ExpScore: Learning Metrics for Recommendation Explanation

Bingbing Wen
University of Washington
Seattle, WA, US
bingbw@uw.edu

Yunhe Feng
University of Washington
Seattle, WA, US
yunhe@uw.edu

Yongfeng Zhang
Rutgers University
New Brunswick, NJ, US
yongfeng.zhang@rutgers.edu

Chirag Shah
University of Washington
Seattle, WA, US
chirags@uw.edu

ABSTRACT

Many information access and machine learning systems, including recommender systems, lack transparency and accountability. High-quality recommendation explanations are of great significance to enhance the transparency and interpretability of such systems. However, evaluating the quality of recommendation explanations is still challenging due to the lack of human-annotated data and benchmarks. In this paper, we present a large explanation dataset named *RecoExp*, which contains thousands of crowdsourced ratings of perceived quality in explaining recommendations. To measure explainability in a comprehensive and interpretable manner, we propose *ExpScore*, a novel machine learning-based metric that incorporates the definition of explainability from various perspectives (e.g., relevance, readability, subjectivity, and sentiment polarity). Experiments demonstrate that *ExpScore* not only vastly outperforms existing metrics and but also keeps itself explainable. Both the *RecoExp* dataset and open-source implementation of *ExpScore* will be released for the whole community. These resources and our findings can serve as forces of public good for scholars as well as recommender systems users.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Evaluation of retrieval results**; • **Computing methodologies** → *Natural language generation*.

KEYWORDS

Metric, Evaluation, Explainable Recommendation

ACM Reference Format:
Bingbing Wen, Yunhe Feng, Yongfeng Zhang, and Chirag Shah. 2022. ExpScore: Learning Metrics for Recommendation Explanation. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3485447.3512269

1 INTRODUCTION

As explainable recommendation has drawn increasing attention in recent years [4, 14, 27], many studies explored the explanation generation for recommendation systems [9, 11, 26, 30]. However, existing task-agnostic text quality evaluation methods, such as BLEU [23], METEOR [3], and ROUGE [19], are not flexible or eligible

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04...$15.00
https://doi.org/10.1145/3485447.3512269

Figure 1: Recommendation explanation survey. MTurkers first read the description of a given item and then rate three explanations from 1 to 5.

to evaluate such explanations because they fail to consider the context of recommendation systems. According to [7, 18, 24, 28, 32, 33], explainable recommendations should serve to improve the transparency, persuasiveness, effectiveness, trustworthiness, efficiency, scrutability and user satisfaction of the recommendation systems. In addition, a good explanation should be easy to read (e.g., concise), consistent with the rating (consistency), and be sufficient for predicting users' preference on items (explainability) [29].

An ideal way of evaluating the explainability of machine generated explanations is through online user-study. Balog et al [13] measured recommendation explanation quality by collecting users judgment on seven pre-designed goals. Though such human-centric evaluation is a desirable way, it costs extensive labors and time, and is not always repeatable or scalable. In most cases, offline evaluation is a more usable solution for general research scenarios. The most commonly used metrics for evaluating machine generated explanation sentences are BLEU [23], METEOR [3] or ROUGE [19] scores, which consider the word-level precision and recall of sentences. They can reflect the quality of a generated sentence on readability. However, these measures do not consider how well a sentence can be used as an explanation.

To the best of our knowledge, a general and commonly accepted metric for explanation evaluation in recommendation systems is still missing. Substantive and foundational research often depends on solid evaluation metrics [1]. A lack of suitable metrics hinders our ability to assess the performance of explanation generation models and push them for further improvements. In this paper, we discuss the construction of a human-labeled dataset, *RecoExp*, built by asking users to rate the perceived quality of recommendation explanations. Specifically, we adopt Neural Template (NETE) method [16], a state-of-the-art explanation generation model, to create recommendation explanation candidates. Through exploring and analyzing *RecoExp*, we explicate vital factors that may affect human evaluations towards explainability of recommendation explanations. Based on these factors, we further develop *ExpScore*[1], an extendable and adaptable learning metric, to evaluate recommendation explanations. The main contributions of the paper are summarized as follows.

