香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

Generate Neural Template Explanations for Recommendation

Lei Li[1], Yongfeng Zhang[2], Li Chen[1]

[1] Hong Kong Baptist University, [2] Rutgers University

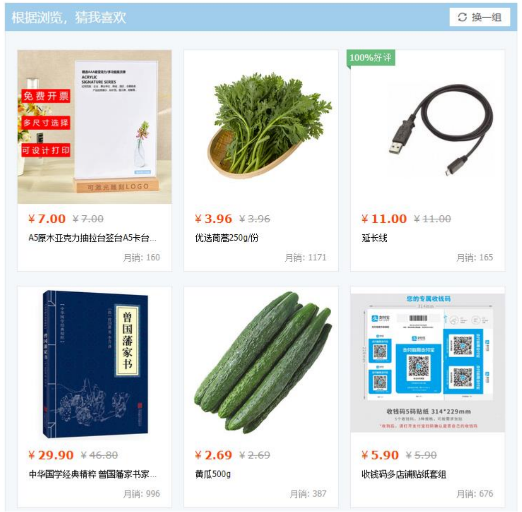
csleili@comp.hkbu.edu.hk

October 21, 2020

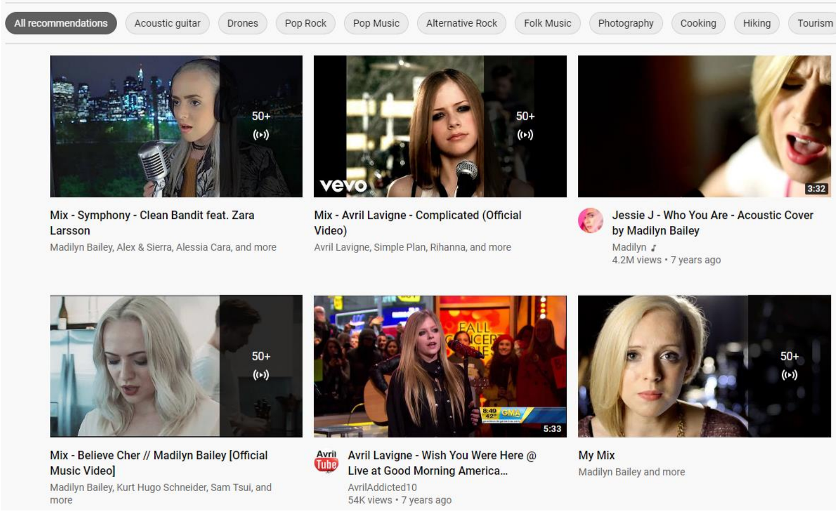
The 29th ACM International Conference on Information and Knowledge Management (CIKM'20)

Recommendation Everywhere

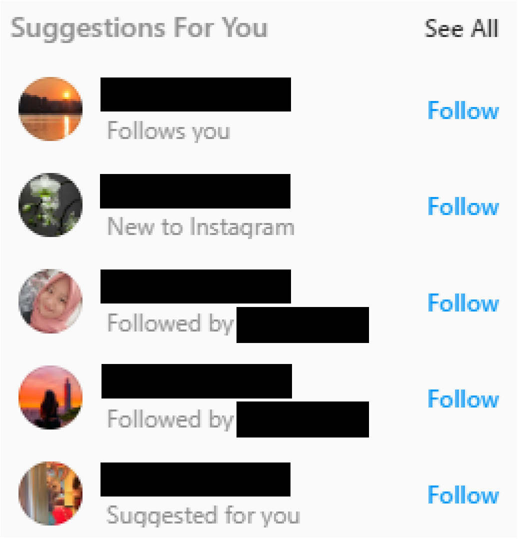
E-commerce
(taobao.com)



Video
(youtube.com)



Social network
(instagram.com)



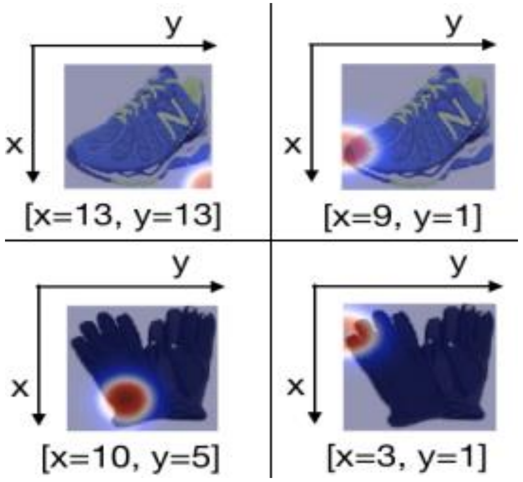
Movie
(movie.douban.com)

