Plug and Play Explainable Recommendations

Shirish K. Shevade
Vijaikumar M.
Deepesh Hada
shirish@iisc.ac.in
vijaikumar@iisc.ac.in
deepeshhada@iisc.ac.in
Indian Institute of Science
Bangalore, India

ABSTRACT

Explainable Recommendations provide the reasons behind why an item has been recommended to the user, leading to increased user satisfaction and persuasiveness. We propose a novel way of explaining recommendations by generating text reviews on behalf of the user for an item that is recommended to him. These generated sentences have resemblance to the way the user has written his past reviews, while also talking about the concept of the item being recommended.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Natural language processing**.

KEYWORDS

Explainable Recommendations, PPLM, Neural Networks, Controlled Text Generation

ACM Reference Format:

Shirish K. Shevade, Vijaikumar M., and Deepesh Hada. 2018. Plug and Play Explainable Recommendations. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/1122445.1122456

1 MODEL OVERVIEW

To recommend items that a user has never interacted with, Recommendation Systems predict ratings corresponding to a user-item pair. Evidently, items having a higher predicted rating is recommended to the user. But not just the past ratings, models can exploit another rich source of information, the *user reviews*. A single review contains much more information about the concerned user and item than a single rating. Intuitively, reviews also *explain* why the user had bought the item, and what he feels about the item after buying it.

A significant amount of work has been done in the field of controlled text generation. Controlled text generation models need the output sentences to be fluent, while also satisfying the controlled attribute being passed to them. Most of these models are very complex and have a large number of parameters to tune/train, which is a very costly and time-consuming task. An alternative and a simple approach was proposed by [2]. The Plug and Play Language Models work on top of a mammoth language model like GPT-2 [5], but have a shallow network which is very easy to train. The mammoth ensures fluency in the generated sentence, while the lightweight attribute model satisfies the attribute (like sentiment).

The ratings can be thought of as sentiments; higher the rating, more positive is the sentiment. To generate new sentences based on a given sentiment, the trained PPLM model accepts two arguments: **sentiment** and **conditional text**. The sentiment comes from the ratings predicted by the **Rating Predictor**, which forms the first module of the framework. The second module involves generating an appropriate **conditional text sentence**. This conditional text is used to kick-start the **PPLM**, and has around 15-20 words.

We now meticulously describe the three modules:

(1) **Rating Predictor:**
It was shown in a recent work by [6] that using text reviews as regularizers outperform many complex models. To generate explainable recommendations, we first build a hybrid network, inspired by [3], while using the text reviews as a regularizer to learn better user and item embeddings. This is done by putting up two parallel neural networks, both of which accept the same learnable user and item embeddings as inputs. The first network is a simple MLP which does regression over the ratings. The parallel network acts as the regularizer network. It accepts the same user and item embeddings that were fed to the first MLP, and tries to predict the corresponding review text which is represented as a vector.
To represent the review text as a fixed-dimensional vector, we have used the Deep Averaging Network [4] based Universal Sentence Encoder [1] to get a 512-dimensional vector. This regularizing model outputs a 512-dimensional vector, and tries to make the prediction as close as possible to the review encoded by the Universal Sentence Encoder. Note that this parallel network appears only during the training phase, to learn better user and item embeddings. During the testing phase, only the first MLP is used to predict the ratings.
The loss functions for both of these networks is the MSE Loss. The overall loss of this module is given by a linear

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

combination of the losses of the two networks, *i.e.*,

$$L_{mod_1} = L_1 + \lambda L_2 \tag{1}$$

where λ is a hyperparameter.

(2) **Conditional Text Generator:**
From the given set of training reviews, this module identifies the most important reviews corresponding to the user-item pair in question. For each such pair, we have two sequences/list of reviews from the training set. The first list corresponds to the reviews that the user has written, and the second list corresponds to the reviews the item has received. Some recent models like MPCN [7] use the attention mechanism to achieve this.
*This is still **Work in Progress**. This module may be clubbed with the first module later.*

(3) **Generating Controlled Text Reviews via PPLM:**
We first train the attribute model that belongs to the PPLM. This attribute model is a sentiment classifier, which accepts a textual sentence as input, and predicts a sentiment class. In our case, we treat this attribute model as a binary (0/1) sentiment predictor. The same dataset which was used in the first module must be used here.
In our experiments, we found out that a binary sentiment predictor resulted in generating remarkable sentences that satisfy the input sentiment. However, in most datasets, the ratings are present in the range of 1-5. We converted these to binary ratings by mapping $\{1, 2\} \rightarrow 0$ (indicating negative sentiment) and $\{4, 5\} \rightarrow 1$. Reviews corresponding to the neutral rating (3) were removed as they tend to induce ambiguity, leading to slightly poorer performance of the attribute model.
The PPLM model accepts the rating predicted by the first module as a sentiment and the conditional text from the second module. It then generates multiple candidate sample reviews which entail the conditional text and satisfy the input sentiment.

REFERENCES

[1] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *CoRR* abs/1803.11175 (2018). arXiv:1803.11175 http://arxiv.org/abs/1803.11175

[2] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. arXiv:1912.02164 [cs.CL]

[3] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat Seng Chua. 2017. Neural Collaborative Filtering. *CoRR* abs/1708.05031 (2017). arXiv:1708.05031 http://arxiv.org/abs/1708.05031

[4] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, 1681–1691. https://doi.org/10.3115/v1/P15-1162

[5] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).

[6] Noveen Sachdeva and Julian McAuley. 2020. How Useful are Reviews for Recommendation? A Critical Review and Potential Improvements. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Jul 2020). https://doi.org/10.1145/3397271.3401281

[7] Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Multi-Pointer Co-Attention Networks for Recommendation. *CoRR* abs/1801.09251 (2018). arXiv:1801.09251 http://arxiv.org/abs/1801.09251