- We develop a new *RecoExp* dataset to facilitate the progress on the recommendation explanation evaluation. *RecoExp* is designed to work when no ground-truth explanation is available so as to alleviate ground truth dependency, which is closer to the real-world explainable recommendation scenarios.
- We present a novel machine learning-based metric *ExpScore* for evaluating recommendation explanations. Experiments show that *ExpScore* vastly outperforms existing metrics and correlates better with human judgments.
- We propose an interpretable and easily extendable factor-based framework for *ExpScore* that explores the definition of explainability from various perspectives. We also provide a comprehensive analysis of domain-independent explainability factors.

2 DATASET

We conducted an IRB-approved Amazon Mechanical Turk survey to collect a large dataset called *RecoExp*.

2.1 Survey Setup

Source data preparation We used the Movies and TV category of Amazon Review Dataset [21] as the data source for our survey. Specifically, we randomly extracted 634 product items with the required information about product name, description, and the corresponding human reviews (used as the reference corpus for calculating metrics such as BLEU). The solutions for generating textual explanations can be categorized as template-based [5, 6, 17] and generation-based [8, 18, 22]. The Neural Template (NETE) method [16] integrates template-based and generation-based approaches to make the explanation generation process more controllable, which is the state-of-the-art approach for recommendation explanation generation. Therefore, we adopted NETE [16], a state-of-the-art neural template explanation generation framework, to create three explanations for each product item.

Survey design Figure 1 shows a micro rating task example. Each micro rating task contains one product as a recommendation and three explanations for the recommendation. When doing each micro rating task, MTurk workers (MTurkers) were requested to read the product item name and description and evaluate three machine-generated explanations for their quality on a scale from 1 to 5, with 1 being the lowest quality and 5 being the highest quality. Personal information about the workers was not collected because it was not judged essential in this task, and this also helps to protect the workers' privacy. Each survey consists of ten consecutive rating tasks and one mandatory question of what factors contributed to the decision-making process. 634 products were randomly assigned to each survey. We assured that each product had received five responses which is the critical criterion of valid product responses to mitigate the subjectivity of rating scores. Participants were paid $2.00 USD for their participation. The average task completion time was 12.28 minutes. A response would not be considered valid if the micro task completion time was less than 5s.

[1]Code and dataset are available at https://github.com/bbwen/ExpScore

2.2 *RecoExp* Dataset

The collected *RecoExp* dataset contains 579 product items, 1,737 machine-generated explanations, and factors affecting ratings. We attempted to collect five responses for each explanation, considering that people might hold different opinions on the same explanation. Initially, 317 participants took our survey. A total of 288 participants provided valid responses for the 1,737 machine-generated explanations, reporting 8,685 explanation quality ratings. At the end of the survey, we asked an open question to the workers "what factors affected your ratings?" and 288 answers were collected.

Ratings Collection For rating tasks, we explore the distributions of rating scores and time cost from three perspectives (i.e., overall, by MTurker, by product item) The overall distributions aggregated all ratings (time cost) generated by all participants for all product items, while the MTurker (product item) distributions reported the average measurements per MTurker (product item). Surprisingly, the mean and median of all the three rating distributions in Figure 2 are above the average score of 3, indicating an acceptable or even good quality of machine-generated explanations. As expected, the rating distributions by MTurker and by product item in Figure 2 follow a normal distribution.

Factors Collection We employed a qualitative approach, based on open coding and constant comparison to understand factors affecting ratings from participants' perspectives. We conclude 10 categories: good grammar, length, no repetitions, making sense, expressing opinions, detail, relevance, and emotion. We illustrate popular words mentioned in MTurkers' answers in Figure 3. The highlighted words, such as "relevant," "spelling," "logic" inspired us to consider the corresponding factors in our explanation metric.

3 EXPSCORE METRIC

Inspired by factors we collected in the *RecoExp* dataset, we first explain the factor-based framework of *ExpScore* and further present the explainability factors that serve as the basic modules of the framework. To the best of our knowledge, *ExpScore* is the first offline metric designed for evaluating recommendation explanations.

