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

Natural Language Explanation for Recommendations and Beyond

Lei Li

Supervisor: Dr. Li Chen

Ph.D. Thesis Defense

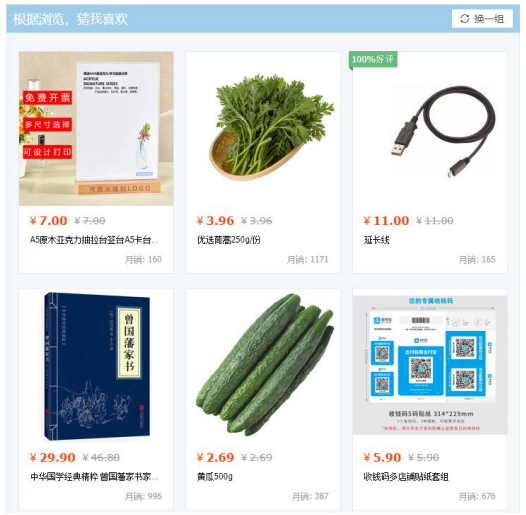
Apr. 27, 2022

Outline

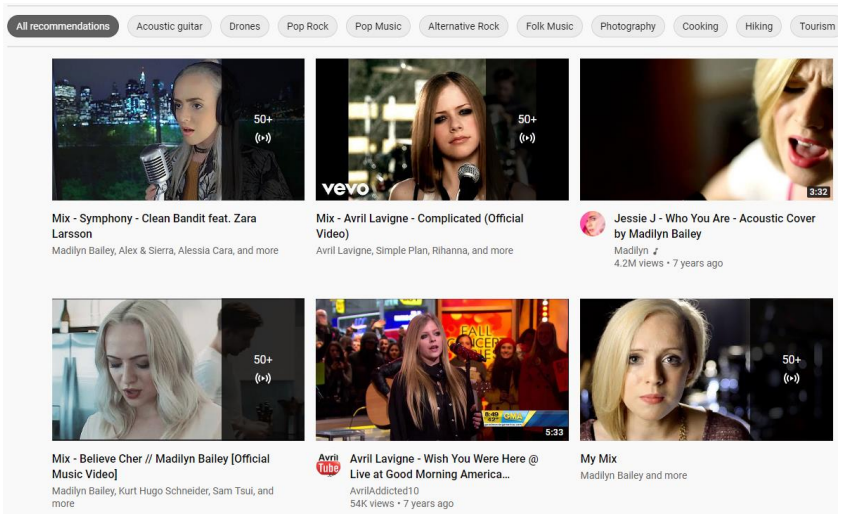
- **Introduction**
  - Explainable Recommendation
- Context-aware Explanation
- Neural Template Explanation Generation
- Natural Language Explanation Generation
- Explanation Ranking
- Future Work

Recommendations Everywhere

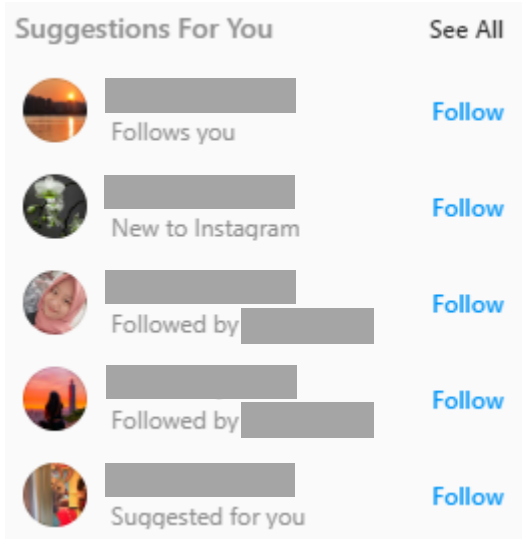
E-commerce
(taobao.com)



Video
(youtube.com)



Social Network
(instagram.com)



Movie
(movie.douban.com)

Explainable Recommendation

- Provide an explanation to justify why an item is recommended to a user (Zhang and Chen, 2020)
  - The style of the jacket is fashionable

Explanatory Goals (Tintarev and Mashoff, 2015)

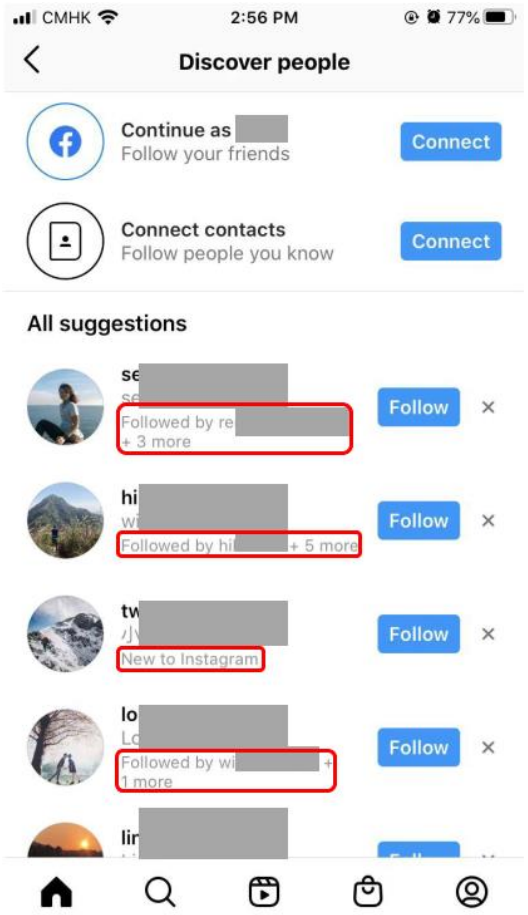
- **Trust**: increase users' confidence in the system
- **Effectiveness**: help users make good decisions
- **Persuasiveness**: convince users to try or buy
- **Efficiency**: help users make decisions faster
- **Satisfaction**: increase the ease of use or enjoyment
- **Transparency**: explain how the system works
- **Scrutability**: allow users to tell the system it is wrong

User-centric

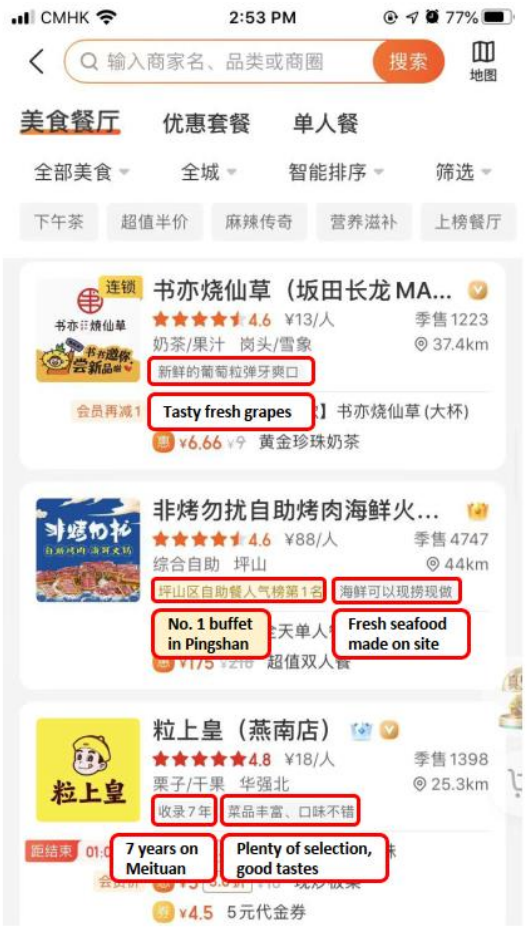
System-centric

Why Natural Language Explanation?

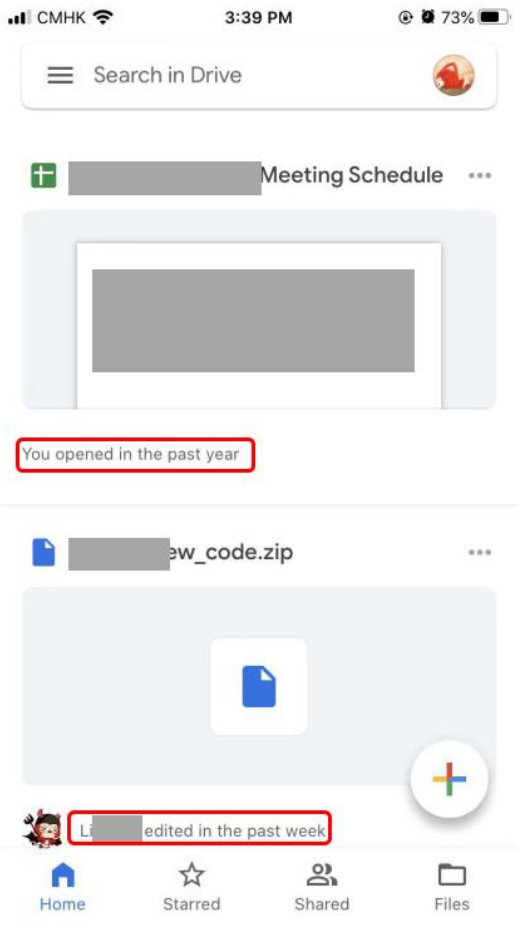
- Able to communicate rich information to users
- Massive textual data available online
  - User reviews



Instagram
(instagram.com)



Meituan
(meituan.com)

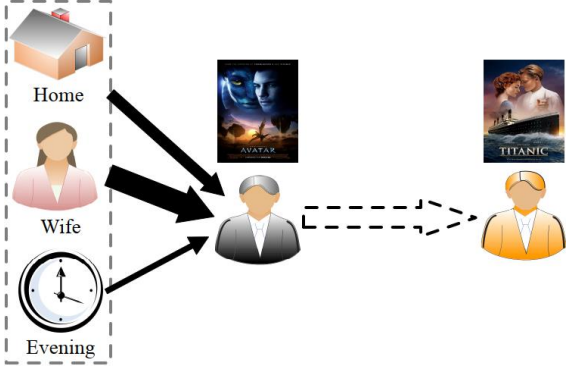


Google Drive
(drive.google.com)

Outline

- Introduction
- **Context-aware Explanation**
  - JIIS 2021
- Neural Template Explanation Generation
- Natural Language Explanation Generation
- Explanation Ranking
- Future Work

