Open-Domain Generation and Automatic Evaluation

Nanyun (Violet) Peng
Nov 4th, The 3rd Workshop on Neural Generation and Translation
Open-Domain Generation

Open-domain dialog system

System: What do you want to talk about?
User: Let’s talk about comic books.
System: I’m a pretty avid comic book reader! Who is your favorite character?
User: I really like Moon Knight
System: Moon Knight is a fictional superhero… Does it sound like we are thinking of the same character? …
User: Yep that’s the one.

Example credit: Zhang et al., 2017

Story generation given a title

Title (given): rock jumping

Story (generated):
One day, I decided to go rock jumping.
I jumped and fell.
I fell and broke my ankle.
I had to go to the hospital.
I learned to be more careful next time.
Open-Domain Generation

Text style transfer

[Source]: [Fake] mexican food and [expensive].

[Target]: [Inexpensive] and [traditional] mexican food!

[Source]: I [could barely get] it [though] they taste so [nasty].

[Target]: I [love] it [because] they taste so [great].

Creative pun generation

- **Input**: hare, hair
- **Output**: The greyhound stopped to get a hare cut.

Figure credit: Zhang et al., 2018
Open-domain Creative Generation is Challenging

- The content is under-specified. To many possibilities, high conditional entropy given the input.
  - How to generate creative and coherent contents?
Some of Our Work on Open-Domain Generation

Open-domain dialog system
Shang et al. (IJCAI 2018)

Text style transfer
Fu et al. (AAAI 2018)

Story generation given a title
Yao et al. (AAAI 2019)

Creative pun generation
He et al. (NAACL 2019)
Outline

• Creative pun generation
  – Pun generation with surprise (He et al. NAACL 2019)

• Evaluating open-domain dialog systems
  – Contextualized relevance (Ghazarian et al. NAACL NeuralGen 2019)
  – Predictive engagement (Ghazarian et al. AAAI 2020)
Some of My Favorite Puns

The greyhound stopped to get a hare cut.

The coloring was just as he said it would be, to dye for.
What Does GPT-2 Say?

<table>
<thead>
<tr>
<th>the greyhound stopped</th>
<th>after running into the crowd of people.</th>
</tr>
</thead>
<tbody>
<tr>
<td>the hound stopped</td>
<td>There are three of them, but they have gone to one of the many houses, ...</td>
</tr>
</tbody>
</table>

The coloring was just as he said, it was perfect the color.

Example credit: generated using https://talktotransformer.com/
Generating Puns is Challenging

Creative Composition

- No large corpus of puns (poems, jokes, stories) to train a generative model.
- Even if a large pun corpus exists, learning the distribution of existing data and sampling from it will likely to just *mimic/memorize*, rather than generate truly *novel, creative* sentences.
Surprise in Puns

Yesterday I accidentally swallowed some food coloring. The doctor says I’m OK, but I feel like I’ve dyed a little inside.  

Pun word: dyed. Alternative word: died

• In the local context:
  — died a little inside. 😄 dyed a little inside. 😕 😐

• In the global context: swallowed some food coloring
  — dyed a little inside. 😄 died a little inside. 😄
The Surprisal Principle

• Two ways to realize the surprisal principle for humor in pun.
  – Quantitative instantiate of the surprisal principle to measure funniness in pun.
  – Instantiate the principle in pun generation
    • Retrieve and edit
Prior Theories - Funniness in Puns

• Kao et. al. 2015 proposes two principles to quantify funniness in puns.
  – **Ambiguity**: The sentence has two meanings (necessary but insufficient condition)
    • the person died/dyed.
  – **Distinctiveness**: The two sentence meanings are supported by distinct subsets of words in the sentence.

• The incongruity theory.
Quantifying Local-Global Surprisal

• We quantify surprise using a pre-trained language model.  
  
