Deep Learning in Open-Domain Dialogue Systems

Yanju Chen
Google Duplex: A Task-Driven Bot

“What? Where is John? I cannot believe this. This is bullshit.”

https://www.youtube.com/watch?v=qB9sYGZJdbs
When a bot meets ... another bot ...

JaviAir
@Javi

Google Duplex: "Hi! Uhm... I'd like to make a dinner reservation for 3."
Restaurant: "Sure! What time would you like?"
Google Duplex: "it's, uhh... for tomorrow May the 11th at NULL POINTER EXCEPTION."
Restaurant: "Internal Exception: Invalid parameter not satisfying: time".

10:03 AM - May 10, 2018 · Berkeley, CA

10.9K 4,742 people are talking about this
Incorporated Dialogue Systems

Why chit-chat bot? What’s the meaning of it?

- entertainment value
- new interaction design
- scientific usage
- decrease human workload
- important part of task bot
ELIZA (1966): Psychological Therapist

Weizenbaum, “ELIZA – A computer program for the study of natural language communication between man and machine”, Communications of ACM 1966

http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm
Two Branches of Dialogue Systems

- **Task-Oriented Bot**
  - Personal assistant, helps users achieve a certain task
  - Combination of rules and statistical components
    - POMDP for spoken dialog systems (Williams and Young, 2007)
    - End-to-end trainable task-oriented dialogue system (Wen et al., 2016)
    - End-to-end reinforcement learning dialogue system (Li et al., 2017; Zhao and Eskenazi, 2016)

- **Chit-Chat Bot**
  - No specific goal, focus on natural responses
  - Using variants of seq2seq model
    - A neural conversation model (Vinyals and Le, 2015)
    - Reinforcement learning for dialogue generation (Li et al., 2016)
    - Conversational contextual cues for response ranking (Al-Rfou et al., 2016)

Outline

- PART I. Open-Domain Dialogue Systems
- PART II. Open-Domain Dialogue Evaluations
Open-Domain Dialogue Systems
Problem Formalization (Generative Model)

- Single-Turn Dialogue
  - (message, response) – \((m, r)\)

- Multi-Turn Dialogue
  - (context, message, response) – \((c, m, r)\)

- Goal of Generative Dialogue Model
  - to generate entirely new sentences that are unseen in the training set

Sequence-To-Sequence Model (seq2seq)

- Vanilla Sequence-To-Sequence Model

single-turn situation \( f(\text{context, message}) \)

multi-turn situation

• How about multi-turn situation? (context, message, response)
  • wrap (context, message) into a function and transform to a new sequence?

Sutskever et al., “Sequence to Sequence Learning with Neural Networks”, NIPS 2014
The Blandness Problem (Response Diversity)

- Not active or engaging at all!

- Maybe we should pay attention to:
  - how to capture dialogue **topics**
  - how to make it **human-like**
How to Capture Dialogue Topics

• In fact, there are early works on dialogue topic capturing using deep learning, even before SEQ2SEQ.

Context-Sensitive Generation

• Motivation: explicitly consider and model dialogue context

• Methods: extends the Recurrent Language Model (RLM)
  
  given sentence \( s = s_1, \ldots, s_T \), to estimate:

  \[
  p(s) = \prod_{t=1}^{T} p(s_t|s_1, \ldots, s_{t-1})
  \]

  \[
  p(r|c, m) = \prod_{t=1}^{T} p(r_t|r_1, \ldots, r_{t-1}, c, m)
  \]

  probability of a natural language sentence \( s \)

  What?? No sequence generated? How does this work?

• Dialogue Generation before SEQ2SEQ
  
  • complex systems generate candidate responses
  • use features to re-rank candidate responses
  • RLM provides a feature for a candidate response

---

Mikolov et al., “Recurrent Neural Network Based Language Model”, INTERSPEECH 2010
Context-Sensitive Models: RLMT

- **Tripled** Language Model (RLMT) - Baseline
  - **concatenate the triple** (context, message, response): \( s = [c; m; r] \)

\[ (RLM) \ s = \text{ok good luck !} \]
\[ (RLMT) \ s = \text{because of your game ? yeah i 'm on my way now . ok good luck !} \]

Figure 1: Example of three consecutive utterances occurring between two Twitter users A and B.

Figure 2: Compact representation of an RLM (left) and unrolled representation for two time steps (right).