Explanation for Recommendation

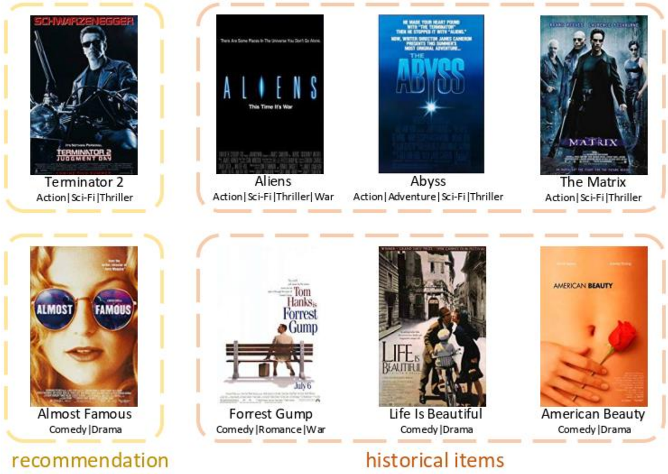
- Help users understand recommendations
- Benefits of Explanation (Tintarev and Mashoff. Handbook’15)
  - **Trust**: increase users’ confidence in the system
  - **Effectiveness**: help users make good decisions
  - **Persuasiveness**: convince users to try or buy
  - **Efficiency**: help users make decisions faster
  - **Satisfaction**: increase the ease of use or enjoyment

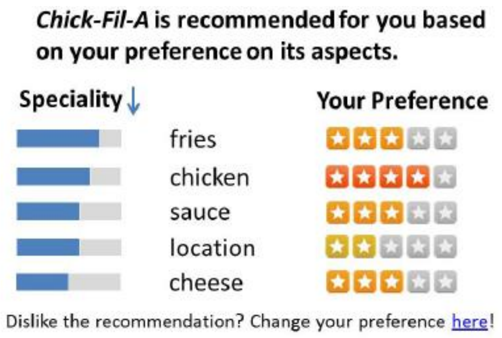
Typical Explanation Styles

Images (Chen et al. SIGIR’19)



Neighbors (Li et al. WWW’20)





Features (He et al. CIKM’15)

```
You might be interested in [feature],
on which this product performs well.
```

```
You might be interested in [feature],
on which this product performs poorly.
```

Templates (Zhang et al. SIGIR’14)

- **Textual Sentences**
  - Able to communicate rich information to users
  - Massive textual data available (e.g., user reviews)

Limitations of Existing Textual Explanations

- Pre-defined templates
  - Require human effort to create
  - Restrict the sentence expressiveness
- Free-style sentences
  - Topics of sentences sometimes irrelevant to the recommendation
  - Sentences similar or even identical

CF (Sarwar et al. WWW'01)	Customers who bought this item also bought.
EFM (Zhang et al. SIGIR'14)	You might be interested in [***feature***], on which this product performs well.

Reference	They have a huge ***variety*** of things.
NRT (Li et al. SIGIR'17)	The food is good.
Att2Seq (Dong et al. EACL'17)	I'm not sure if I need to go back.
Reference	The black garlic ***ramen*** was good as well.
NRT	The food is good.
Att2Seq	The food was great.