Motivation



Courtesy image from (Mei et al., CIKM'18)

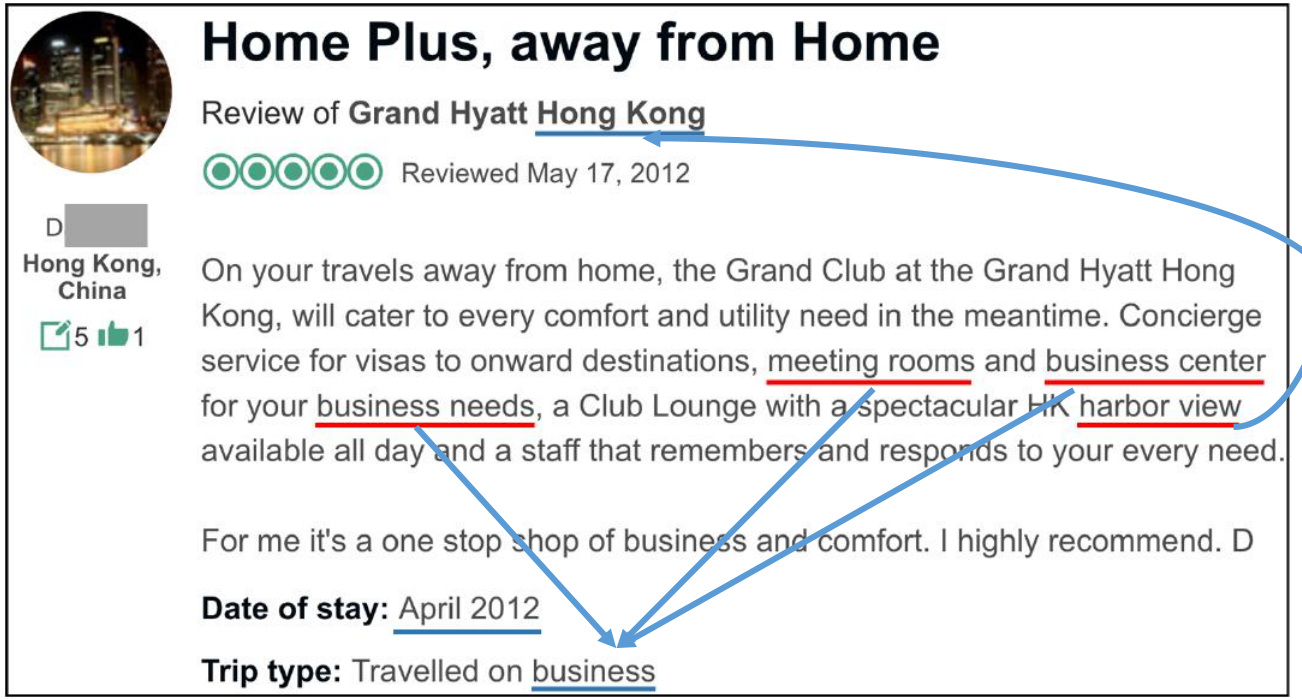
- "Context is any information that can be used to characterize the situation of an entity." (Abowd et al., HUC'99)
  - Location
  - Companion
  - Time
- Context-aware recommendation has been extensively studied.
- Context-aware explanation received relatively less attention.
  - This movie [Titanic] is recommended to you, because its **features** [*plot and music*] are suitable for your current context [wife].

How to obtain such features?

Contextual Features in User Reviews

- User reviews contain rich contextual features.
  - Contexts
  - Contextual features

How to correlate a feature with a context?



A hotel review (tripadvisor.com)

Contextual Feature Mining

- Extract features from user reviews via a toolkit (Zhang et al., SIGIR'14)
- Measure the relevance between a feature f and a context c

$$PMI_f^c = \frac{freq_f^c}{freq_f \cdot freq^c}$$

$$avg_f = \frac{1}{|\mathcal{C}_j|} \sum_{c \in \mathcal{C}_j} PMI_f^c$$

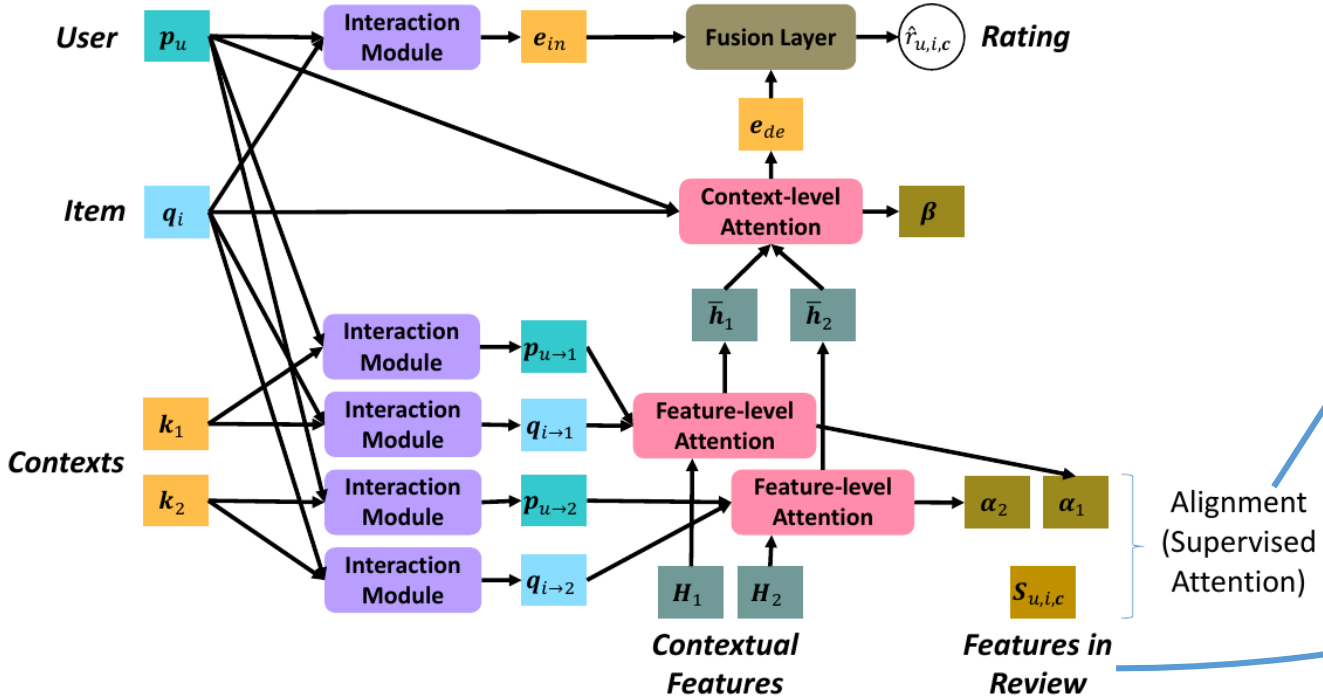
$$err_f^c = PMI_f^c - avg_f$$

$$w_f^c = \left| err_f^c \right|$$

- The larger the weight, the closer the feature to the context
- Select top features for each context

Attention based Explanation

- **Two-level attention** mechanism (Luong et al., EMNLP'15) for selecting important context and its contextual features
- **Supervised attention** mechanism (Liu et al., ACL'17) for matching to user's preference on ground-truth features



$$\mathcal{L}_e = \sum_{(u,i,\mathbf{c})\in\mathcal{T}} \sum_{j=1}^{m} \sum_{k=1}^{n} (s_j^k - \alpha_j^k)^2$$

2. Align attention score with distribution

$$s_j^f = freq_f^{c_j} \Big/ \sum_{f'\in\mathcal{F}_j^{c_j}} freq_{f'}^{c_j}$$

1. Feature distribution in target review

Datasets

- Two typical service domains
  - Hotel
  - Restaurant

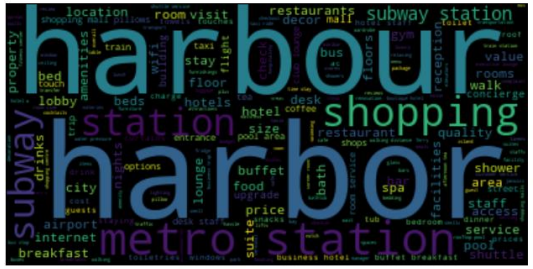
	TripAdvisor	Yelp
# of users	9,765	27,147
# of items	6,280	20,266
# of reviews	320,023	1,293,247
Avg. # of reviews / user	32.77	47.64
Avg. # of reviews / item	50.96	63.81
# of contextual variables in companion	6	-
# of contextual variables in day of a week	-	7
# of contextual variables in month	13	12
# of contextual variables in destination	415	242

Contextual Feature Analysis

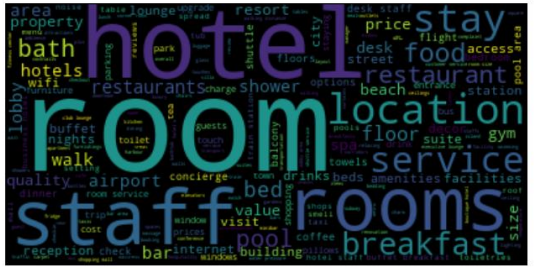
- The contextual feature mining approach is capable of discovering context-aware features.
  - Harbor, shopping, and metro station for Hong Kong
- Those adopted in existing work are context-unaware features.
  - Room, hotel, and staff

