On Faithfulness and Coherence of Language Explanations for Recommendation Systems

Zhouhang Xie  Julian McAuley  Bodhisattwa Prasad Majumder
University of California, San Diego
{zhx022, jmcauley, bmajumde}@ucsd.edu

Abstract

Reviews contain rich information about product characteristics and user interests and thus are commonly used to boost recommender system performance. Specifically, previous work show that jointly learning to perform review generation improves rating prediction performance. Meanwhile, these model-produced reviews serve as recommendation explanations, providing the user with insights on predicted ratings. However, while existing models could generate fluent, human-like reviews, it is unclear to what degree the reviews fully uncover the rationale behind the jointly predicted rating. In this work, we perform a series of evaluations that probes state-of-the-art models and their review generation component. We show that the generated explanations are brittle and need further evaluation before being taken as literal rationales for the estimated ratings.

1 Introduction

Product reviews capture rich information about user preferences and thus improve recommender system performance (McAuley et al., 2012; McAuley and Leskovec, 2013; Zheng et al., 2017; Tay et al., 2018; Chen et al., 2018; Pugoy and Kao, 2020, 2021). Meanwhile, advancements in text generation enable generating realistic synthetic reviews conditioning on user and item identifiers, as well as additional features such as historical reviews (Li and Tuzhilin, 2019), product metadata (Ni and McAuley, 2018; Dong et al., 2017), knowledge graph embedding (Li et al., 2021a), and sometimes the rating itself (Chen et al., 2021). Recently, there has been increasing interest in coupling rating estimation and review generation, treating generated reviews as explanations for model recommendations (Ni et al., 2017; Sun et al., 2020; Li et al., 2020, 2021b; Hada et al., 2021).

In the current literature, the quality of the generated explanations are usually measured by perplexity and overlapping-based metrics such as Distinct-N (Li et al., 2016), Rouge score (Lin, 2004), and BLEU score (Papineni et al., 2002) with respect to the ground truth reviews. However, while these evaluations measure fluency and word-overlapping, they do not warrant the quality of the generated reviews’ quality as explanations.

Specifically, overlapping metrics overlook two core aspect of natural language explanations (NLEs): (1) faithfulness, how truthfully do the generated explanations reflect the decision process for the models rating prediction, and (2) semantic coherence, how well the model capture the users’ true interest towards the product. To highlight the potential issue associated with current evaluation, consider the review text for a restaurant "I love this hotel because it has great service" with a rating of 5, where the explanation generated is "I love this hotel because it has great cookies" with the correct predicted rating. The generated explanation deviates from the ground truth sentence by only one word, yet completely changes the rationale for the rating. However, the currently widely used automatic metrics will still assign a high score to the generated review. Further, there is no guarantee that even cookie is truly accountable for the predicted rating.
To address these discrepancies, we argue that NLEs for recommendation systems should be evaluated as explanations, similar to NLEs in NLP tasks. In this work, we probe review-as-explanation models in explainable recommendation literature. Our results show a concerning trend that current models struggle to produce reviews that are semantically coherent with the ground truth reviews, and are inconsistent with the explanation they produce. We encourage researchers and practitioners to consider the explanations generated by the models and not just the reviews. Better evaluations could lead to deeper understanding of capabilities of these generated rationales, and foster more trustworthy explainable recommendation systems.

2 Problem Definition and Models

This section is organized as follows: we first cover the task of joint review-rating generation, then introduce the models used in our experiments.

2.1 Problem Setup

Given a user \( u \) and an item \( i \), the task of joint review-rating generation aims at predicting an associated rating \( \hat{r} \) as well as a natural language explanation \( \hat{e} \). During training, the model jointly minimizes the negative log likelihood (NLL) of ground truth reviews in the corpus, as well as the mean squared error (MSE) of their associated rating.

2.2 Models

We compare four recent models in the literature that covers a variety of commonly used architectures in natural language generation: Att2Seq (Dong et al., 2017), NRT (Li et al., 2017), PEPLER (Li et al., 2022), and PETER (Li et al., 2021b). Among the models, Att2Seq and NRT are based on Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), while PEPLER combines a pre-trained GPT-2 model (Radford et al., 2019) with prompt tuning. Finally, PETER adopts non-auto-regressive transformer architecture.