Motivation

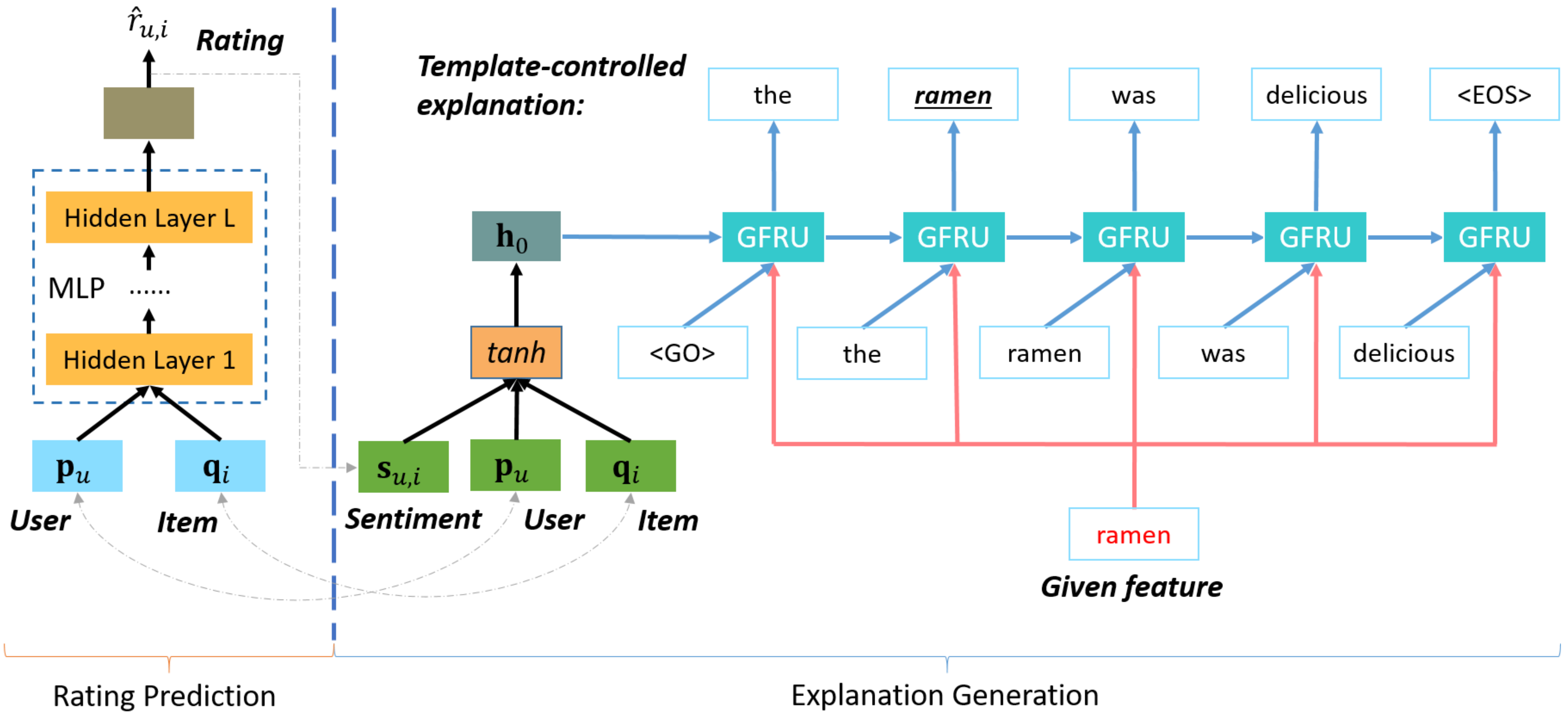
- Improve overall user experience and the recommendation acceptance
- Propose a **Neural Template** (NETE) approach that can produce expressive and high-quality explanations
- Bridge the benefits of template and generation approaches
  - Learn templates from data
  - Generate template-shaped explanations about specific features

Reference	They have a huge ***variety*** of things.
NETE	They have a **<u>variety</u>** of things to choose from.
Reference	The black garlic ***ramen*** was good as well.
NETE	The **<u>ramen</u>** was delicious.

Problem Formulation

- Recommendation
  - Predict a rating $\hat{r}_{u,i}$, given a user u and an item i
- Explanation
  - Generate an explanation sentence $\hat{S}_{u,i}$, given a feature $f_{u,i}$
- Feature Prediction
  - The feature $f_{u,i}$ can be either manually set by the user u
  - Or predicted by a prediction method based on the user's interests

