Generate Neural Template Explanations for Recommendation

Lei Li\textsuperscript{1}, Yongfeng Zhang\textsuperscript{2}, Li Chen\textsuperscript{1}

\textsuperscript{1}Hong Kong Baptist University, \textsuperscript{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)

Social network
(instagram.com)

Video
(youtube.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

<table>
<thead>
<tr>
<th>Images (Chen et al. SIGIR’19)</th>
<th>Neighbors (Li et al. WWW’20)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image Example" /></td>
<td><img src="image2.png" alt="Neighbor Example" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Image Example" /></td>
<td><img src="image4.png" alt="Neighbor Example" /></td>
</tr>
</tbody>
</table>

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

#### Features (He et al. CIKM’15)

<table>
<thead>
<tr>
<th>Speciality</th>
<th>Your Preference</th>
</tr>
</thead>
<tbody>
<tr>
<td>fries</td>
<td>🍔amburgers 🍔burgers</td>
</tr>
<tr>
<td>chicken</td>
<td>🍗ay 🔞ay 🔞ay</td>
</tr>
<tr>
<td>sauce</td>
<td>🍤sauce 🔞sauce</td>
</tr>
<tr>
<td>location</td>
<td>📍location 🔞location</td>
</tr>
<tr>
<td>cheese</td>
<td>🧀cheese 🔞cheese</td>
</tr>
</tbody>
</table>

Dislike the recommendation? Change your preference here.

#### Templates (Zhang et al. SIGIR’14)

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF (Sarwar et al. WWW’01)</td>
<td>Customers who bought this item also bought.</td>
</tr>
<tr>
<td>EFM (Zhang et al. SIGIR’14)</td>
<td>You might be interested in [feature], on which this product performs well.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>They have a huge <strong>variety</strong> of things.</td>
</tr>
<tr>
<td>NRT (Li et al. SIGIR’17)</td>
<td>The food is good.</td>
</tr>
<tr>
<td>Att2Seq (Dong et al. EACL’17)</td>
<td>I’m not sure if I need to go back.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>The black garlic <strong>ramen</strong> was good as well.</td>
</tr>
<tr>
<td>NRT</td>
<td>The food is good.</td>
</tr>
<tr>
<td>Att2Seq</td>
<td>The food was great.</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Reference</th>
<th>They have a huge <strong>variety</strong> of things.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NETE</td>
<td>They have a <strong>variety</strong> of things to choose from.</td>
</tr>
<tr>
<td>Reference</td>
<td>The black garlic <strong>ramen</strong> was good as well.</td>
</tr>
<tr>
<td>NETE</td>
<td>The <strong>ramen</strong> was delicious.</td>
</tr>
</tbody>
</table>
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

Rating Prediction

Explanation Generation

Template-controlled explanation:

\[
\hat{f}_{u,i} \quad \text{Rating}
\]

Hidden Layer 1

MLP

Hidden Layer 1

User

Item

Sentiment

User

Item

\[
\begin{align*}
\mathbf{p}_u & \quad \mathbf{q}_i \\
\mathbf{s}_{u,i} & \quad \mathbf{p}_u & \quad \mathbf{q}_i
\end{align*}
\]

\[
\mathbf{h}_0 \quad \text{tanh}
\]

GFRU

Given feature

\[
\text{the} \quad \text{ramen} \quad \text{was} \quad \text{delicious} \quad \text{<EOS>}
\]
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)

\[
\begin{align*}
    z_1 &= \sigma(W_1[p_u, q_i] + b_1) \\
    z_2 &= \sigma(W_2 z_1 + b_2) \quad \text{and} \quad \hat{r}_{u,i} = w_r z_L + b_r \\
    \vdots \\
    z_L &= \sigma(W_L z_{L-1} + b_L)
\end{align*}
\]

• 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
    \[
    h_0 = \tanh(W_e[p_u, q_i, s_{u,i}] + b_e)
    \]
- 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

\[
\begin{align*}
\hat{h}_t^\alpha &= \tanh(W_\alpha h_t^\alpha) \\
\hat{h}_t^\beta &= \tanh(W_\beta h_t^\beta) \\
k &= \sigma(w_k[\hat{h}_t^\alpha, \hat{h}_t^\beta]) \\
h_t &= (1 - k) \odot h_t^\alpha + k \odot h_t^\beta
\end{align*}
\]
Explanation Generation (2)

• Hidden states of each time step can be computed by GFRU
  \[ h_t = g(x_{t-1}, h_{t-1}, x_f) \]

• During decoding, a word with the largest probability over the vocabulary is sampled
  \[ p(y_t | y_{<t}, h_0) = \text{softmax}_{y_t}(W_v h_t + b_v) \]

• Cross-entropy loss function
  \[ \mathcal{L}_e = \frac{1}{|\mathcal{T}|} \sum_{u, i \in \mathcal{T}} \frac{1}{|S_{u,i}|} \sum_{i=1}^{|S_{u,i}|} - \log p(y_t) \]
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

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

\[
\hat{f}_i = \arg\max_{f \in F_i} \text{PMI}(F_u, f)
\]
\[
\text{PMI}(F_u, f) = \log \frac{p(F_u|f)}{p(F_u)} \approx \log \frac{\prod_{f' \in F_u} p(f'|f)}{\prod_{f' \in F_u} p(f')} = \sum_{f' \in F_u} \log \frac{p(f'|f)}{p(f')} = \sum_{f' \in F_u} \text{PMI}(f', f)
\]
\[
\text{PMI}(f_u, f_i) = \log \frac{p(f_u, f_i)}{p(f_u)p(f_i)} = \log \frac{p(f_u|f_i)}{p(f_u)}
\]