(a) Contextual features for business

(b) Contextual features for couples



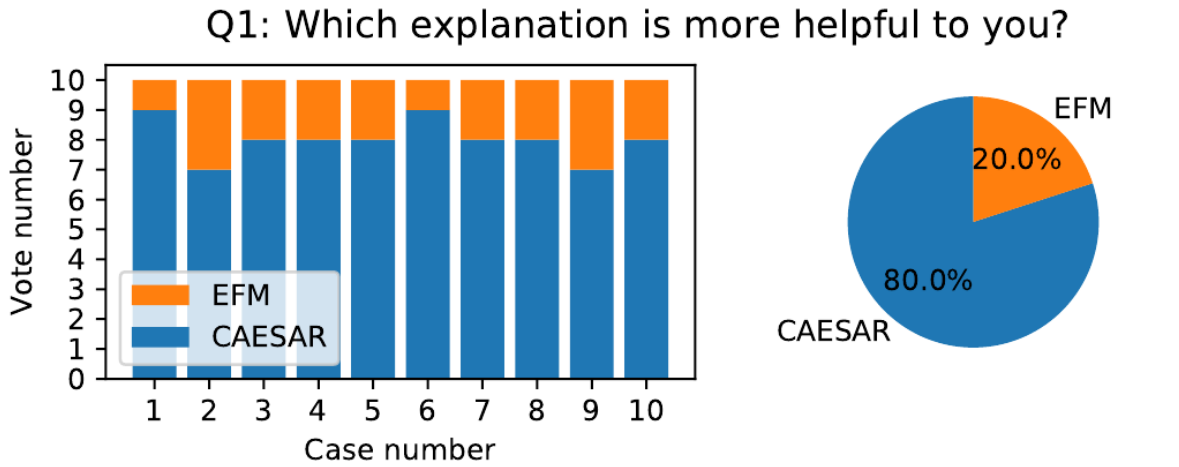
(c) Contextual features for Hong Kong

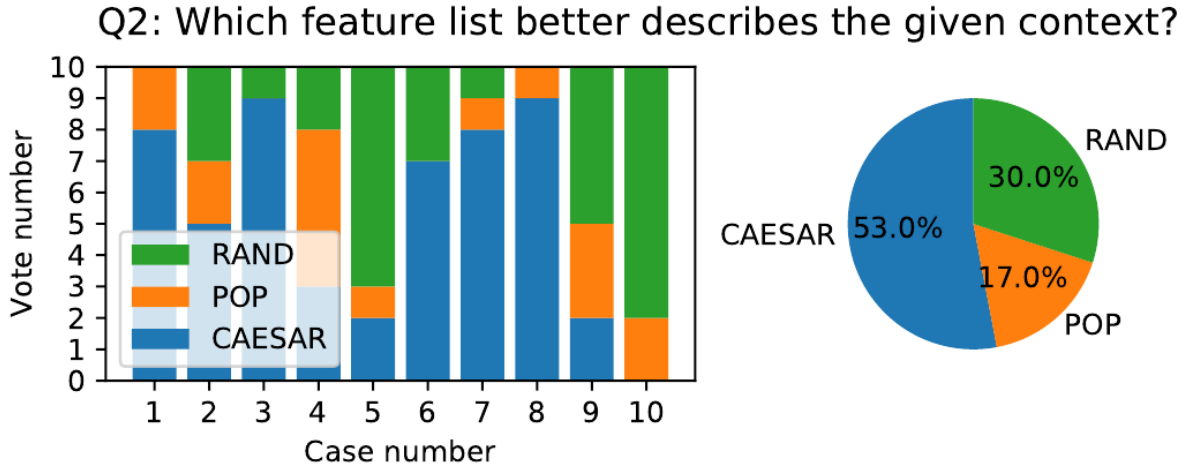


(d) Features according to occurring frequency

Human Evaluation on Explanations

- Context-aware explanations are more helpful than context-unaware explanations.





Summary

- Existing explainable recommendation approaches rarely consider context for producing explanations.
- We developed a new recommendation approach based on attention mechanism that can produce context-aware feature-level explanations.
- We also designed an effective contextual feature mining approach to identify context-aware features from user reviews.

Outline

- Introduction
- Context-aware Explanation
- **Neural Template Explanation Generation**
  - WWW'20 (demo) & CIKM'20
- Natural Language Explanation Generation
- Explanation Ranking
- Future Work

Existing Natural Language Explanation

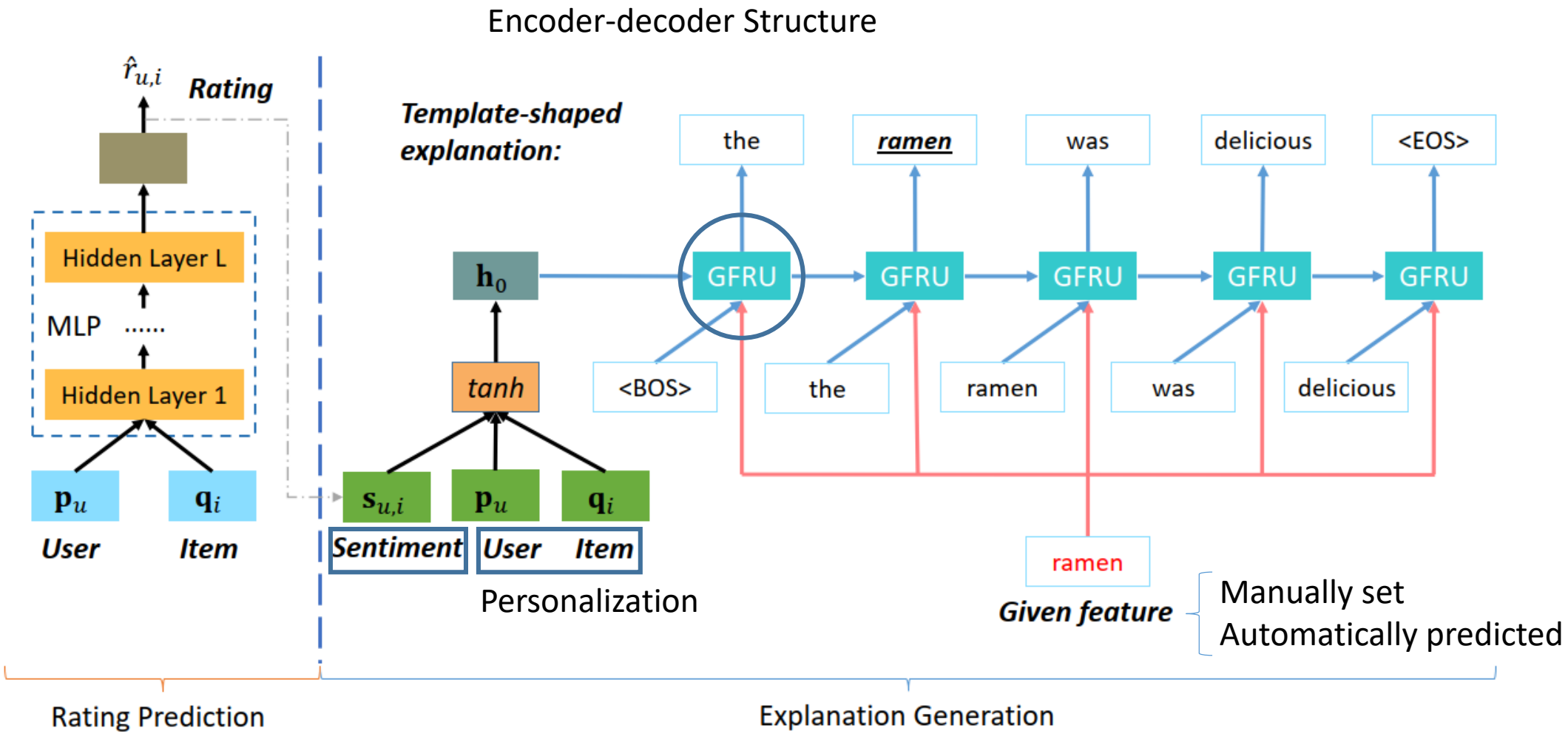
- Pre-defined templates
  - Human effort required
  - Explanation expressiveness limited
- Generated sentences
  - Similar or even identical
  - Sometimes irrelevant to the recommendation

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.

How to combine them?

Overview of Our Neural Template Approach



Gated Fusion Recurrent Unit (GFRU)

- Two Gated Recurrent Units (GRU) (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

$$\hat{\mathbf{h}}_t^{\alpha} = \tanh(\mathbf{W}_{\alpha}\mathbf{h}_t^{\alpha})$$

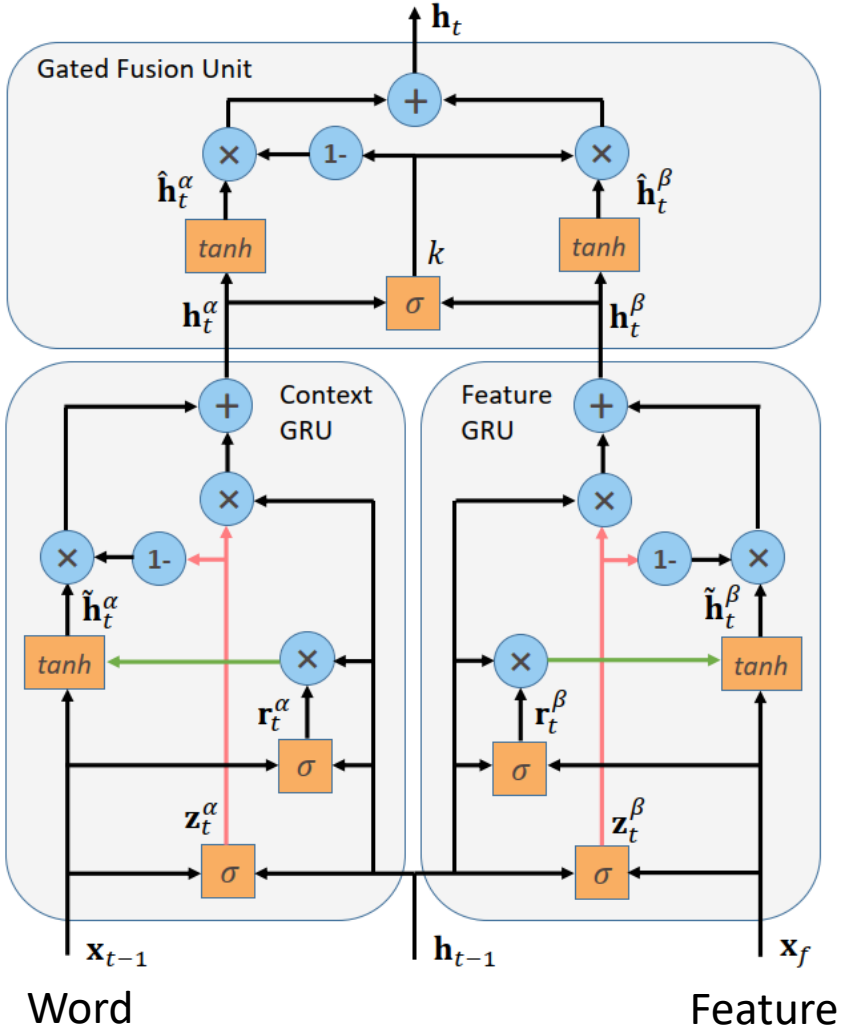
$$\hat{\mathbf{h}}_t^{\beta} = \tanh(\mathbf{W}_{\beta}\mathbf{h}_t^{\beta})$$

$$k = \sigma(\mathbf{w}_k^{\top}[\hat{\mathbf{h}}_t^{\alpha}, \hat{\mathbf{h}}_t^{\beta}])$$

$$\mathbf{h}_t = (1-k) \odot \mathbf{h}_t^{\alpha} + k \odot \mathbf{h}_t^{\beta}$$