Overview of Our Neural Template Model



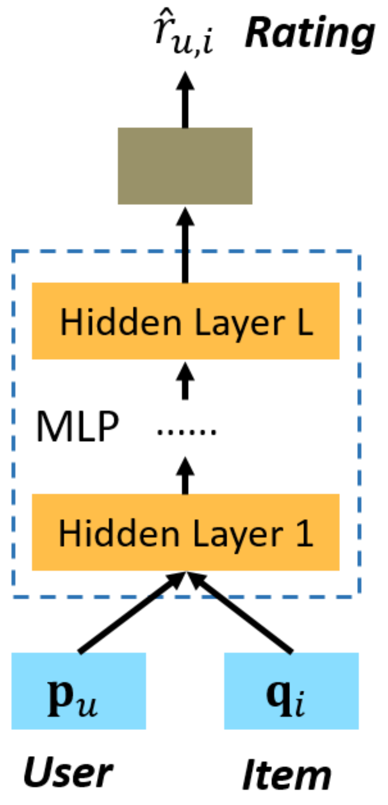
Personalized Recommendation

- Capture the interactions between users and items via MLP
- The non-linear transformations of MLP have better representation ability than linear models, e.g., MF (Mnih and Ruslan. NIPS'08)

$$\begin{cases} \mathbf{z}_1 = \sigma(\mathbf{W}_1[\mathbf{p}_u, \mathbf{q}_i] + \mathbf{b}_1) \\ \mathbf{z}_2 = \sigma(\mathbf{W}_2\mathbf{z}_1 + \mathbf{b}_2) \\ \dots \\ \mathbf{z}_L = \sigma(\mathbf{W}_L\mathbf{z}_{L-1} + \mathbf{b}_L) \end{cases} \quad \text{and } \hat{r}_{u,i} = \mathbf{w}_r\mathbf{z}_L + b_r$$

- Mean squared error loss function

$$\mathcal{L}_r = \frac{1}{|\mathcal{T}|}\sum_{u,i\in\mathcal{T}}(r_{u,i} - \hat{r}_{u,i})^2$$



Explanation Generation (1)

- Essentially a table-to-text generation task (Wiseman et al. EMNLP'18)
- Encoder
  - MLP encodes user u and item i for personalization, and the sentiment (derived from the predicted rating $\hat{r}_{u,i}$) for sentiment control

$$\mathbf{h}_0 = \tanh(\mathbf{W}_e[\mathbf{p}_u, \mathbf{q}_i, \mathbf{s}_{u,i}] + \mathbf{b}_e)$$

- Decoder
  - Sentences from vanilla decoder could be irrelevant to the recommendation
  - Propose a **Gated Fusion Recurrent Unit (GFRU)** that could include a given feature in the generated sentence

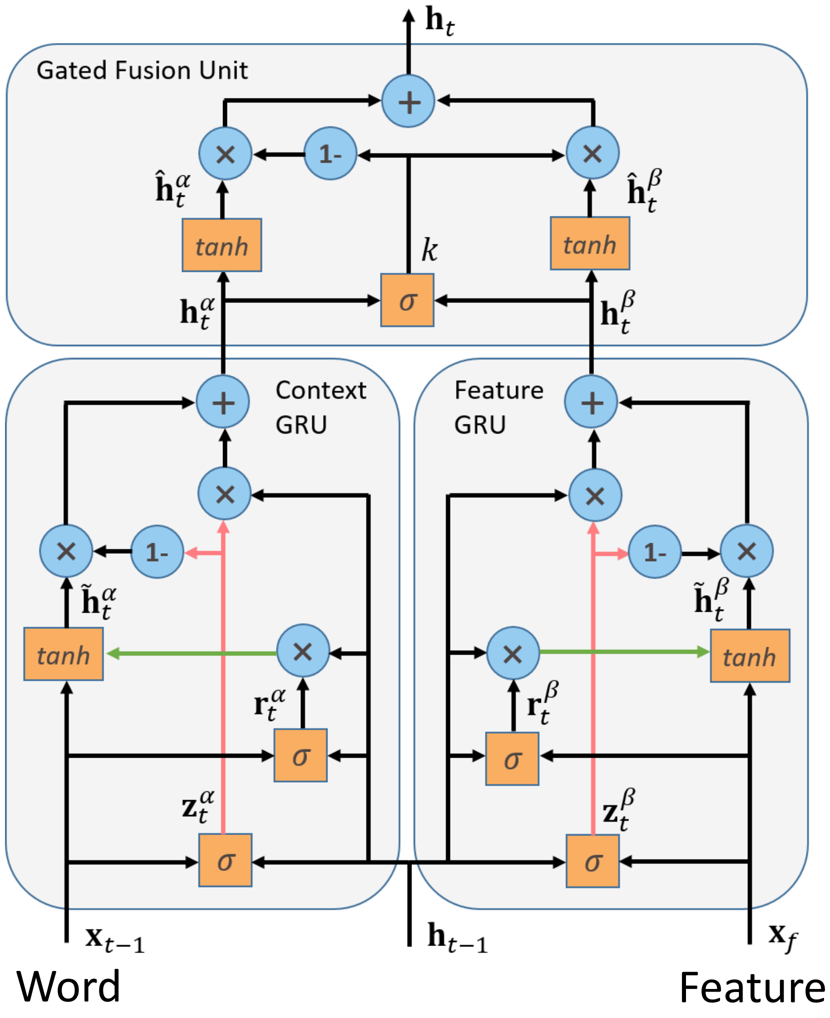
Gated Fusion Recurrent Unit (GFRU)