Meanwhile, since Li et al. noted incorporating specific content words significantly improves generation quality, we follow the original paper and condition the model on a content word, denoted by PETER\(_{cond}\). Note that following the previous implementation, the aspect word in the dataset is extracted from ground truth review, giving PETER\(_{cond}\) an unfair advantage. We thus use PETER\(_{cond}\) as an upper-bound baseline.

2.3 Datasets

We conduct our experiments on Yelp\(^2\) (Y.), TripAdvisor\(^3\) (T.), and Movies and TV category from Amazon dataset (He and McAuley, 2016) (M.). These are standard dataset commonly used to benchmark joint review-rating estimation models (Li et al., 2020).

3 Evaluating Faithfulness and Semantic Coherence

We generate 10,000 explanations for each model on each dataset, and perform a set of evaluations as described in this section.

3.1 Faithfulness

When reviews are treated as natural language explanations, joint review-rating prediction models could be categorized into self-rationalizing models. Jacovi and Goldberg argues that the quality of NLE should be evaluated by both their plausibility, how convincing the explanations are to

\(^2\)https://www.yelp.com/dataset/challenge

\(^3\)https://www.tripadvisor.com

\(^4\)NLR in Jacovi and Goldberg’s work
humans, and faithfulness, how truthful they reflect the models’ decision process. We focus on model faithfulness in this section.

By definition, a faithful explanation will truthfully represent the decision process of a model. However, directly measuring faithfulness is infeasible due to the black-box nature of deep neural networks. We instead design a set of proxy tasks that test unfaithful behavior of joint review-rating estimation models.

**Adversarial Invariance Ratio (AIR).** Since the explanation generation by the model is representative of the model’s belief of the reasons behind the rating prediction, we argue that such belief must be robust to sentiment perturbations. In other words, assume a model generates a sequence $\hat{e}_{u,i}$ as an explanation, the sentiment-negated counter explanation $\neg\hat{e}_{u,i}$ should not receive a higher likelihood (lower perplexity) than the original review. Illustration of selected sub-experiments are as shown in Figure 2.

Concretely, we take 4 or 5-star (positive sentiment) ratings from the test set and rewrite their sentiment to negative using a pretrained BART model, and let the target model rank the ground truth and rewritten review with perplexity. We mark the models’ decision as flipped if it assigns lower perplexity to the rewritten review with the negated sentiment. In this case, the model’s explanation is thus unfaithful. Note that this means a random baseline would achieve 50 percent in AIR.

**Mean Reciprocal Rank against Alternative Explanations (MRR-AE).** As pointed out in Jacovi and Goldberg, a model is unfaithful if it provides a different interpretation for the same decision by the same model. That is, the model should be able to differentiate its generated review from other candidate explanations. Following this intuition, we argue that the model should have the ability to pick out its generated review from other reviews, such as random reviews drawn from the dataset or adversarially constructed ones, as shown in Figure 2.

To measure this, we sample 100 reviews randomly from the test dataset for each gold review, and replace the aspect in the sampled sentences with the aspect covered by the ground truth. We then let the target model rank the 100 sentences along with the gold review with perplexity score and measure its performance with mean reciprocal rank (MRR). The random baseline for MRR-AE is thus around 5 percent.

**Text-label Agreement Error (TLAE).** As faithful explanations, the generated reviews should strongly correlate with predicted ratings. To measure this, we train a BERT (Vaswani et al., 2017) based auxiliary rating regressor based on only user reviews on the training set of the models being evaluated. At test time, we measure the Mean Squared Error of the auxiliary predictor on generated reviews and regressor-predicted ratings.