  - Local surprisal $S_{local}$
    
    • $\log \frac{p(\text{like I've dyed a little inside})}{p(\text{like I've died a little inside})}$
  
  - Global surprisal $S_{global}$
    
    • $\log \frac{p(\text{Yesterday I ... like I've dyed a little inside.})}{p(\text{Yesterday I ... like I've died a little inside.})}$
  
  - Local-global surprisal ratio: $S_{ratio} \overset{\text{def}}{=} \frac{S_{local}}{S_{global}}$ (the larger the better)

Inspired by Levy, 2015

Humor theory about incongruity resolution, Tony, 2004
Evaluating The Surprisal Principle

• Goal: compute the correlation between human ratings of funniness and the scores produced by our principle.

• Three types of sentences:
  – **Pun**: The magician got so mad he pulled his hare out.
  – **Swap pun**: The magician got so mad he pulled his hair out.
  – **Non-pun**: Look at that hare.

• Datasets
  – Derived from SemEval 2017 (130 sentences including the three cases)
  – Human rating of funniness from 1 (not at all) to 7 (extremely)
No Single Metric Works Well Across the Board

Directly optimizing the surprisal score cannot work well for generation.
A Retrieve-and-Edit Framework for Pun Generation

- Generating local-surprisal: retrieve and swap

- Generating global-local contrast: inserting a topic word at the beginning.
  - Relativeness measured by a distance skip-gram.

\[
\begin{align*}
  w^p &= \text{hare} \\
  w^a &= \text{hair}
\end{align*}
\]

Retrieve using hair

the man stopped to get a hair cut.

Swap hair → hare

the man stopped to get a hare cut.

Insert topic man → greyhound

the greyhound stopped to get a hare cut.
Evaluating the Generation System

• Selected 150 pun-alternative pairs.
• Each system generates puns from these words.
• Human ratings:
  • funniness (1-5),
  • grammaticality (1-5)
  • success (yes/no), given the formal definition in D. Aarons. 2017
Evaluating Generated Puns

Success Rate

- Neural Joint Decoder: 9.20%
- Retrieve: 4.60%
- Our best system SURGEN: 31.40%
- Human: 78.90%

Yu et al. ACL 2018
Evaluating Generated Puns

He et al. (NAACL 2019)

Funniness and Grammaticality

- Funniness
  - Neural Joint Decoder: 1.4
  - Retrieve: 1.3
  - Our best system SURGEN: 1.7
  - Human: 3

- Grammar
  - Neural Joint Decoder: 2.6
  - Retrieve: 3.9
  - Our best system SURGEN: 3
  - Human: 3.8
<table>
<thead>
<tr>
<th>Method</th>
<th>Example</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pun/alternative word pair: butter – better</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NJD</td>
<td>He is going to come up with the butter a ‘very good’ approach to the world’s economic crisis, the world’s biggest economic climate.</td>
<td>1</td>
</tr>
<tr>
<td>SURGEN</td>
<td>Well, gourmet did it, he thought, it’d butter be right.</td>
<td>2</td>
</tr>
<tr>
<td>Human</td>
<td>Why did the dairy churn? The less said, the butter...</td>
<td>1.5</td>
</tr>
<tr>
<td><strong>Pun/alternative word pair: peace – piece</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NJD</td>
<td>Further, he said, at the end of the peace, it’s not clear that it will be a good example.</td>
<td>1</td>
</tr>
<tr>
<td>SURGEN</td>
<td>That’s because negotiator got my car back to me in one peace.</td>
<td>1.5</td>
</tr>
<tr>
<td>Human</td>
<td>Life is a puzzle; look here for the missing peace.</td>
<td>3</td>
</tr>
</tbody>
</table>
Outline

• Creative pun generation
  – Pun generation with surprise (He et al. NAACL 2019)

• Evaluating open-domain dialog systems
  – Contextualized relevance (Ghazarian et al. NAACL workshop 2019a)
  – Predictive engagement (Ghazarian et al. AAAI 2020)
MT metric validation studies

(source text, reference text, system output, human quality score, ... + BLEU score, new metric score)

i.e.

(“我吃了榛子”, “I ate hazelnuts”, “I ate”, 3/10, 0.36, 0.2777)

To make contributions proposing a new metric:

$\text{corr}(\text{human scores, new metric}) > \text{corr}(\text{human scores, BLEU score})$
Related Work on MT Evaluation Metrics

Ma et al., WMT 2019

Metric performance (correlation)

# of top-N systems included in analysis
The Meta Challenge of Open-domain Generation Evaluation

Reference-based methods do not work for open-domain generation evaluation.