Large -> Template

Small -> Feature



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 a feature's relevance to the user's preferred features

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

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

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

- Two times better than randomly selecting target item's features

Datasets Construction

- Three domains
  - Hotel
  - Restaurant
  - Movies & TV
- Explanations are sentences extracted from reviews
  - Contain item features

	TripAdvisor	Yelp	Amazon
# 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

Adopted by (Cai, ICDM'21; Zhou et al., 2021; Hu et al., 2021)

Evaluation Metrics

- Text quality
  - BLEU (Papineni et al., ACL'02) in machine translation
  - ROUGE (Lin, ACL'04 Workshop) in text summarization
- **Explainability**: previous work mostly ignored, so we design 4 new metrics
  - Unique Sentence Ratio (USR)
  - Feature Matching Ratio (FMR)
  - Feature Coverage Ratio (FCR)
  - Feature Diversity (DIV)

$$USR = |\mathcal{E}| / N$$

$$FMR = \frac{1}{N} \sum_{u,i} \delta(f_{u,i} \in \hat{E}_{u,i})$$

Adopted by (Hu et al., 2021)

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

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

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
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-GRU	0.38	-	0.11	2.34	12.10	0.95	20.16	12.93	14.93	2.63	1.75	1.97
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 method consistently achieves the best performance on three datasets

Quantitative Analysis on Explanations (2)

	Personalization				BLEU (%)		ROUGE-1 (%)			ROUGE-2 (%)		
	USR	FMR	FCR	DIV	BLEU-1	BLEU-4	Precision	Recall	F1	Precision	Recall	F1
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-GRU	0.27	-	0.15	3.00	13.84	0.92	18.55	13.64	15.02	2.23	1.76	1.86
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

- USR different but BLEU and ROUGE close
  - BLEU and ROUGE cannot properly evaluate sentence diversity
  - We are motivated 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
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-GRU	0.27	-	0.15	3.00	13.84	0.92	18.55	13.64	15.02	2.23	1.76	1.86
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

GFRU

- Most similar to ground-truth
  - Informativeness of the features
  - Effectiveness of our GFRU

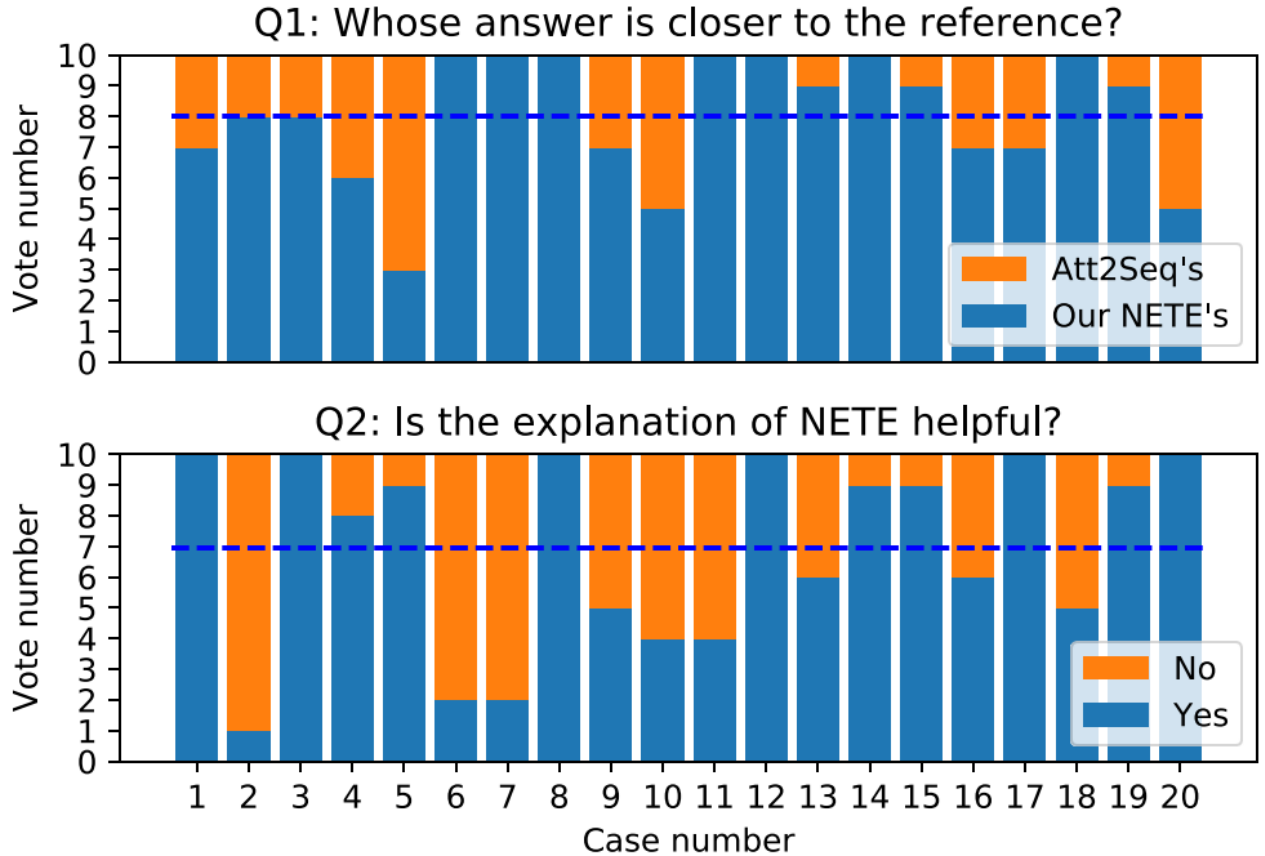
Qualitative Case Study on Explanations

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

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**.

Human Evaluation on Yelp

- High-quality explanations relative to baseline
- Helpful to better understand the recommendations



Summary

- Bridge the merits of template and generation approaches
  - Generate neural template explanations
  - Improve the expressiveness and quality of explanations
- Design four novel metrics
  - Particularly care about the explainability of generated explanations
- Show the controllability of our model
  - Generate explanations about the given user, item, sentiment, and features

Outline

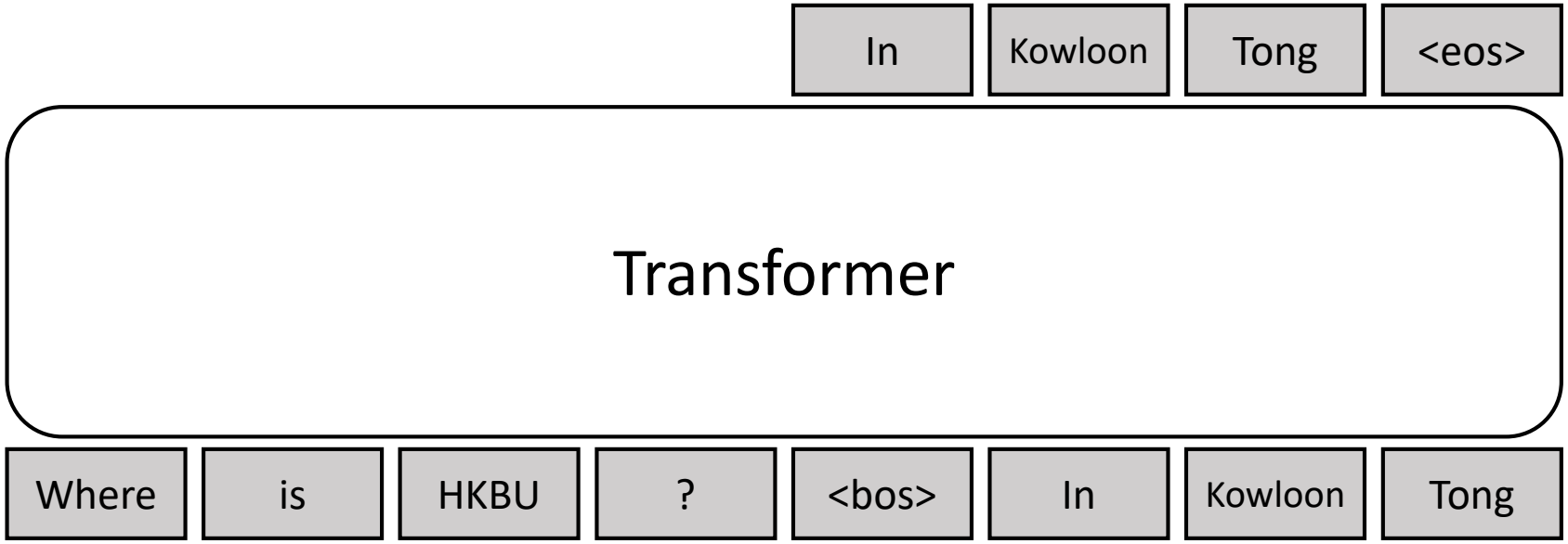
- Introduction
- Context-aware Explanation
- Neural Template Explanation Generation
- **Natural Language Explanation Generation**
  - ACL'21
- Explanation Ranking
- Future Work

Motivation

- To generate neural template explanation, an item feature must be specified
  - Location
  - Breakfast
- Problems
  - What if there is no feature?
  - What if there are multiple features?
  - How to accommodate any number of features?