Context-Sensitive Models: DCGM-I

- Dynamic-Context Generative Models I (DCGM-I)
  - model **word occurrences** in context
  - \( b_{cm} \in \mathbb{R}^V \): bag-of-words representation
  - \( k_L \): context-message encoding
  - adding context vector as additional bias to RLM:

\[
  h_t = \sigma(h_{t-1}^T W_{hh} + k_L + s_t^T W_{in})
\]

**additional bias**

- do not distinguish between \( c \) and \( m \)
- \( m \) and \( r \) have stronger dependency

---


---
Context-Sensitive Models: DCGM-II

- Dynamic-Context Generative Models II (DCGM-II)
  - model **word occurrences**, **distinguish** context and message
  - $b_c, b_m \in \mathbb{R}^V$: bag-of-words representation
  - $k_L$: context-message encoding
  - adding context vector as additional bias to RLM

$$h_t = \sigma(h_{t-1}^\top W_{hh} + k_L + s_t^\top W_{in})$$

Dataset & Evaluation Settings

- **Dataset**: selected 4,232 Twitter \((c, m, r)\) triplets, 2,118/2,114 for train/test
- **Automatic Evaluations**
  - BLEU
  - METEOR
- **Multi-Reference Extraction**
  - **Why?** The set of reasonable responses is vast and diverse.
  - **How?** Use Information Retrieval method to select more candidate response.
  - Retain high-quality candidates by human evaluation.

### Experiment Results (Automatic Evaluation)

<table>
<thead>
<tr>
<th>MT (n)-best</th>
<th>BLEU (%)</th>
<th>METEOR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT 9 \text{ feat.}</td>
<td>3.60 (-9.5%)</td>
<td>9.19 (-0.9%)</td>
</tr>
<tr>
<td>CMM 9 \text{ feat.}</td>
<td>3.33 (-16%)</td>
<td>9.34 (+0.7%)</td>
</tr>
<tr>
<td>▷ MT + CMM 17 \text{ feat.}</td>
<td>3.98 (-)</td>
<td>9.28 (-)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IR (n)-best</th>
<th>BLEU (%)</th>
<th>METEOR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR 2 \text{ feat.}</td>
<td>1.51 (-55%)</td>
<td>6.25 (-22%)</td>
</tr>
<tr>
<td>CMM 9 \text{ feat.}</td>
<td>3.39 (-0.6%)</td>
<td>8.20 (+0.6%)</td>
</tr>
<tr>
<td>▷ IR + CMM 10 \text{ feat.}</td>
<td>3.41 (-)</td>
<td>8.04 (-)</td>
</tr>
</tbody>
</table>

Table 3: Context-sensitive ranking results on both MT (left) and IR (right) \(n\)-best lists, \(n = 1000\). The subscript \text{ feat.} indicates the number of features of the models. The log-linear weights are estimated by running one iteration of MERT. We mark by \((\pm \%)\) the relative improvements with respect to the reference system (▷).
Experiment Results (Human Evaluation)

<table>
<thead>
<tr>
<th>System A</th>
<th>System B</th>
<th>Gain (%)</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUMAN</td>
<td>MT+CMM</td>
<td>13.6*</td>
<td>[12.4, 14.8]</td>
</tr>
<tr>
<td>DCGM-II</td>
<td>MT</td>
<td>1.9*</td>
<td>[0.8, 2.9]</td>
</tr>
<tr>
<td>DCGM-II+CMM</td>
<td>MT</td>
<td>3.1*</td>
<td>[2.0, 4.3]</td>
</tr>
<tr>
<td>DCGM-II+CMM</td>
<td>MT+CMM</td>
<td>1.5*</td>
<td>[0.5, 2.5]</td>
</tr>
<tr>
<td>DCGM-II</td>
<td>IR</td>
<td>5.2*</td>
<td>[4.0, 6.4]</td>
</tr>
<tr>
<td>DCGM-II+CMM</td>
<td>IR</td>
<td>5.3*</td>
<td>[4.1, 6.6]</td>
</tr>
<tr>
<td>DCGM-II+CMM</td>
<td>IR+CMM</td>
<td>2.3*</td>
<td>[1.2, 3.4]</td>
</tr>
</tbody>
</table>

Table 4: Pairwise human evaluation scores between System A and B. The first (second) set of results refer to the MT (IR) hypothesis list. The asterisk means agreement between human preference and BLEU rankings.
Sample Responses from the System

**context** no one can help this time its just i need a support system
**message** well i m here to support you. whenever you need me.
**response** thank you, i appreciate it thank you

**context** nothin much, and how's the book?!
**message** its good but i’m only like halfway through cuz i don’t feel like reading. i’m so bored ...
**response** that’s good! i have the book but i’m bored too.

**context** thanks, your way better though
**message** i’m just good at expressing it that’s but trust me you are freakin amazing and waay better
**response** i know, i don’t trust you, i love you too! thank you

**context** today i will walk for miles. clear sky clear mind.
**message** love walking. it’s too hot to walk this month. i haven’t moved. frozen. heat freeze.
**response** it’s too hot for you to get up to honey if you have frozen yogurt to you.