- Two GRUs (Cho et al. EMNLP’14) process two types of information
  - The context GRU takes the previously generated word as input
  - The feature GRU takes the given feature
- One Gated Fusion Unit (GFU) (Arevalo. ICLR’17 Workshop) merges them

$$\begin{cases} \hat{\mathbf{h}}_t^{\alpha} = \tanh(\mathbf{W}_{\alpha}\mathbf{h}_t^{\alpha}) \\ \hat{\mathbf{h}}_t^{\beta} = \tanh(\mathbf{W}_{\beta}\mathbf{h}_t^{\beta}) \\ k = \sigma(\mathbf{w}_k[\hat{\mathbf{h}}_t^{\alpha}, \hat{\mathbf{h}}_t^{\beta}]) \\ \mathbf{h}_t = (1-k) \odot \mathbf{h}_t^{\alpha} + k \odot \mathbf{h}_t^{\beta} \end{cases}$$

Large -> Template

Small -> Feature



Explanation Generation (2)

- Hidden states of each time step can be computed by GFRU

$$\mathbf{h}_t = g(\mathbf{x}_{t-1}, \mathbf{h}_{t-1}, \mathbf{x}_f)$$

- During decoding, a word with the largest probability over the vocabulary is sampled

$$p(y_t | y_{<t}, \mathbf{h}_0) = \mathrm{softmax}_{y_t}(\mathbf{W}_v \mathbf{h}_t + \mathbf{b}_v)$$

- Cross-entropy loss function

$$\mathcal{L}_e = \frac{1}{|\mathcal{T}|} \sum_{u,i \in \mathcal{T}} \frac{1}{|S_{u,i}|} \sum_{t=1}^{|S_{u,i}|} -\log p(y_t)$$

Model Training

- Two tasks
  - Recommendation
  - Explanation
- Little research studies if and how the two tasks are compatible in a joint learning framework
- Investigate the influence of different learning frameworks
  - Single-task learning
  - Multi-task learning

$$\mathcal{L}_r \rightarrow \mathcal{L}_e$$

$$\mathcal{J} = \min_{\Theta}(\lambda_r \mathcal{L}_r + \lambda_e \mathcal{L}_e + \lambda_n ||\Theta||^2)$$

Feature Prediction

- Extract features from user reviews via a toolkit (Zhang et al. SIGIR’14)
- Utilize point-wise mutual information (PMI) to predict a user’s interest to each feature
  - Measure its relationship with the user’s preferred features

$$\hat{f}_i = \mathrm{argmax}_{f \in \mathcal{F}_i} \mathrm{PMI}(\mathcal{F}_u, f)$$

$$\mathrm{PMI}(\mathcal{F}_u, f) = \log \frac{p(\mathcal{F}_u | f)}{p(\mathcal{F}_u)} \approx \log \frac{\prod_{f' \in \mathcal{F}_u} p(f' | f)}{\prod_{f' \in \mathcal{F}_u} p(f')} = \sum_{f' \in \mathcal{F}_u} \log \frac{p(f' | f)}{p(f')} = \sum_{f' \in \mathcal{F}_u} \mathrm{PMI}(f', f)$$

$$\mathrm{PMI}(f_u, f_i) = \log \frac{p(f_u, f_i)}{p(f_u) p(f_i)} = \log \frac{p(f_u | f_i)}{p(f_u)}$$

- Two times better than randomly selecting the target item’s features

Datasets

- TripAdvisor
  - Hotel
- Yelp
  - Restaurant
- Amazon
  - Movie & TV
- The explanation is a review sentence containing features

	TA-HK	YELP19	AZ-MT
# of users	9,765	27,147	7,506
# of items	6,280	20,266	7,360
# of reviews	320,023	1,293,247	441,783
# of features	5,069	7,340	5,399
Avg. # of reviews / user	32.77	47.64	58.86
Avg. # of reviews / item	50.96	63.81	60.02
Avg. # of words / explanation	13.01	12.32	14.14

* **TA** and **AZ** denote **TripAdvisor** and **Amazon**, respectively.