3.1 Factor-based Framework

The key idea of our evaluation framework is to learn a unified evaluation model that aggregates the scores of an explanation on various factors such as relevance, length, subjectivity, popularity, and grammar correctness.

Implementation details Our proposed framework shown in Figure 4 first extracts a set of numeric factors using machine-generated explanations and human reviews (as reference). The extracted factors are then fed as an input to our model. We adopt several simple models such as linear regression, logistic regression, and neural networks to examine the effectiveness of our framework. Figure 4 only shows the neural network as an illustration. For the neural network model configuration, we adopt two hidden layers with 6 and 3 hidden neurons. The learning rate is set to 0.01 and the L_2 regularization parameter is fixed to 0.01. As for the configuration of linear regression and logistic regression, we also set the L_2 regularization parameter to 0.01. We used Adam optimizer in PyTorch and stopped training until the loss does not decrease. Each factor is normalized and concatenated before being fed into the model. *ExpScore* as the output of the framework will be used to measure the

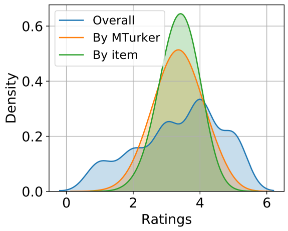

Figure 2: Rating distributions

Figure 3: Wordcloud of factors

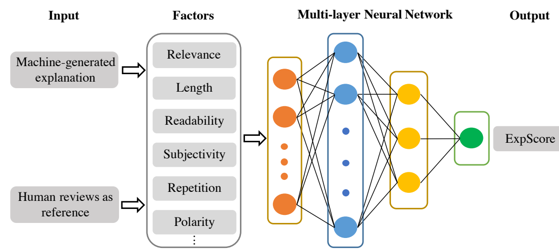

Figure 4: Factor-based framework for *ExpScore*

quality of explanations. The design of our evaluation framework enables the following merits for explanation evaluation.

Extendability We think any single factor is insufficient to measure the explanation quality comprehensively because each such factor can only evaluate explanation from one particular perspective. Therefore, our factor-based framework aims to gather the strengths of multiple text quality factors and generate a high-quality aggregated explainability score for the explanation. However, our framework leaves spaces for additional factor discovery and improvement in the future.

Interpretability In this study, we choose a factor-based framework for better interpretability compared to some BERT-like frameworks such as BertScore [31] and BLUERT [25]. We further examine the effectiveness of our framework by adopting several simple models. Inspired by the two typical paradigms (linear and non-linear) for model design, we adopt three machine learning methods, including linear and non-linear models, to explore the relationship between standalone text quality factors and *ExpScore*. Specifically, we trained linear regression, multinomial logistic regression, and multi-layer neural network models to fit the human ratings on the training dataset. We use cross-entropy as the loss function when training multinomial logistic regression and use mean squared error (MSE) for training linear regression and multi-layer neural networks.

Adaptability In general, our framework adaptability lies in two main aspects. First, it does not need any ground-truth explanations. It can be adopted in many real-world recommendation systems as long as it has machine-generated explanations and human reviews. Second, it is domain-independent. Although the *RecoExp* Dataset is based on a movie recommendation scenario, the factors we use in the framework are not specific to movies, and can be easily transferred to other recommendation systems.

3.2 Explainability Factors

As we know, factors are important modules of our *ExpScore* framework. However, there is no widely-acknowledged definition of explainability of explanation. We decided to let users define what makes a good explanation. From the survey conducted, we learned about many factors, such as good grammar, length, no repetitions, making sense, expressing opinions, detail, relevance, and emotion. We extracted frequently-mentioned and domain-indepedent factors, as we will explain in the following.

To detail the implementation of our method, let us see the notations used by *RecoExp*. For each explanation exp_i, we have its corresponding item as $item_i$, one human review of $item_i$ as rev_i, the feature of $item_i$ as fea_i which is introduced in [16] (such as the color of a phone), and we take the average of five rating scores of the explanation as y_{label}. The following evaluation factors are considered in this work.