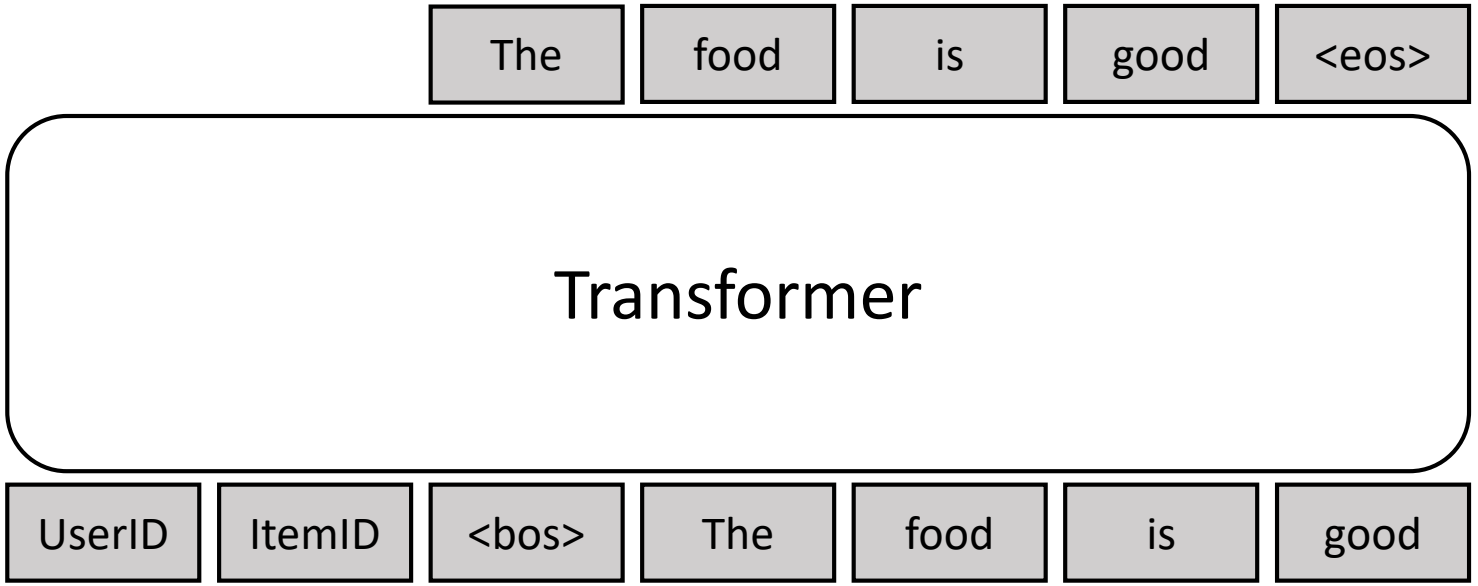
Transformer (Vaswani et al., NIPS'17)

- A well-known model employed in many fields
- Auto-regressive natural language generation
  - Predict future tokens based on past tokens



Problem for Explanation Generation

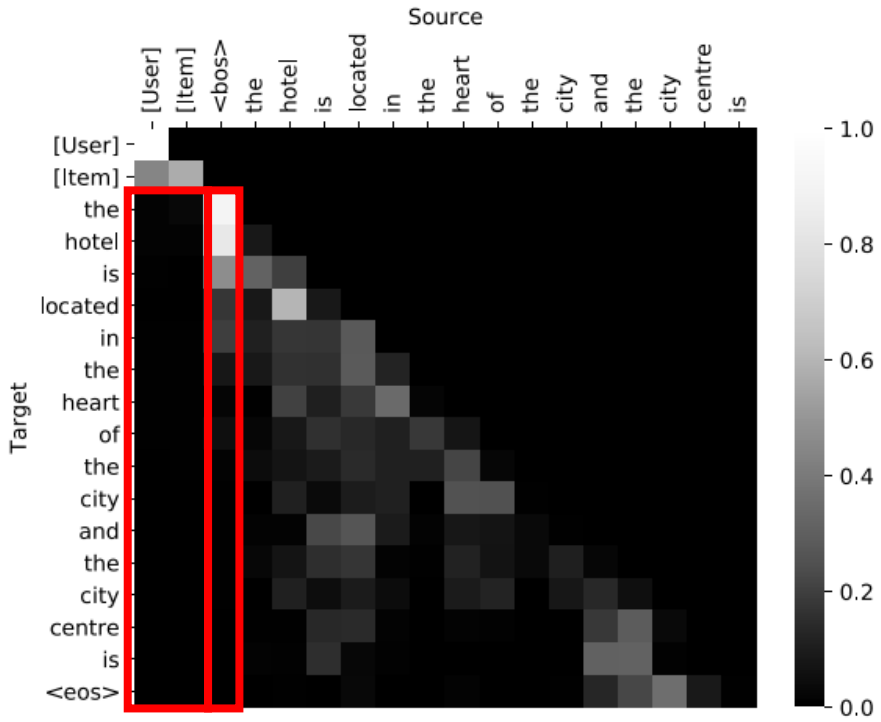
- Consider IDs as tokens, like words, and perform auto-regressive generation



Why “the food is good” for almost every user-item pair?

Attention Visualization

- The model relies heavily on <bos> for generation
- Attention weights of userID and itemID are 0
  - Model insensitive to IDs



Why insensitive?

Problem Analysis

- Frequency mismatch between IDs and words
  - One user/item ID vs. hundreds of words in a review
  - An ID appears only a few times
- IDs being regarded as uncommon words (OOV tokens)

12/4/2019

6 photos

Ho Lee Fook was one of the best food spots I went to in HK. At first I was skeptical because sometimes the fusion or westernized type Asian restaurants are all for the look but don't taste great. But, Ho Lee Fook was beautiful inside and the food was amazing. We ordered the pan fried thick rolled noodles and the massive bone steak (forgot the actually name) but you won't miss it on the menu. The noodles were crispy and seasoned just right. The steak was so tender and delicious. It came with a jalapeño sauce on the plate which complimented it so well.

While being here I forgot I was in HK because everyone spoke English and the menu was also in English! The entrance is so cute with the lucky cats all on the walls.

If you are visiting HK or live there I definitely recommend giving this place a try! It is a little on the pricey side but for the atmosphere it is expected.

A restaurant review (yelp.com)

Solution: Context Prediction

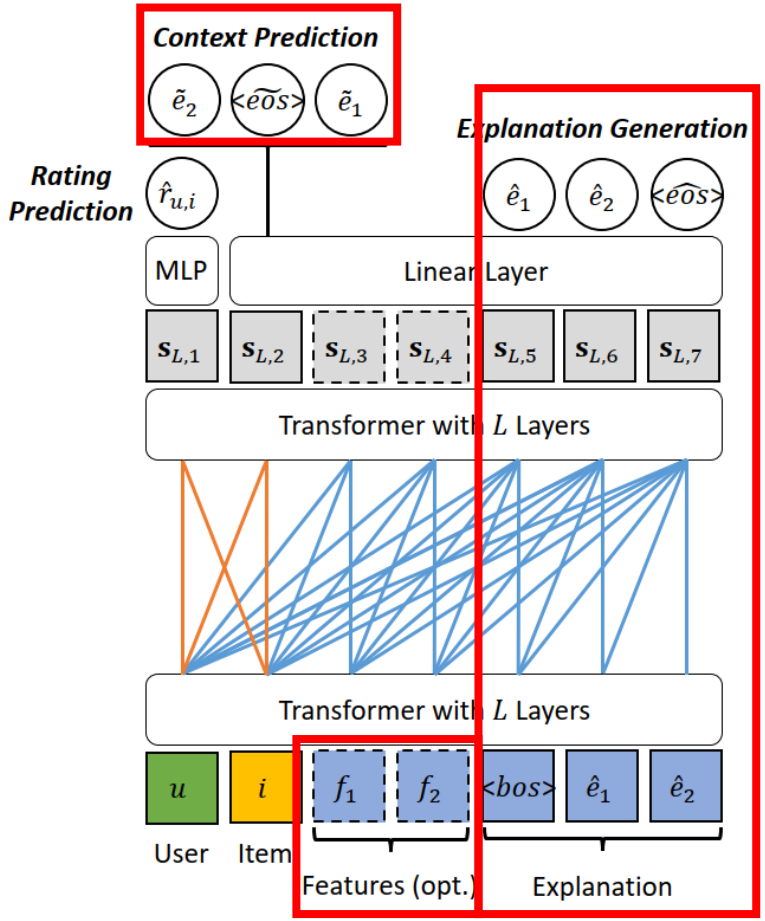
- Bridge IDs and words, and give the former linguistic meanings
- Difference
  - Context prediction: predict explanation words in one step

$$\mathcal{L}_c = \frac{1}{|\mathcal{T}|} \sum_{(u,i)\in\mathcal{T}} \frac{1}{|E_{u,i}|} \sum_{t=1}^{|E_{u,i}|} -\log c_2^{e_t}$$

  - Explanation generation: generate them one by one

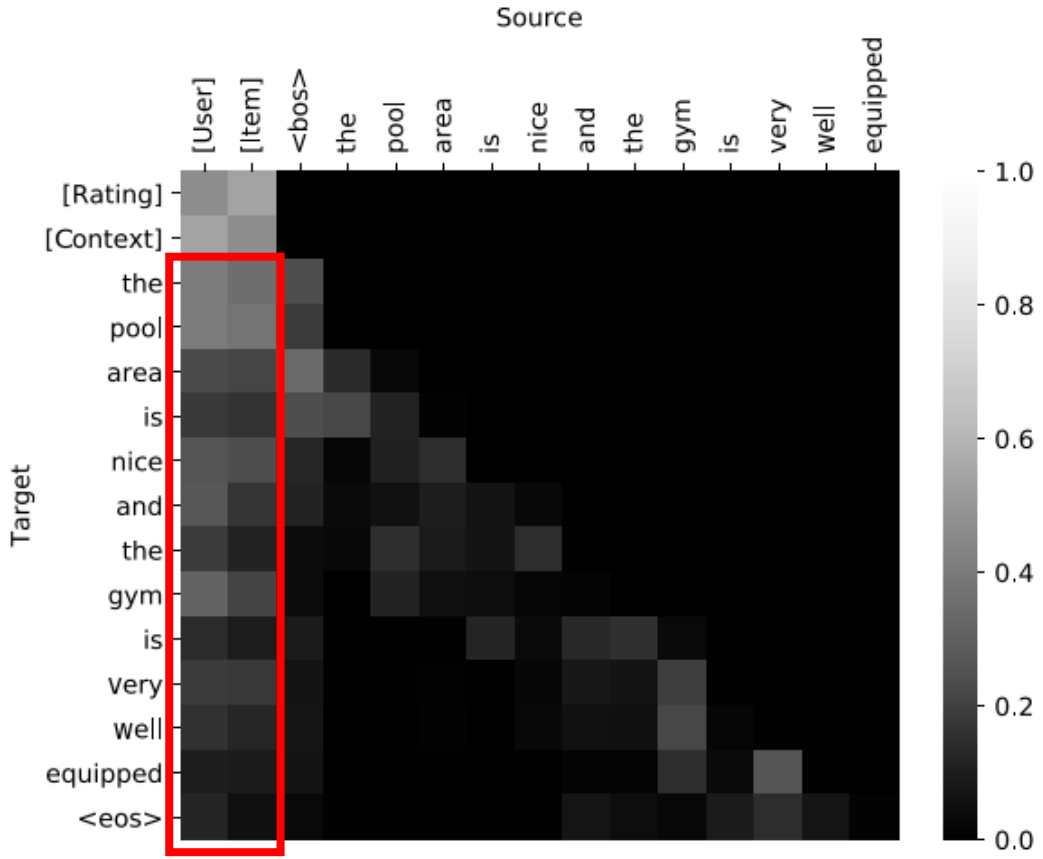
$$\mathcal{L}_e = \frac{1}{|\mathcal{T}|} \sum_{(u,i)\in\mathcal{T}} \frac{1}{|E_{u,i}|} \sum_{t=1}^{|E_{u,i}|} -\log c_{2+|F_{u,i}|+t}^{e_t}$$