Evaluation Metrics

- Recommendation
  - Rating prediction: RMSE and MAE
  - Personalized ranking: NDCG and HR
- Explanation
  - Text quality: BLEU (Papineni et al. ACL'02) and ROUGE (Lin. ACL'04 Workshop)
  - **Explainability**: previous work mostly ignored
  - Design 4 metrics
    - Unique Sentence Ratio (USR)
    - Feature Matching Ratio (FMR)
    - Feature Coverage Ratio (FCR)
    - Feature Diversity (DIV)

$$USR = |\mathcal{S}| / N \qquad FMR = \frac{1}{N} \sum_{u,i} \delta(f_{u,i} \in \hat{S}_{u,i})$$

$$FCR = N_g / |\mathcal{F}|$$

$$DIV = \frac{2}{N \times (N-1)} \sum_{u,u',i,i'} \left| \hat{\mathcal{F}}_{u,i} \cap \hat{\mathcal{F}}_{u',i'} \right|$$

Ablation Study

- Investigate the impacts of different settings

	Learning Framework		Decoder		Given Features	
	Single-task	Multi-task	GFRU	GRU	In ground-truth	By PMI
NETE	√		√		√	
NETE-GRU	√			√	√	
NETE-MUL		√	√		√	
NETE-GM		√		√	√	
NETE-PMI	√		√			√

Quantitative Analysis on Explanations (1)

	Personalization				BLEU (%)		ROUGE-1 (%)			ROUGE-2 (%)		
	USR	FMR	FCR	DIV	BLEU-1	BLEU-4	Precision	Recall	F1	Precision	Recall	F1
	TA-HK dataset											
	YELP19 dataset											
	AZ-MT dataset											
NRT	0.00	-	0.01	5.46	14.02	0.57	23.57	14.24	16.87	2.53	1.70	1.92
Att2Seq	0.34	-	0.18	2.81	12.78	1.01	20.53	13.49	15.42	2.77	1.87	2.09
NETE-GM	0.00	-	0.01	4.12	12.31	0.50	22.77	13.43	16.18	2.40	1.51	1.76
NETE-GRU	0.38	-	0.11	2.34	12.10	0.95	20.16	12.93	14.93	2.63	1.75	1.97
NETE-MUL	0.05	0.61	0.03	2.63	17.20	1.94	33.79	20.01	24.17	7.50	4.32	5.16
NETE-PMI	**0.72**	0.50	**0.19**	3.06	13.02	0.82	20.93	12.76	14.99	2.36	1.63	1.81
NETE	0.57**	**0.71**	**0.19***	**1.93****	**18.76****	**2.46****	**33.87****	**21.43****	**24.81****	**7.58****	**4.77****	**5.46****
Improvement (%)	+69.1	-	+5.6	+45.2	+33.8	+143.6	+43.7	+50.5	+47.1	+174.3	+154.9	+161.2

Our methods consistently achieve the best performance on three datasets

Quantitative Analysis on Explanations (2)

Multi-task

	Personalization				BLEU (%)		ROUGE-1 (%)			ROUGE-2 (%)		
	USR	FMR	FCR	DIV	BLEU-1	BLEU-4	Precision	Recall	F1	Precision	Recall	F1
TA-HK dataset												
NRT	0.00	-	0.00	13.61	14.26	0.80	17.57	16.52	16.56	2.45	2.64	2.48
Att2Seq	0.18	-	0.17	3.93	14.76	1.01	19.26	14.45	15.83	2.43	1.96	2.06
NETE-GM	0.00	-	0.00	14.40	14.01	0.83	17.55	16.19	16.42	2.50	2.60	2.50
NETE-GRU	0.27	-	0.15	3.00	13.84	0.92	18.55	13.64	15.02	2.23	1.76	1.86
NETE-MUL	0.02	0.66	0.07	3.92	22.09	3.33	32.59	23.96	26.30	8.87	6.51	7.00
NETE-PMI	**0.79**	0.38	**0.30**	2.92	14.55	0.82	17.84	13.96	14.90	2.01	1.70	1.74
NETE	0.57**	**0.78**	0.27**	**2.22****	**22.39****	**3.66****	**35.68****	**24.86****	**27.71****	**10.20****	**6.98****	**7.66****
Improvement (%)	+210.7	-	+57.1	+77.1	+51.7	+261.3	+85.2	+50.5	+67.3	+317.0	+164.0	+209.1