• **Relevance** Relevance score (REL) indicates if the explanation is relevant to the corresponding item. Since the item's reviews are informative and reflect users' opinions on the item, we use the item reviews as a reference. Specifically, we compute the semantic similarity between the explanation and the item review as a relevance score. For the implementation method, we use the sentence-BERT model [31] to get the embedding vectors of the explanation and the item review, and then compute the cosine similarity of the two embedding vectors Emb_{exp_i} and Emb_{rev_i}.

• **Length** Length (LEN) of the explanation in this work is defined as the number of words after removing stop words since the length of explanations may influence how users perceive the explanations.

• **Readability** The readability (REA) score of the explanation can be calculated based on the Flesch-Kincaid readability test. Higher scores indicate that the material is easier to read in the Flesch reading ease test [12].

• **Word importance** Word importance (WI) allows us to generate the explanation importance score by adding up the individual WI scores. We simplify the implementation of word importance with inverted term frequency.

• **Repetition** Repetition (REP) refers to how many duplicate segments one explanation has. Significant Repetition in the sentence may lead to low-quality explanations.

• **Subjectivity** Subjectivity (SUB) [20] is one sentiment analysis attribute reflecting whether explanation contains personal opinion, emotion, or judgment. We use Textblob[2]to compute Subjectivity.

• **Polarity** Polarity (POL) [20] indicates the confidence level that explanations are positive or negative. Good explanations may persuade users not to buy an item rather than always giving positive opinions to "cheat" users. Similar to Subjectivity, we use Textblob[2] to compute Polarity.

• **Grammar Correctness** Grammar correctness (GC) reflects the grammar quality of the generated explanations. Too many typos or grammar errors may confuse and frustrate readers. Also, grammar errors make the generated explanations less reliable. We use the Python Language Tool [3] to compute Grammar Correctness.

• **Feature appearance** Feature appearance (FA) measures if an explanation sentence captures item features. It checks whether the explanation contains feature words of the item.

4 RESULTS AND INSIGHTS

In this section, we present the results from our crowdsourced experiments. We start by assessing the proposed evaluation metric. Then we provide a comprehensive analysis of domain-independent explainability factors. Last we compare the performance of three implementations of *ExpScore*.

[2]https://textblob.readthedocs.io/en/dev/
[3]https://pypi.org/project/language-tool-python/

Table 1: Pearson's and Kendall's correlation co-efficient of baseline metrics and *ExpScore* against human quality judgements, divided by 100. All correlations are statistically significant at p<0.001.

Metrics	Pearson r	Kendall τ
BLEU	16.3	10.9
ROUGE	11.2	7.1
METEOR	10.3	6.0
***ExpScore* (linear)**	**44.1**	**30.2**
***ExpScore* (logistic)**	**34.1**	**25.2**
***ExpScore* (NN)**	**44.4**	**30.0**

Table 2: Pearson's and Kendall's correlation coefficients of factors against human quality judgements, divided by 100. All correlations are statistically significant at p<0.001.

Factors	Pearson r	Kendall τ
Relevance	**19.8**	**12.8**
Length	**-16.9**	**-12.1**
Readability	12.0	12.1
Word importance	13.0	16.2
Repetition	-10.0	-6.5
Subjectivity	**30.0**	**18.1**
Polarity	**39.7**	**28.8**
Grammar correctness	8.4	4.6
Feature appearance	14.5	11.4

Table 3: Factor weights generated by *ExpScore* (linear)

REL	**LEN**	REA	WI	REP	**SUB**	**POL**	**GC**	FA
0.081	**-0.063**	0.023	0.052	-0.062	**0.078**	**0.222**	**0.130**	0.022

Table 4: Loss and accuracy on *RecoExp* test dataset reported by various models.