### 3.2 Semantic Coherence

Traditional evaluation metrics use in the literature focuses on word-overlapping, and thus would be insensitive to mismatched content words. To address this issue, we argue that generated explanations should be evaluated by its semantic coherence. Concretely, we adopt two recent, state-of-the-art semantic evaluation metrics: BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021). Further, we use a pre-trained entailment model to check whether the generated content entails the ground truth review. We report the percentage of entailment (Entail), where a good model should have high ratio.

### 4 Empirical Results

**Faithfulness.** Our main evaluation results are as shown in table 1. While model-generated explanation generally matches the predicted rating (TLAE), most models have near random performance against sentiment perturbations (AIR). Meanwhile, although PEPLER is the most robust to sentiment perturbation, it is not as competitive as other models in terms of recommendation performance (RMSE). This illustrates the potential risk of powerful language models giving a false sense of explainability simply due to their strong language modeling ability. In other words, the explanations are plausible but not faithful under (Wiegreffe et al., 2021)”s framework.

**Semantic Coherence.** From coherence evaluations, we could see the model generally struggle to capture the exact aspect that the user cares about, resulting in a low entailment ratio (Entail) compared to PETER$_{cond}$. This can be corroborated by

---

5$^5$dapang/yelp_pos2neg_lm_bart_large from huggingface.

6BERTScore is cosine similarity-based (larger means better) and BARTScore is NLL based (smaller means better).
Table 1: Evaluation results on the datasets. **Bold text** denotes best performance (except PETER_{cond} model), **bold and underlined** text means the performance exceeded PETER_{cond} model, which has access to ground truth aspect covered in review. M., T. and Y. denotes Amazon Movies and TV, TripAdvisor, and Yelp dataset, respectively.

| Metric | Faithfulness | Semantic Coherence | Rec. |
| --- | --- | --- | --- |
| | AIR↑ | MRR-AE↑ | TLAE↓ | Entail↑ | BERTS↑ | BARTS↓ | RMSE↓ |
| Model | M. | T. | Y. | M. | T. | Y. | M. | T. | Y. | M. | T. | Y. | M. | T. | Y. | M. | T. | Y. | M. | T. | Y. |
| Att2Seq | 14.6 | 43.5 | 47.9 | 23.5 | 19.2 | 23.5 | n/a | n/a | n/a | 6.6 | 2.7 | 5.6 | 0.08 | 0.16 | 0.10 | 5.95 | 5.97 | 5.97 | n/a | n/a | n/a |
| NRT | 55.8 | 43.0 | 55.8 | 15.1 | 18.2 | 22.6 | **1.39** | 0.84 | **1.18** | 3.2 | 1.5 | 1.4 | -0.21 | -0.15 | -0.19 | 6.54 | 6.72 | 6.6 | 0.95 | 0.79 | 1.01 |
| PETER | 60.3 | 47.3 | 49.4 | 18.1 | **22.0** | 26.3 | **1.39** | 0.84 | **1.18** | 8.2 | 3.5 | 7.9 | 0.11 | 0.17 | 0.12 | 5.95 | 5.96 | 5.88 | 0.95 | 0.81 | 1.01 |
| PEPLER | 70.0 | **63.3** | 21.0 | 16.6 | 16.3 | 7.3 | **1.39** | 0.84 | **1.18** | 4.4 | 3.4 | 4.6 | 0.13 | 0.19 | 0.23 | 5.95 | 5.93 | 6.99 | 1.25 | 1.71 | 1.69 |
| PETER_{cond} | 19.1 | 52.0 | 55.7 | 27.9 | 20.2 | 27.9 | **1.39** | 0.76 | **1.18** | 27.7 | 24.2 | 24.3 | 0.18 | 0.30 | 0.25 | 5.47 | 5.93 | 5.17 | 0.95 | 0.81 | 1.02 |

6 Implications and Recommendations

**Calibrate user expectations.** End users often trust that algorithm explanations are faithful (Jin et al., 2022) and hope they could receive reliable explanations (Lakkaraju et al., 2022), we recommend practitioners inform the users about the probabilistic nature of generated reviews as explanations. For example, if a system generates an explanation related to ‘bbq pork ribs’, it would be more of an indication of user’s interest in smokehouse cuisines rather than the dish itself.