- Less than 3% unique sentences across the whole dataset
  - Multi-task learning is harmful to sentence diversity

- USR different but BLEU and ROUGE close
  - BLEU and ROUGE cannot properly evaluate sentence diversity
  - It motivates us to design new metrics

Quantitative Analysis on Explanations (3)

	Personalization				BLEU (%)		ROUGE-1 (%)			ROUGE-2 (%)		
	USR	FMR	FCR	DIV	BLEU-1	BLEU-4	Precision	Recall	F1	Precision	Recall	F1
	TA-HK dataset											
NRT	0.00	-	0.00	13.61	14.26	0.80	17.57	16.52	16.56	2.45	2.64	2.48
Att2Seq	0.18	-	0.17	3.93	14.76	1.01	19.26	14.45	15.83	2.43	1.96	2.06
NETE-GM	0.00	-	0.00	14.40	14.01	0.83	17.55	16.19	16.42	2.50	2.60	2.50
NETE-GRU	0.27	-	0.15	3.00	13.84	0.92	18.55	13.64	15.02	2.23	1.76	1.86
NETE-MUL	0.02	0.66	0.07	3.92	22.09	3.33	32.59	23.96	26.30	8.87	6.51	7.00
NETE-PMI	**0.79**	0.38	**0.30**	2.92	14.55	0.82	17.84	13.96	14.90	2.01	1.70	1.74
NETE	0.57**	**0.78**	0.27**	**2.22****	**22.39****	**3.66****	**35.68****	**24.86****	**27.71****	**10.20****	**6.98****	**7.66****
Improvement (%)	+210.7	-	+57.1	+77.1	+51.7	+261.3	+85.2	+50.5	+67.3	+317.0	+164.0	+209.1

GRU

- Diverse sentences
- Given features mostly included
- Improved feature coverage ratio & diversity
  - Single-task learning
  - GFRU

- Most similar to ground-truth
  - Informativeness of given features
  - Effectiveness of GFRU

Quantitative Analysis on Explanations (4)

	Personalization				BLEU (%)		ROUGE-1 (%)			ROUGE-2 (%)		
	USR	FMR	FCR	DIV	BLEU-1	BLEU-4	Precision	Recall	F1	Precision	Recall	F1
	TA-HK dataset											
NRT	0.00	-	0.00	13.61	14.26	0.80	17.57	16.52	16.56	2.45	2.64	2.48
Att2Seq	0.18	-	0.17	3.93	14.76	1.01	19.26	14.45	15.83	2.43	1.96	2.06
NETE-GM	0.00	-	0.00	14.40	14.01	0.83	17.55	16.19	16.42	2.50	2.60	2.50
NETE-GRU	0.27	-	0.15	3.00	13.84	0.92	18.55	13.64	15.02	2.23	1.76	1.86
NETE-MUL	0.02	0.66	0.07	3.92	22.09	3.33	32.59	23.96	26.30	8.87	6.51	7.00
NETE-PMI	**0.79**	0.38	**0.30**	2.92	14.55	0.82	17.84	13.96	14.90	2.01	1.70	1.74
NETE	0.57**	**0.78**	0.27**	**2.22****	**22.39****	**3.66****	**35.68****	**24.86****	**27.71****	**10.20****	**6.98****	**7.66****
Improvement (%)	+210.7	-	+57.1	+77.1	+51.7	+261.3	+85.2	+50.5	+67.3	+317.0	+164.0	+209.1

Predicted features may not match those in the ground-truth explanations

Qualitative Case Study on Explanations

- Good linguistic quality
  - Learn templates from data, e.g., "__ *was large/comfortable*"
- Good controllability
  - Generate more targeted explanations for specific features
  - Produce personalized explanations for different user-item pairs
  - Take into account the sentiment of the predicted ratings

Rating	Feature	Explanation
4		***The rooms are spacious and the bathroom has a large tub.***
3.90	bathroom	The **bathroom** was large and had a separate shower.
	tub	The bathroom had a separate shower and **tub**.
	rooms	The **rooms** are large and comfortable.
4		***The rooms are brilliant and ideal for business travellers.***
4.13	rooms	The **rooms** are very spacious and the **rooms** are very comfortable.
2		***The broken furniture and dirty surfaces are a dead giveaway.***
2.96	furniture	The **furniture** is worn.
4		***Ideal for plane spotters and very close to the airport.***
2.76	airport	It is not close to the **airport**.