Model	Accuracy	Loss Type	Loss
Linear Regression	0.550	MSE	0.4295
Logistic Regression	0.521	CE	2.7036
Neural Network	0.551	MSE	0.4296

4.1 Metrics Correlation

The most desirable characteristic of an evaluation metric is its strong correlation with human scores. A stronger correlation with human judgment indicates that the metric captures the information humans use to assess an explanation. We compare *ExpScore* with the following metrics in assessing explanation quality.

- **BLEU** The BLEU method uses a modified form of precision to compare a candidate against multiple references [23].
- **ROUGE** The Rouge score of the explanation indicates how the explanation summarises the user review [19].
- **METEOR** METEOR [3, 15] is based on the harmonic mean of unigram precision and recall.

We conducted Kendall's and Pearson's correlation analysis on the above evaluation metrics and three basic *ExpScore* approaches (linear, logistic, NN) against human judgments. The experiments demonstrate all three *ExpScore* metrics vastly outperform existing metrics, as Table 1 shows. Even BLEU performs best among existing metrics, the correlation strength of *ExpScore* is about two times larger than that of BLEU.

4.2 Factors Analysis

We decompose the abstract concept of explainability of recommendation explanations into various factors, each of which describes one aspect of the explanation quality. In Table 2, we calculate the Kendall's and Pearson's correlation coefficient of all factors against human quality judgments. Polarity, Subjectivity, Relevance, and Length have stronger correlations with human assessments, indicating a better explainability when explanations have high Relevance and high emotional preference. However, Length is negatively correlated with the human judgment of explanation quality. One possible reason is that longer explanations are more likely to suffer from repetitions, low readability, and even grammatical errors. On the other hand, Table 3 shows the importance of all factors generated by *ExpScore* (linear). Polarity, Subjectivity, Relevance, Length, and Grammar correctness are the top essential factors among linear weights. We find that it is consistent with correlation strength in terms of positive and negative relationships. However, the weight importance ranking of factors is slightly different from correlation rankings might because *ExpScore* (linear) factors are entangled.

4.3 Model Accuracy

We compare the average test accuracy of the three implementations of *ExpScore* in Table 4. Since we both have regression and multiclass tasks, we decide to adopt a custom accuracy to measure the learning performance for a fair comparison. Accuracy is calculated by considering that the model's *ExpScore* is correct if it falls into the range of $y_{label} \pm 0.5$, where y_{label} is the average human's evaluation score of the corresponding explanation (range is 1 to 5). We could see that these three approaches achieve comparable performance, and the accuracy is not very high. However, high accuracy is not the ultimate goal of this paper. We could add more factors and utilize more complex models like BERT [10] in future work. As the first paper addresses learning metrics for recommendation explanations, we focus more on the interpretability of the framework.

5 CONCLUSION AND FUTURE WORK

Web systems, including search and recommender systems, have been prone to various forms of biases [2]. Transparency is one of the ways we can promote these complex, black-box systems for fairness and social good. Creating and providing meaningful explanations is a primary user-centric way that can be accomplished [32]. In this paper, we introduced a new *RecoExp* dataset to facilitate research concerning the evaluation of recommendation explanations. We presented a novel machine learning-based metric *ExpScore* for evaluating recommendation explanations. For *ExpScore*, we proposed an interpretable and extendable factor-based framework that explores the definition of explainability from various perspectives. We showed that *ExpScore* vastly outperforms existing metrics and correlates better with human evaluation.

In the future, we plan to further explore this research direction in several different dimensions. For instance, here we only compared the explanations generated by NETE, while in the future, we will extend *ExpScore* to support additional explanation generation models and explanation datasets. Besides, since *ExpScore* is model-independent, it can provide a better reference for comparing explanation models than BLEU and ROUGE. We will also consider more evaluation factors to improve the accuracy of the evaluation model further. For example, informativeness and concreteness are highly preferred for a good explanation. Finally, we only considered text explanations in this work, while we will further consider multiple modalities such as images and knowledge graphs for evaluation in the future. Since textual reviews are often aligned with pictures in many scenarios, such as online shopping or hotel reviewing, we can adopt various modalities for joint learning.

6 ACKNOWLEDGEMENTS

This work is supported by the US National Science Foundation (NSF) award number IIS-1910154.