- Incorporate any number of features for targeted generation: none, one, or multiple



Attention Visualization Again

- Our model can well utilize IDs for generation



Experimental Settings (Li et al., CIKM'20)

- Datasets
  - Yelp
  - Amazon
  - TripAdvisor
- Metrics
  - Text quality: BLEU & ROUGE
  - Explainability from the angle of item features
    - Unique Sentence Ratio (USR)
    - Feature Matching Ratio (FMR)
    - Feature Coverage Ratio (FCR)
    - Feature Diversity (DIV)

Quantitative Analysis on Explanations

	Explainability			Text Quality								
	FMR↑	FCR↑	DIV↓	USR↑	B1↑	B4↑	R1-P↑	R1-R↑	R1-F↑	R2-P↑	R2-R↑	R2-F↑
Transformer	0.06	0.06	2.46	0.01	7.39	0.42	**19.18**	10.29	12.56	1.71	0.92	1.09
NRT	0.07	0.11	2.37	0.12	**11.66**	0.65	17.69	12.11	13.55	1.76	1.22	1.33
Att2Seq	0.07	0.12	2.41	**0.13**	10.29	0.58	18.73	11.28	13.29	1.85	1.14	1.31
PETER	**0.08****	**0.19****	**1.54****	**0.13**	10.77	**0.73****	18.54	**12.20**	**13.77****	**2.02****	**1.38****	**1.49****
ACMLM	0.05	0.31	**0.95**	**0.95**	7.01	0.24	7.89	7.54	6.82	0.44	0.48	0.39
NETE	0.80	0.27	1.48	0.52	19.31	2.69	33.98	22.51	25.56	8.93	5.54	6.33
PETER+	**0.86****	**0.38****	1.08	0.34	**20.80****	**3.43****	**35.44****	**26.12****	**27.95****	**10.65****	**7.44****	**7.94****

IDs only

With features

Less useful, if unable to guarantee text quality

Ours the best or comparable

Qualitative Case Study on Explanations

- Context prediction task can indeed give IDs linguistic meanings
- Two tasks resemble one's drafting-polishing process
- The incorporated features further improve text quality

	Top-15 Context Words	Explanation
Ground-truth		the **rooms** are spacious and the bathroom has a large tub
PETER	<eos> the and a pool was with nice is very were to good in of	the pool area is nice and the gym is very well equipped <eos>
PETER+	<eos> the and a was pool with to nice good very were is of in	the rooms were clean and comfortable <eos>
Ground-truth		beautiful **lobby** and nice bar
PETER	<eos> the and a was were separate bathroom with shower large very had in is	the bathroom was large and the shower was great <eos>
PETER+	<eos> the and a was bathroom shower with large in separate were room very is	the lobby was very nice and the rooms were very comfortable <eos>

Summary

- Propose a general explanation generation approach
  - Accommodate any number of item features
- Enable Transformer with personalized natural language generation
  - Shed light on other fields that also need personalization, e.g., personalized conversational systems
- Design a task to connect IDs and words
  - Point out a way for Transformer to deal with heterogeneous data, e.g., image generation based on text in multi-modal AI

Beyond: Image Generation

- Adopt our PETER model as the backbone (Geng et al., ACL'22)
- Key idea: convert an image into a sequence of tokens as if a sentence

Inputs:
User A, Item 1, Feat. word: floors
Outputs:
Pred. rating: 4.62
Gen. explanation: higher floors have better view
Image visualization:

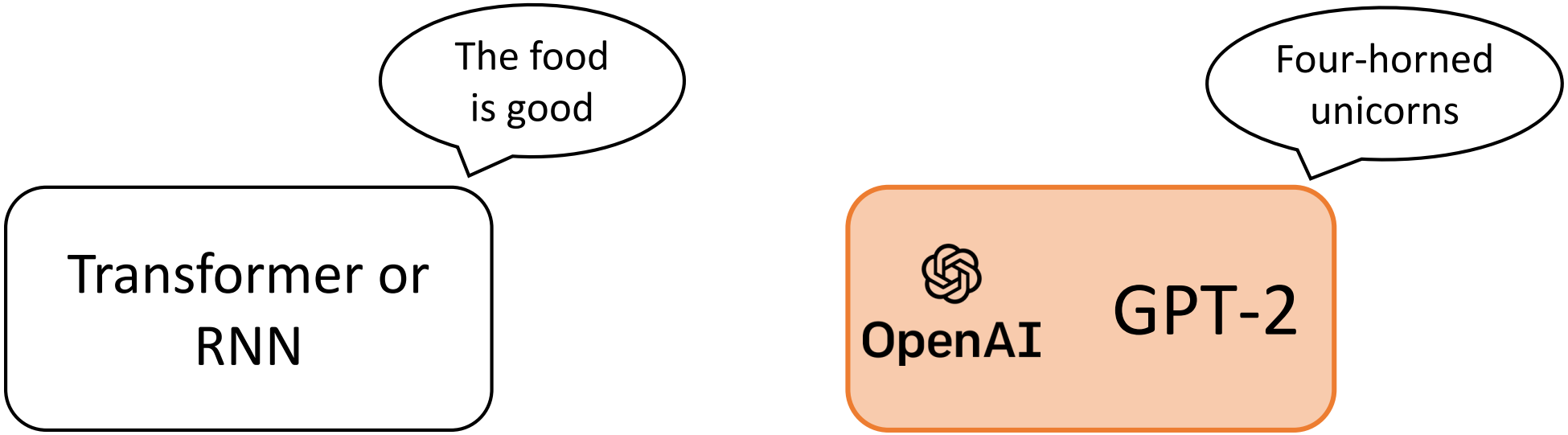
Inputs:
User B, Item 2, Feat. word: seat
Outputs:
Pred. rating: 4.15
Text explanation: we were seated immediately and ordered our food
Image visualization:

Outline

- Introduction
- Context-aware Explanation
- Neural Template Explanation Generation
- Natural Language Explanation Generation
- **Explanation Ranking**
  - SIGIR'21 (resource) & TIST 2022 (submitted)
- Future Work

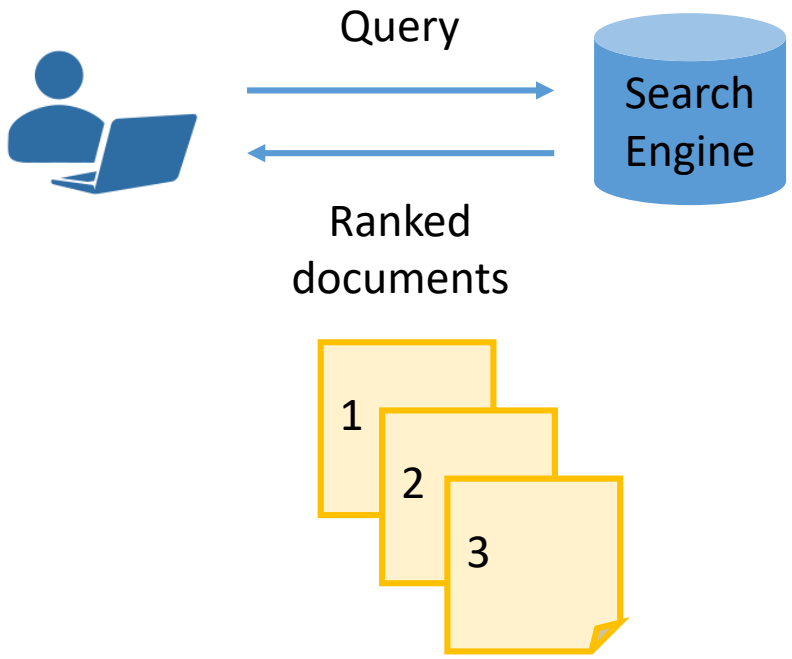
Problems of Natural Language Generation

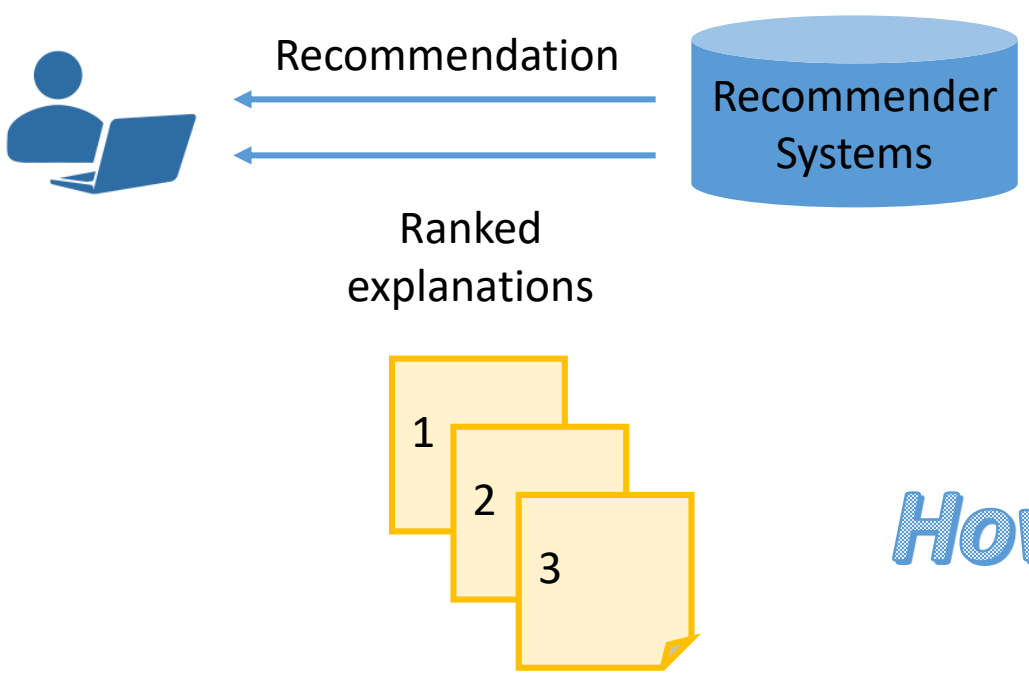
- Fit the given samples rather than creating new explanations
- Sometimes deviate from the facts