<table>
<thead>
<tr>
<th>Query</th>
<th>Dialogue Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker A: Why don’t we go see a movie?</td>
<td>Speaker 1: Hey! What are you doing here?</td>
</tr>
<tr>
<td>Ground-truth Reply</td>
<td>Speaker 2: I’m just shopping.</td>
</tr>
<tr>
<td>Speaker B: I don’t like watching movies at all.</td>
<td>Query: What are you shopping for?</td>
</tr>
<tr>
<td>Generated Reply</td>
<td>Generated Response: Some new clothes.</td>
</tr>
<tr>
<td>Speaker B: Heard the film about Turing is out!</td>
<td>Reference Response: I want buy gift for my mom!</td>
</tr>
</tbody>
</table>

(“What’s your plan tonight”, [“Watch TV”], “Going for a movie”, 3/10, 0.36, 0.2777)
Previous Work on Dialog Evaluation

- RUBER: a Referenced metric and Unreferenced metric Blended Evaluation Routine

- The unreferenced score:
  - A trainable metric measure relevance.

Tao et al. (AAAI 2018)
Contextualized Relevance Evaluation

Contextualized RUBER

Ghazarian et al. (NAACL 2019 NeuralGen)
Contextualized Relevance Score Improves Non-contextualized One

Evaluation on 300 human-labeled query-response pairs.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>RUBER</td>
<td>0.28</td>
<td>0.30</td>
</tr>
<tr>
<td>RUBER_Relevance_Only</td>
<td>0.22</td>
<td>0.23</td>
</tr>
<tr>
<td>Ctx_RUBER (Ours)</td>
<td>0.49</td>
<td>0.44</td>
</tr>
<tr>
<td>Ctx_Relevance (Ours)</td>
<td>0.55</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Contextualized word embeddings help the relevance score. Reference score does not help contextualized relevance.
The Importance of Engagement Metric

*Engagement* is defined as a user’s inclination to continue interacting with a dialogue system.

Only relevancy can not enough!

⭐ We Investigate the efficiency of engagement score for dialog evaluation
Engagement Measurement

• We investigate the efficacy of estimating *utterance-level* engagement and define a novel metric, *predictive engagement*.
• **Benefit**: real-time feedback for training dialogue models, more effective in providing signals to adjust the system as it proceeds.
Utterance-level Engagement Scores

Ghazarian et al. (AAAI 2020)

- In chatbot competitions like NeurIPS ConvAI and Amazon Alexa prize, users are asked to evaluate conversations based on how engaging they are.

- **Question 1:** whether the engagement score can be measured at the utterance level?
  - **Answer:** Yes, users agreement on utterance-level engagement score

- AMT experiments on randomly selected 50 conversations from ConvAI

<table>
<thead>
<tr>
<th>Utterances</th>
<th>Annotators</th>
<th>Kappa Agreement</th>
<th>Pearson</th>
</tr>
</thead>
<tbody>
<tr>
<td>297</td>
<td>49</td>
<td>0.52</td>
<td>0.93</td>
</tr>
</tbody>
</table>
**Utterance and Conversation level Engagement Scores**

**Question 2:** Is the utterance-level engagement score for each utterance in a conversation predictive for the conversation-level engagement score?
Utterance and Conversation level Engagement Scores

Question 2: Is the utterance-level engagement score for each utterance in a conversation predictive for the conversation-level engagement score?

<table>
<thead>
<tr>
<th>Aggregation Method</th>
<th>Pearson Correlation (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.49 (&lt;3e-4)</td>
</tr>
<tr>
<td>Max</td>
<td>0.72 (4e-9)</td>
</tr>
<tr>
<td>Mean</td>
<td>0.85 (&lt;9e-15)</td>
</tr>
</tbody>
</table>

Utterance 1
Utterance 2
...
Utterance n

S1
S2
Sn

g
S'

AMT

ConvAI dataset

Information Sciences Institute
**Question 3:** Can we simply assign the conversation-level engagement score to each utterances in the conversation to get noisy training data for utterance-level engagement?