REFERENCES

[1] Enrique Amigo, Julio Gonzalo, Jesus Giménez, and Felisa Verdejo. 2011. Corroborating text evaluation results with heterogeneous measures. *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (2011), 455–466.

[2] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.

[3] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[4] Oren Barkan, Yonatan Fuchs, Noam Koenigstein, and Avi Caciularu. 2020. Explainable Recommendations via Attentive Multi-Persona Collaborative Filtering. (2020). https://doi.org/10.1145/3383313.3412226 arXiv:2010.07042v1

[5] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 World Wide Web Conference*. 1583–1592.

[6] Hanxiong Chen, Xu Chen, Shaoyun Shi, and Yongfeng Zhang. 2019. Generate natural language explanations for recommendation. *SIGIR 2019 Workshop on ExplainAble Recommendation and Search* (2019).

[7] Xu Chen, Zheng Qin, Yongfeng Zhang, and Tao Xu. 2016. Learning to rank features for recommendation over multiple categories. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 305–314.

[8] Xu Chen, Yongfeng Zhang, and Zheng Qin. 2019. Dynamic explainable recommendation based on neural attentive models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 53–60.

[9] Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. 2018. Automatic Generation of Natural Language Explanations. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*. ACM, 57.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv* abs/1810.04805 (2019).

[11] Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 623–632.

[12] James N Farr, James J Jenkins, and Donald G Paterson. 1951. Simplification of Flesch Reading Ease Formula. *Journal of Applied Psychology* 35, 5 (1951), 333–337. https://doi.org/10.1037/h0062427

[13] Krisztian Balog Google and Filip Radlinski. [n.d.]. Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations. ([n. d.]), 10. https://doi.org/10.1145/3397271.3401032

[14] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.

[15] Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*. 228–231.

[16] Lei Li, Yongfeng Zhang, and Li Chen. 2020. Generate Neural Template Explanations for Recommendation. (2020). https://doi.org/10.1145/3340531.3411992

[17] 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*. 755–764.

[18] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 345–354.

[19] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://www.aclweb.org/anthology/W04-1013

[20] Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing, Second Edition* January 2010 (2010), 627–666.

[21] Jianmo Ni, Jiacheng Li, and Julian Mcauley. [n.d.]. *Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects*. Technical Report.

[22] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. 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)*. 188–197.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[24] Zhaochun Ren, Shangsong Liang, Piji Li, Shuaiqiang Wang, and Maarten de Rijke. 2017. Social collaborative viewpoint regression with explainable recommendations. In *Proceedings of the tenth ACM international conference on web search and data mining*. ACM, 485–494.

[25] Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. *arXiv* (2020), 7881–7892. https://doi.org/10.18653/v1/2020.acl-main.704 arXiv:2004.04696

[26] Quoc-Tuan Truong and Hady Lauw. 2019. Multimodal review generation for recommender systems. In *The World Wide Web Conference*. 1864–1874.

[27] Kosetsu Tsukuda and Masataka Goto. 2020. Explainable Recommendation for Repeat Consumption. In *RecSys 2020 - 14th ACM Conference on Recommender Systems*. 462–467. https://doi.org/10.1145/3383313.3412230

[28] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable Recommendation via Multi-Task Learning in Opinionated Text Data. *arXiv preprint arXiv:1806.03568* (2018).

[29] Xiting Wang and Yiru Chen. 2018. 2018 IEEE International Conference on Data Mining A Reinforcement Learning Framework for Explainable Recommendation. *IEEE International Conference on Data Mining* (2018). https://www.microsoft.com/en-us/research/publication/a-reinforcement-learning-framework-for-explainable-recommendation/

[30] Zhongqing Wang and Yue Zhang. 2017. Opinion recommendation using neural memory model. *arXiv preprint arXiv:1702.01517* (2017).

[31] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv* (2019), 1–43. arXiv:1904.09675

[32] Yongfeng Zhang and Xu Chen. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval* (2020). https://doi.org/10.1561/9781680836592

[33] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 83–92.