Information Retrieval vs. Explanation Ranking

- Rank available documents
- Enable standard evaluation via ranking metrics



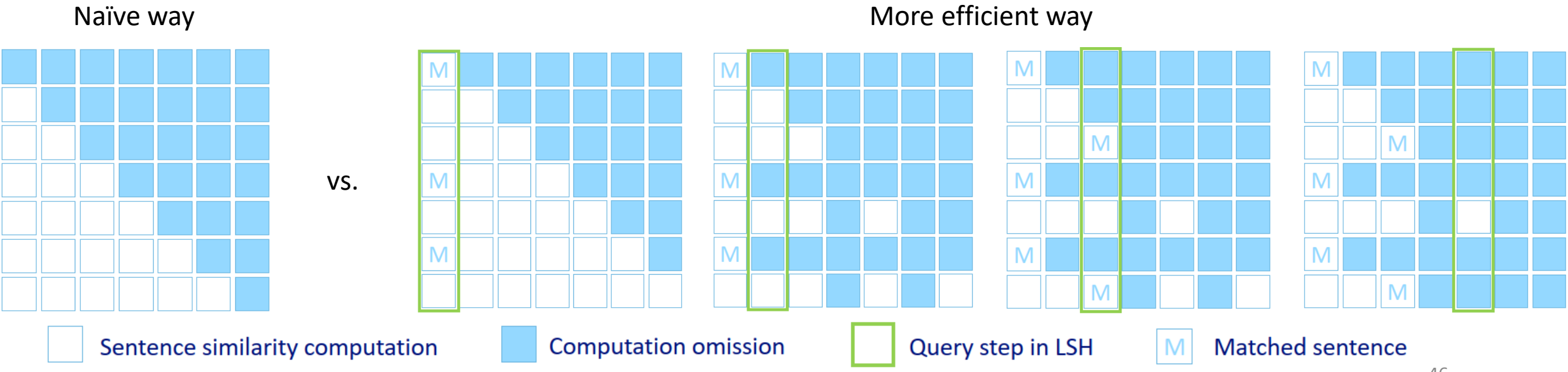


Wisdom of the Crowd

- Detect co-occurring sentences across reviews
- Create user-item-explanation interactions
- Allow to design collaborative filtering algorithms

Near-duplicate Detection

- Quadratic time complexity for comparing any two sentences
  - Conduct near-duplicate detection in sub-linear time with Locality-Sensitive Hashing (LSH) (Rajaraman and Ullman, 2011)
  - Remove already matched sentences



Datasets Construction

- Explanations
  - Concise and informative
  - Well suit target application domains
- Interaction records very sparse

	Amazon	TripAdvisor	Yelp
# of users	109,121	123,374	895,729
# of items	47,113	200,475	164,779
# of explanations	33,767	76,293	126,696
# of (u, i) pairs	569,838	1,377,605	2,608,860
# of (u, i, e) triplets	793,481	2,618,340	3,875,118
# of explanations / (u, i) pair	1.39	1.90	1.49
Density ($\times 10^{-10}$)	45.71	13.88	2.07

Explanation	Occurrence
Amazon Movies & TV	
Great story	3307
Don't waste your money	834
The acting is great	760
The sound is okay	11
A wonderful movie for all ages	6

Explanation	Occurrence
TripAdvisor	
Great location	61993
The room was clean	6622
The staff were friendly and helpful	2184
Bad service	670
Comfortable hotel with good facilities	8

Explanation	Occurrence
Yelp	
Great service	46413
Everything was delicious	5237
Prices are reasonable	2914
This place is awful	970
The place was clean and the food was good	6

Problem Formulation

- Item recommendation

$$\mathrm{Top}(u, M) := \underset{i \in \mathcal{I}/\mathcal{I}_u}{\arg\max^{M}} \hat{r}_{u,\underline{i}}$$

- Explanation ranking

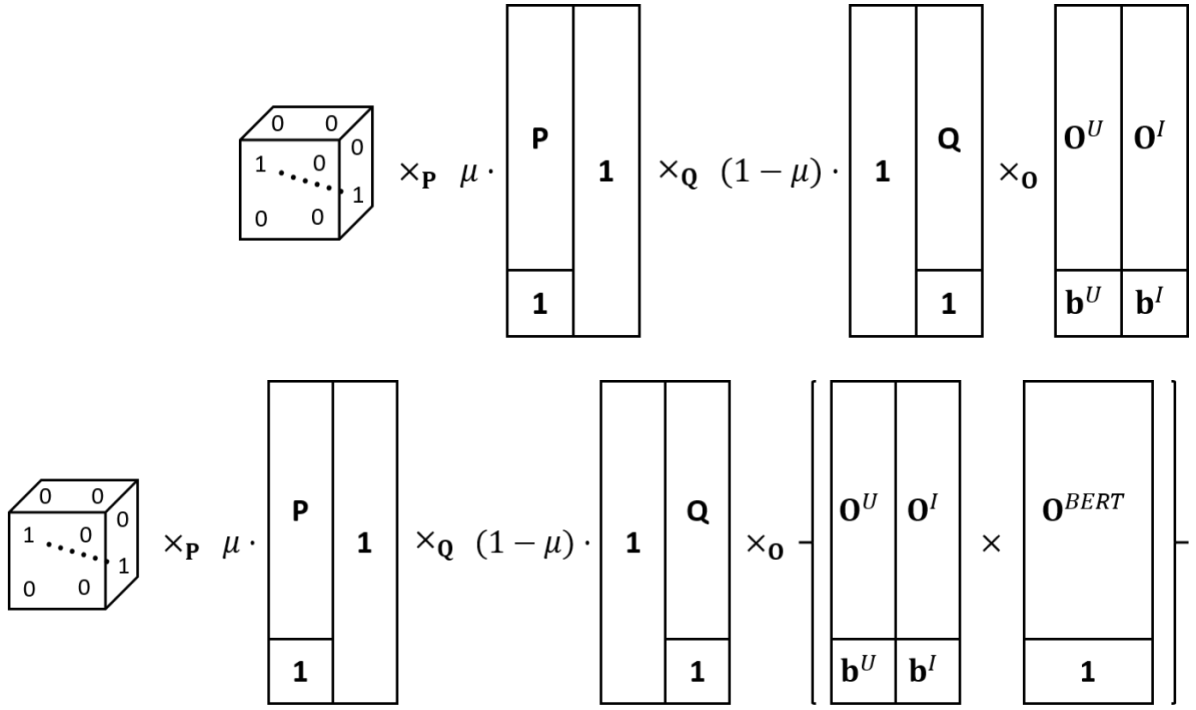
$$\mathrm{Top}(u, i, N) := \underset{e \in \mathcal{E}}{\arg\max^{N}} \hat{r}_{u,i,\underline{e}}$$

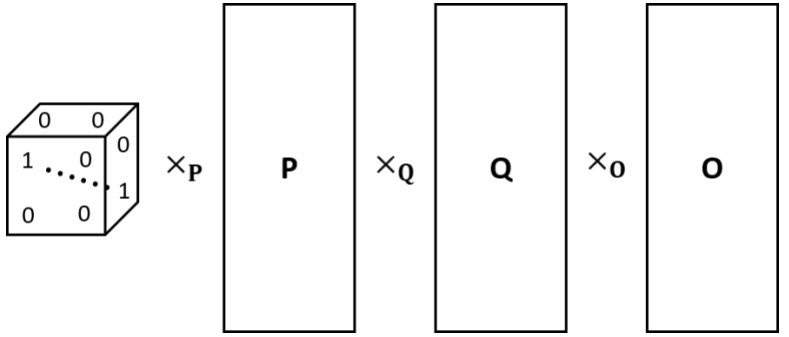
- Item-explanation joint-ranking

$$\mathrm{Top}(u, M) := \underset{i \in \mathcal{I}/\mathcal{I}_u, e \in \mathcal{E}}{\arg\max^{M}} \hat{r}_{u,\underline{i,e}}$$

Tensor Factorization vs. Matrix Factorization

- Decompose user-item-explanation (TF) into user-explanation (MF) and item-explanation (MF) to address data sparsity issue
  - Leverage user, item, and explanation IDs only
  - Incorporate explanation text with BERT (Devlin et al., NAACL'19)