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

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

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

* 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)

\[
USR = \frac{|S|}{N} \quad FMR = \frac{1}{N} \sum_{u,i} \delta(f_{u,i} \in \hat{S}_{u,i}) \\
FCR = \frac{N_g}{|\mathcal{F}|} \\
DIV = \frac{2}{N \times (N - 1)} \sum_{u,u',i,i'} \left| \hat{\mathcal{F}}_{u,i} \cap \hat{\mathcal{F}}_{u',i'} \right|
\]
Ablation Study

• Investigate the impacts of different settings

<table>
<thead>
<tr>
<th></th>
<th>Learning Framework</th>
<th>Decoder</th>
<th>Given Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single-task</td>
<td>Multi-task</td>
<td>GFRU</td>
</tr>
<tr>
<td>NETE</td>
<td>√</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>NETE-GRU</td>
<td>√</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>NETE-MUL</td>
<td></td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>NETE-GM</td>
<td></td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>NETE-PMI</td>
<td>√</td>
<td></td>
<td>√</td>
</tr>
</tbody>
</table>
### Quantitative Analysis on Explanations (1)

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

Our methods consistently achieve the best performance on three datasets.
<table>
<thead>
<tr>
<th>Personalization</th>
<th>BLEU (%)</th>
<th>ROUGE-1 (%)</th>
<th>ROUGE-2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>USR</td>
<td>FMR</td>
<td>FCR</td>
</tr>
<tr>
<td>NRT</td>
<td>0.00</td>
<td>-</td>
<td>0.00</td>
</tr>
<tr>
<td>Att2Seq</td>
<td>0.18</td>
<td>-</td>
<td>0.17</td>
</tr>
<tr>
<td>NETE-GM</td>
<td>0.00</td>
<td>-</td>
<td>0.00</td>
</tr>
<tr>
<td>NETE-GRU</td>
<td>0.27</td>
<td>-</td>
<td>0.15</td>
</tr>
<tr>
<td>NETE-MUL</td>
<td>0.02</td>
<td>0.66</td>
<td>0.07</td>
</tr>
<tr>
<td>NETE-PMI</td>
<td>0.79</td>
<td>0.38</td>
<td>0.30</td>
</tr>
<tr>
<td>NETE</td>
<td>0.57**</td>
<td>0.78</td>
<td>0.27**</td>
</tr>
<tr>
<td>Improvement (%)</td>
<td>+210.7</td>
<td>-</td>
<td>+57.1</td>
</tr>
</tbody>
</table>

- 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)

<table>
<thead>
<tr>
<th>Personalization</th>
<th>BLEU (%)</th>
<th>ROUGE-1 (%)</th>
<th>ROUGE-2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>USR</td>
<td>FMR</td>
<td>FCR</td>
</tr>
<tr>
<td>NRT</td>
<td>0.00</td>
<td>-</td>
<td>0.00</td>
</tr>
<tr>
<td>Att2Seq</td>
<td>0.18</td>
<td>-</td>
<td>0.17</td>
</tr>
<tr>
<td>NETE-GM</td>
<td>0.00</td>
<td>-</td>
<td>0.00</td>
</tr>
<tr>
<td>NETE-GRU</td>
<td>0.27</td>
<td>-</td>
<td>0.15</td>
</tr>
<tr>
<td>NETE-MUL</td>
<td>0.02</td>
<td>0.66</td>
<td>0.07</td>
</tr>
<tr>
<td>NETE-PMI</td>
<td>0.79</td>
<td>0.38</td>
<td>0.30</td>
</tr>
<tr>
<td>NETE</td>
<td>0.57**</td>
<td>0.78</td>
<td>0.27**</td>
</tr>
<tr>
<td>Improvement (%)</td>
<td>+210.7</td>
<td>-</td>
<td>+57.1</td>
</tr>
</tbody>
</table>

- 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)

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

<table>
<thead>
<tr>
<th>Rating</th>
<th>Feature</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>rooms</td>
<td>The rooms are spacious and the bathroom has a large tub.</td>
</tr>
<tr>
<td>3.90</td>
<td>bathroom</td>
<td>The bathroom was large and had a separate shower.</td>
</tr>
<tr>
<td></td>
<td>tub</td>
<td>The bathroom had a separate shower and tub.</td>
</tr>
<tr>
<td></td>
<td>rooms</td>
<td>The rooms are large and comfortable.</td>
</tr>
<tr>
<td>4</td>
<td>rooms</td>
<td>The rooms are brilliant and ideal for business travellers.</td>
</tr>
<tr>
<td>4.13</td>
<td>rooms</td>
<td>The rooms are very spacious and the rooms are very comfortable.</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>The broken furniture and dirty surfaces are a dead giveaway.</td>
</tr>
<tr>
<td>2.96</td>
<td>furniture</td>
<td>The furniture is worn.</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Ideal for plane spotters and very close to the airport.</td>
</tr>
<tr>
<td>2.76</td>
<td>airport</td>
<td>It is not close to the airport.</td>
</tr>
<tr>
<td>Rating Prediction</td>
<td>Personalized Ranking</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>----------------------</td>
<td></td>
</tr>
<tr>
<td><strong>TA-HK</strong></td>
<td><strong>YELP19</strong></td>
<td><strong>AZ-MT</strong></td>
</tr>
<tr>
<td><strong>TA-HK</strong></td>
<td><strong>YELP19</strong></td>
<td><strong>AZ-MT</strong></td>
</tr>
<tr>
<td><strong>Baselines</strong></td>
<td><strong>NETE outperforms the others</strong></td>
<td><strong>The advantage of single-task learning</strong></td>
</tr>
</tbody>
</table>

- **Accuracy is close**
- **Not all items evaluated**
  - Data selection bias ([Harald. RecSys’13](#))
- **Performance gap widens**
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)


References (2)

Thank you!