**Develop hybrid systems.** While PETER_{cond} acts as an “unfair” baseline in our experiment section, the model would be a great tool in a larger pipeline, where users actively provide feedback. Similarly, recent works starts to explore natural language as an **interface** in pipeline systems for non-language based explanations (Slack et al., 2022). We encourage the community to consider conditional and pipeline systems in addition to end-to-end models.

**Better evaluations.** We note that although evaluations generally depend on the use-cases of the model, and a powerful model does not necessarily need to satisfy faithfulness, plausibility, and semantic coherence simultaneously, it is advisable to perform beyond-overlapping evaluations before assuming the literal validity of generated recommendation NLRs.

6.1 Related Works

**Faithfulness of Natural Language Explanations.** Jacovi and Goldberg argue that the quality of NLE from this class of models should be evaluated by their faithfulness, how truthful (and thus consistent) do they reflect the models’ decision

---

7Jin et al. study was based on medical image
process. Under this setup, our work’s evaluation differs from existing evaluations in the literature in that we clearly distinguish faithfulness from general language quality. Wiegreffe et al. approach this problem by measuring the connection between labels and explanations, yet their evaluation does not take the semantics of the generated explanation itself into account.

**Analyzing Model Decision in NLP.** Another related line of work is analysis of model decision boundaries. Common strategies usually involve adversarially probing the model, such as using counterfactual data (Wu et al., 2021), contrast sets (Gardner et al., 2020) and semantically preserving modifications of sentence characteristics (Ribeiro et al., 2018, 2020; Longpre et al., 2021). Our work deviates from prior works as we establish connections of adversarial evaluation directly with model faithfulness.

7 Conclusions

Joint review-rating prediction models could generate high-quality reviews while producing accurate rating estimations. However, it is unclear whether the generated reviews could be leveraged as precise recommendation rationales. We conduct a set of evaluation that benchmark faithfulness and semantic coherence of state-of-the-art models. We show more careful evaluations are needed before generated reviews could be taken as fully accountable explanations.

**References**

Samuel Carton, Surya Kanoria, and Chenhao Tan. 2022. What to learn, and how: Toward effective learning from rationales. In Findings of ACL.

Chong Chen, Min Zhang, Yiqun Li, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. WWW.

Huimin Chen, Yankai Lin, Fanchao Qi, Jinyi Hu, Peng Li, Jie Zhou, and Maosong Sun. 2021. Aspect-level sentiment-controllable review generation with mutual learning framework. In AAAI.

Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to generate product reviews from attributes. In EACL.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannanah Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models’ local decision boundaries via contrast sets. In Findings of EMNLP.

Deepshe V. Hada, M Vijaiakumar, and Shirish K. Shevade. 2021. Rexplug: Explainable recommendation using plug-and-play language model. SIGIR.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. WWW.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In ACL.

Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. 2022. Evaluating explainable AI on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements? In AAAI. AAAI Press.

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking explainability as a dialogue: A practitioner’s perspective. *arXiv preprint arXiv:2202.01875*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In NAACL-HLT, San Diego, California.

Junyi Li, Wayne Xin Zhao, Zhicheng Wei, Nicholas Jing Yuan, and Ji-Rong Wen. 2021a. Knowledge-based review generation by coherence enhanced text planning. *SIGIR*.

Lei Li, Yongfeng Zhang, and Li Chen. 2020. Generate neural template explanations for recommendation. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM ’20, page 755–764, New York, NY, USA. Association for Computing Machinery.

Lei Li, Yongfeng Zhang, and Li Chen. 2021b. Personalized transformer for explainable recommendation. In ACL-IJCNLP.

Lei Li, Yongfeng Zhang, and Li Chen. 2022. Personalized prompt learning for explainable recommendation. *arXiv preprint arXiv:2202.07371*.

Pan Li and Alexander Tuzhilin. 2019. Towards controllable and personalized review generation. In EMNLP-IJCNLP, Hong Kong, China.

Piji Li, Zhihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. *SIGIR*. 