Results of Explanation Ranking

- Both approaches are very effective

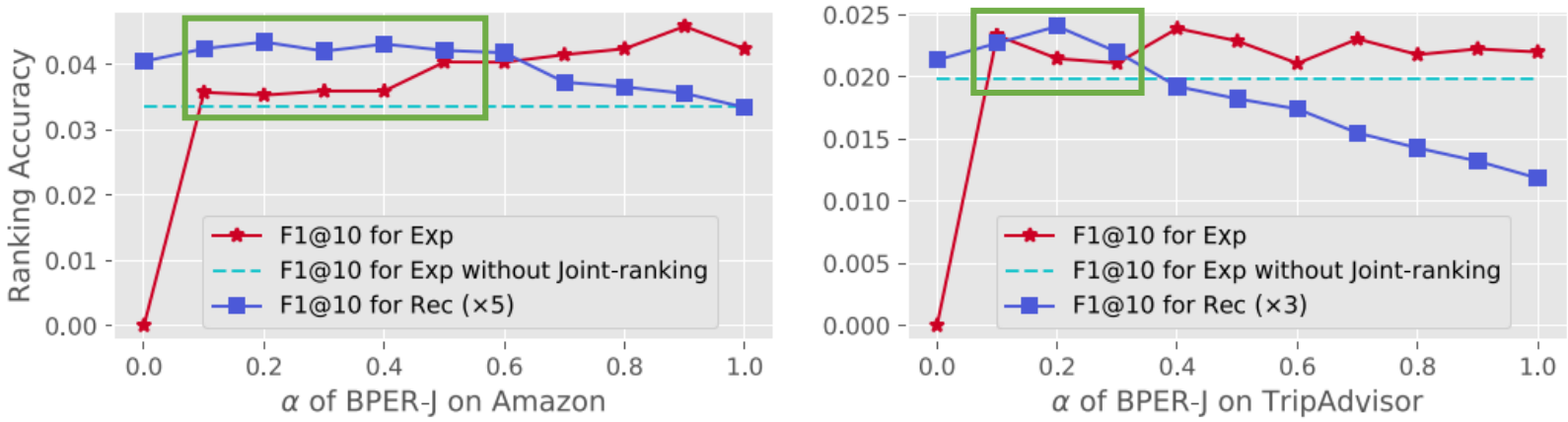
	Amazon				TripAdvisor				Yelp			
	NDCG@10	Pre@10	Rec@10	F1@10	NDCG@10	Pre@10	Rec@10	F1@10	NDCG@10	Pre@10	Rec@10	F1@10
CD	0.001	0.001	0.007	0.002	0.001	0.001	0.003	0.001	0.000	0.000	0.003	0.001
RAND	0.004	0.004	0.027	0.006	0.002	0.002	0.011	0.004	0.001	0.001	0.007	0.002
RUCF	0.341	0.170	1.455	0.301	0.260	0.151	0.779	0.242	0.040	0.020	0.125	0.033
RICF	0.417	0.259	1.797	0.433	0.031	0.020	0.087	0.030	0.037	0.026	0.137	0.042
PITF	2.352	1.824	14.125	3.149	1.239	1.111	5.851	1.788	0.712	0.635	4.172	1.068
BPER	2.630*	**1.942***	**15.147***	**3.360***	1.389*	1.236*	6.549*	1.992*	0.814*	0.723*	**4.768***	1.218*
BPER+	**2.877***	1.919*	14.936*	3.317*	**2.096***	**1.565***	**8.151***	**2.515***	**0.903***	**0.731***	4.544*	**1.220***
Improvement (%)	22.352	5.229	5.739	5.343	69.073	40.862	39.314	40.665	26.861	15.230	8.925	14.228

Item-explanation Joint-ranking

- Purposely select some explanations to improve the chance of clicking/purchasing

$$\min_{\Theta} \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}_u} \Big[\sum_{i' \in \mathcal{I} / \mathcal{I}_u} -\ln \sigma(\hat{r}_{u,ii'}) + \alpha \sum_{e \in \mathcal{E}_{u,i}} \Big(\sum_{e' \in \mathcal{E} / \mathcal{E}_u} -\ln \sigma(\hat{r}_{u,ee'}) + \sum_{e'' \in \mathcal{E} / \mathcal{E}_i} -\ln \sigma(\hat{r}_{i,ee''}) \Big) \Big] + \lambda ||\Theta||_F^2$$

- Improve both recommendation and explanation performance



Summary

- Formulate the recommendation explanation problem as ranking task
- Attempt to achieve standard offline evaluation of explainability
- Construct three large datasets for explanation ranking
- Develop two effective models to address the data sparsity issue
- Study the relation between explanation and recommendation via the item-explanation joint-ranking

Conclusion

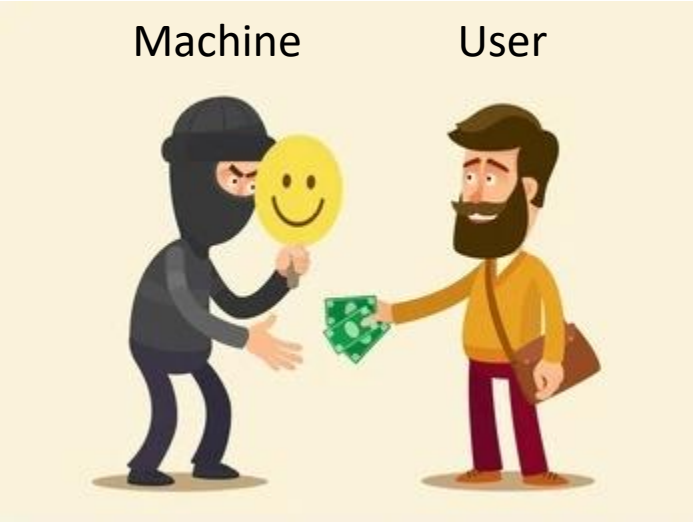
- 1 topic: explainable recommendation
- 2 sets of datasets: natural language generation, explanation ranking
- 3 explanation formats: template, generation, ranking
- 4 approaches: attention, RNN, transformer, tensor factorization
- 5 published papers: JIIS 2021, WWW'20 (demo), CIKM'20, ACL'21, SIGIR'21 (resource), TIST 2022 (submitted)
  - Other first-author papers: ICDE'19 (workshop), RecSys'22 (submitted), TOIS 2022 (submitted)

Outline

- Introduction
- Context-aware Explanation
- Neural Template Explanation Generation
- Natural Language Explanation Generation
- Explanation Ranking
- **Future Work**
  - Ethical Issue
  - Bias and Fairness

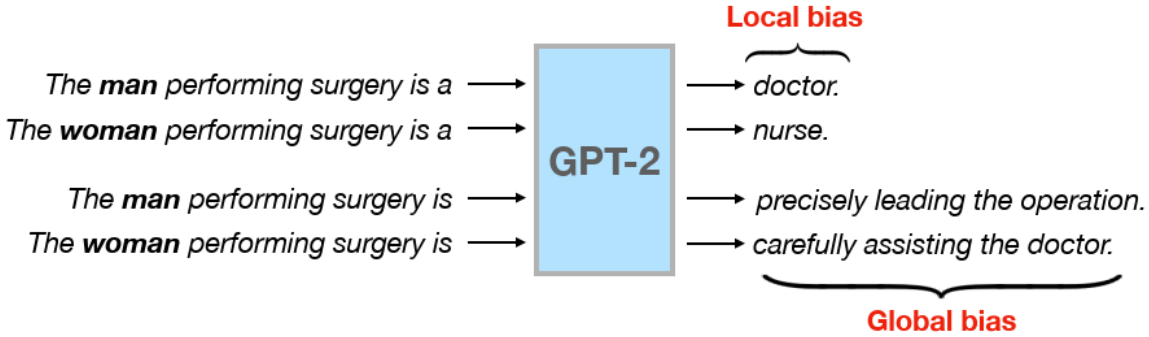
Ethical Issue in Explanation Models

- In the joint-ranking formulation, purposely selected explanations could help improve recommendation accuracy.
  - Are they faithful to the recommendations?
  - What if they are chosen simply because they can lure and manipulate user's clicking/purchasing?



Bias in Natural Language Generation

- Bias in Pre-trained model GPT-2 (Liang et al., ICML'21)



- Does such bias still exist or could it be amplified, when adapted to downstream tasks?
  - Recommendation explanation generation
- How to mitigate the bias in order to achieve fairer and more inclusive machine learning?

Interpretability of Pre-trained Models

- In what form does the bias exist in pre-trained models?
  - Transparency
  - Fairness
- Potential applications
  - Recommender systems
  - Information retrieval systems
  - Conversational systems
  - Image captioning systems

References (1)

- Zhang, Yongfeng, and Xu Chen. "Explainable recommendation: A survey and new perspectives." Foundations and Trends in Information Retrieval. 2020.
- Tintarev, Nava, and Judith Masthoff. "Explaining recommendations: Design and evaluation." Recommender systems handbook. 2015.
- Abowd, Gregory D., et al. "Towards a better understanding of context and context-awareness." HUC'99.
- Mei, Lei, et al. "An attentive interaction network for context-aware recommendations." CIKM'18.
- Zhang, Yongfeng, et al. "Do users rate or review? Boost phrase-level sentiment labeling with review-level sentiment classification." SIGIR'14.
- Luong, Minh-Thang, et al. "Effective approaches to attention-based neural machine translation." EMNLP'15.
- Liu, Shulin, et al. "Exploiting argument information to improve event detection via supervised attention mechanisms." ACL'17.
- Sarwar Badrul, et al. "Item-based collaborative filtering recommendation algorithms." WWW'01.

References (2)

- Zhang, Yongfeng, et al. "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis." SIGIR’14.
- Li, Piji, et al. "Neural rating regression with abstractive tips generation for recommendation." SIGIR’17.
- Dong, Li, et al. "Learning to generate product reviews from attributes." EACL’17.
- Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." EMNLP’14.
- Arevalo, John, et al. "Gated multimodal units for information fusion." ICLR’17 Workshop.
- Cai, Zerui. "Generating Explanations for Recommendation Systems via Injective VAE." ICDM’21.
- Zhou, Yao, et al. "From Intrinsic to Counterfactual: On the Explainability of Contextualized Recommender Systems." arXiv:2110.14844, 2021.
- Hu, Yidan, et al. "Hierarchical Aspect-guided Explanation Generation for Explainable Recommendation." arXiv:2110.10358, 2021.

References (3)

- Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." ACL'02.
- Lin, Chin-Yew. "ROUGE: A Package for Automatic Evaluation of Summaries." ACL'04 Workshop.
- Vaswani, Ashish, et al. "Attention is all you need." NIPS'17.
- Li, Lei, et al. "Generate neural template explanations for recommendation." CIKM'20.
- Geng, Shijie, et al. "Improving Personalized Explanation Generation through Visualization." ACL'22.
- Rajaraman, Anand, and Jeffrey David Ullman. "Finding similar items." Mining of massive datasets. 2011.
- Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." NAACL'19.
- Liang, Paul Pu, et al. "Towards understanding and mitigating social biases in language models." ICML'21.

Q&A

Thank you!