Recommendation Performance

	Rating Prediction						Personalized Ranking					
	TA-HK		YELP19		AZ-MT		TA-HK		YELP19		AZ-MT	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5	HR@5
	Baselines											
MF	0.798	0.613	1.011	0.782	0.963	0.719	0.361	0.559	0.116	0.140	0.449	0.416
SVD++	0.798	0.610	1.011	0.785	0.965	0.718	0.362	0.553	0.116	0.138	0.443	0.350
DeepCoNN	0.796	0.607	1.011	0.789	0.959	0.721	0.630	0.963	0.225	0.216	1.044	1.096
NRT	0.792	**0.605**	**1.007**	0.783	0.957	0.718	0.687	1.074	0.218	0.218	1.305	1.178
	Ours											
NETE-GM	0.793	0.606	1.008	0.785	0.957	**0.713**	0.719	1.119	0.281	0.288	1.616	1.452
NETE-MUL	**0.790**	0.608	1.008	**0.781**	**0.956**	0.717	0.594	0.915	0.234	0.246	1.587	1.507
NETE	0.792	0.608	1.010	0.789	0.961	0.727	**1.039***	**1.509***	**0.484***	**0.515***	**1.671***	**1.578***
Improvement (%)	-	-	-	-	-	-	+51.2	+40.5	+115.1	+136.2	+28.0	+34.0

- Accuracy is close
- Not all items evaluated
    - Data selection bias (Harald. RecSys'13)

- Performance gap widens
- NETE outperforms the others
    - The advantage of single-task learning

Conclusion and Future Work

- Propose a model NETE
  - Generate neural template sentences
  - Improve the expressiveness and quality of explanations
- Design four novel metrics
  - Specifically care about the explainability of the generated sentences
- Show the controllability of NETE
  - Generate explanations about the given user, item, sentiment, and features

- Will increase the expressiveness of the explanations
  - Consider adjective words
  - Extend the model to multiple features

References (1)

- [1] Tintarev, Nava, and Judith Masthoff. "Explaining recommendations: Design and evaluation." Recommender systems handbook. 2015.
- [2] Sarwar Badrul, et al. "Item-based collaborative filtering recommendation algorithms." WWW'01.
- [3] Zhang, Yongfeng, et al. "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis." SIGIR'14.
- [4] Li, Piji, et al. "Neural rating regression with abstractive tips generation for recommendation." SIGIR'17.
- [5] Dong, Li, et al. "Learning to generate product reviews from attributes." EACL'17.
- [6] Li, Xueqi, et al. "Directional and Explainable Serendipity Recommendation." WWW'20.
- [7] Chen, Xu, et al. "Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network: Towards Visually Explainable Recommendation." SIGIR'19.
- [8] He, Xiangnan, et al. "Trirank: Review-aware explainable recommendation by modeling aspects." CIKM'15.
- [9] Mnih, Andriy, and Ruslan R. Salakhutdinov. "Probabilistic matrix factorization." NIPS'08.

References (2)

- [10] Wiseman, Sam, et al. "Learning neural templates for text generation." EMNLP’18.
- [11] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." EMNLP’14.
- [12] Arevalo, John, et al. "Gated multimodal units for information fusion." ICLR’17 Workshop.
- [13] Zhang, Yongfeng, et al. "Do users rate or review? Boost phrase-level sentiment labeling with review-level sentiment classification." SIGIR’14.
- [14] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." ACL’02.
- [15] Lin, Chin-Yew. "ROUGE: A Package for Automatic Evaluation of Summaries." ACL’04 Workshop.
- [16] Steck, Harald. "Evaluation of recommendations: rating-prediction and ranking." RecSys’13.

Q&A

Thank you!

lileipisces.github.io