters used in the experiments are presented in Table 4.

[1] https://tinyurl.com/yd8xtvam

Table 4. Hyperparameters of coherence classification models, determined in cross-validation process

	Yelp	Amazon	TripAdvisor
Learning rate	0.001	0.001	0.001
Optimizer	Adam	Adam	Adam
L2 regularization weight	0.05	0.05	0.1
Minority class weight	2.5	1.8	2
Number of epochs	150	100	150

References

[1] Alon Jacovi and Yoav Goldberg, 'Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, (July 2020). Association for Computational Linguistics.

[2] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung, 'Survey of hallucination in natural language generation', *ACM Comput. Surv.*, **55**(12), (mar 2023).

[3] Lei Li, Yongfeng Zhang, and Li Chen, 'Personalized transformer for explainable recommendation', in *Proceedings of the 59th Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 4947–4957, (2021).

[4] Marco Lippi and Paolo Torroni, 'Argumentation mining: State of the art and emerging trends', *ACM Trans. Internet Technol.*, **16**(2), (mar 2016).

[5] Jianmo Ni, Jiacheng Li, and Julian McAuley, 'Justifying recommendations using distantly-labeled reviews and fine-grained aspects', in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 188–197, (2019).

[6] Zhouhang Xie, Julian McAuley, and Bodhisattwa Prasad Majumder. On faithfulness and coherence of language explanations for recommendation systems, 2022.

[7] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi, 'Bertscore: Evaluating text generation with bert', *arXiv preprint arXiv:1904.09675*, (2019).