![Utterance and Conversation level Engagement Scores](image-url)
**Utterance and Conversation level Engagement Scores**

**Question 3:** Can we simply assign the conversation engagement score to each utterance in the training data for utterance-level engagement?
Engagement Classifier

Baselines:
- SVM classifier
- MLP classifier with words static embeddings
Datasets

- **ConvAI**: to train the utterance-level engagement model

<table>
<thead>
<tr>
<th>Engagement Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>Conversations</td>
</tr>
<tr>
<td>1690</td>
</tr>
<tr>
<td>Utterances</td>
</tr>
<tr>
<td>10122</td>
</tr>
</tbody>
</table>

- **Daily Dialogue Dataset**: train dialogue systems and fine-tune evaluation metrics for measuring relevance score.
  - 300 utterances with generated replies (attention-based seq-to-seq)
  - 300 utterances with human written replies
Experimental Setup

• Imbalanced dataset → weighted loss function, AUC for evaluation
• Eval on automatically annotated ConvAI dataset and Human annotated ConvAI dataset.
• Baselines: SVM, MLP
Results on Predicting Engagement Score

- Results on utterance level engagement classification.

![Graph showing accuracy of different models for predicting engagement scores.](chart.png)
We recruited about 45 participants from AMT to annotate 300 pairs from Daily Dialog dataset as engaging or not engaging.

Mean Kappa agreement = 0.51
Transfer Learning Results

Correlation between predictive engagement score and human evaluation of dialog quality.
## Results on Dialog Evaluation

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>300 Generated Responses</strong></td>
<td>RUBER_relavance</td>
<td>0.28</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Ctx_relevance</td>
<td>0.55</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>MLP BERT(mean)_engagement</td>
<td>0.14</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>MLP BERT(mean) + Ctx_relevance</td>
<td>0.51</td>
<td>0.47</td>
</tr>
<tr>
<td><strong>300 Human-written Responses</strong></td>
<td>RUBER_relavance</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Cont_relevance</td>
<td>0.14</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>MLP BERT(mean)_engagement</td>
<td><strong>0.43</strong></td>
<td><strong>0.42</strong></td>
</tr>
<tr>
<td></td>
<td>MLP BERT(mean) + Ctx_relevance</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td><strong>600 Generated and Human-written Responses</strong></td>
<td>RUBER_relavance</td>
<td>0.24</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Ctx_relevance</td>
<td>0.54</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>MLP BERT(mean)_engagement</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>MLP BERT(mean) + Ctx_relevance</td>
<td><strong>0.61</strong></td>
<td><strong>0.62</strong></td>
</tr>
</tbody>
</table>
### Case Studies

*Ghazarian et al. (AAAI 2020)*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>So how did I do on my driving test?</td>
<td>Do you want the truth?</td>
<td>0.42</td>
<td>0.60</td>
<td>0.31</td>
<td>0.46</td>
</tr>
<tr>
<td>How long are you going to stay here?</td>
<td>Only one night.</td>
<td>0.43</td>
<td>0.99</td>
<td>0.53</td>
<td>0.76</td>
</tr>
<tr>
<td>Well, there are a lot to do, but nothing to worry about. First, make sure they don't go far from the railings.</td>
<td>Oh, yeah, maybe I should ask them just to walk along the railings.</td>
<td>0.58</td>
<td>0.97</td>
<td>0.17</td>
<td>0.57</td>
</tr>
</tbody>
</table>
Conclusion

• Open domain generation and evaluation are challenging
  – The content is under-specified.
  – Conditional entropy is high.
  – Reference-based method evaluation metrics would not work.

• We leverage the surprisal principle for creative pun generation.

• We propose contextualized relevance and predictive engagement metrics for open-domain dialog evaluation.

• Code and Data can be found at: https://violetpeng.github.io/
Thanks!

Questions?

Special thanks to my students and collaborators:
Sarik Ghazarian, Johnny Tian-Zheng Wei, Aram Galstyan, Ralph Weischedel, He He, Percy Liang