Though reasonably plausible in the context, responses are still generic or commonplace.
How to Make Dialogue Human-Like

• Several Factors:
  • Ease of Answering?
  • Information Flow? (contribute new information)
  • Semantic Coherence?
  • ...

• We now use SEQ2SEQ to generate the dialogue responses.

• What if there’s a human-like model to help discriminate all the dialogues?

Generative Adversarial Nets (GAN)
Design a GAN for dialogue generation

• Our Motivation:
  • Produce sequences that are indistinguishable from human-generated dialogue utterances.

• Problem Formalization
  • given dialogue history $x$: a sequence of dialogue utterances
  • to generate response $y = \{y_1, y_2, \ldots, y_T\}$

• What we have:
  • Generator: SEQ2SEQ
  • Discriminator: a binary classifier $Q_+({x, y})$

Li et al., “Adversarial Learning for Neural Dialogue Generation”, EMNLP 2017
Training a generator: maximize the likelihood

\[ p(y_1, \ldots, y_T|x_1, \ldots, x_T) = \prod_{t=1}^{T} p(y_t|y_1, \ldots, y_{t-1}) \]

**objective function**

(x) how are you ?  
(y) I'm fine . EOS

Li et al., “Adversarial Learning for Neural Dialogue Generation”, EMNLP 2017
Training a generator: maximize the rewards

• How to use discriminator signal $Q_+(\{x, y\})$

$Q_+\left(\{"how\ are\ you\?","I'm\ fine\ .\ EOS"\}\right)$

$J(\theta) = \mathbb{E}_{y \sim p(y|x)}(Q_+\left(\{x, y\}\right) | \theta)$

Adversarial REINFORCE Algorithm

only one reward for one sequence
assign the same reward for all tokens

Li et al., “Adversarial Learning for Neural Dialogue Generation”, EMNLP 2017
Reward for Every Generation Step (REGS)

- The Idea (Monte Carlo Search)
  
  estimate the quality of current token by rolling out complete sentence $N$ times

$$J(\theta) = \sum_t \mathbb{E}_{y_t \sim p(y_t|x,Y_{1:t-1})} Q+(x,Y_t) | \theta$$

REGS Monte Carlo

---

Li et al., “Adversarial Learning for Neural Dialogue Generation”, EMNLP 2017
Designing a discriminator

- What we want the discriminator to be:
  - binary classifier: $Q_+({x, y})$
  - context aware: consider history $x$ and response $y$

Li et al., “Adversarial Learning for Neural Dialogue Generation”, EMNLP 2017
Serban et al., “Building end-to-end dialogue systems using generative hierarchical neural network models”, AAAI 2016
Li et al., “A hierarchical neural autoencoder for paragraphs and documents”, ACL 2015
The Adversarial Training Algorithm

Add Teacher Forcing phase to stabilize the training.

• Dataset: OpenSubtitles (22.27M sentence fragments)
• Pre-Train: Generator and Discriminator
• Negative Samples for Discriminator
  • beam search with mutual information reranking (Li et al., 2016)
  • sampling
• Other Tricks
**Sample Responses & Human Evaluations**

(souce) tell me ... how long have you had this falling sickness?

(MLE) i 'm not a doctor.

(Sample) well everything you did was totally untrue.

(REINFORCE) i don 't know how long it 's been.

(REGS Monte Carlo) A few months, I guess.

(source) so i had the doctors test sammy 's response to conditioning.

(MLE) sammy wrote the test sammy wrote the test.

(Sample) objects pick objects up objects objects objects objects objects objects

(REINFORCE) i 'm not sure that 's relevant.

(REGS Monte Carlo) so he took the pills.

(source) the man who does that is not afraid to die.

(MLE) he 's not afraid of the dark.

(Sample) even the program is the only truth

(REINFORCE) he 's also not afraid of me.

(REGS Monte Carlo) he 's such a brave man!

<table>
<thead>
<tr>
<th>Setting</th>
<th>adver-win</th>
<th>adver-lose</th>
<th>tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>single-turn</td>
<td>0.62</td>
<td>0.18</td>
<td>0.20</td>
</tr>
<tr>
<td>multi-turn</td>
<td>0.72</td>
<td>0.10</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Table 4: The gain from the proposed adversarial model over the mutual information system based on pairwise human judgments.
Open-Domain Dialogue Evaluations
The Problems of Dialogue Evaluations

• NLP tasks have their own **automatic** evaluation metrics:
  • **Machine Translation**: BLEU, METEOR
  • **Summarization**: ROUGE
  • **Open-Domain Dialogue Generation**: ???

• Challenges in dialogue evaluation:
  • diversity of valid responses
  • Then how to evaluate a dialogue?
  • ... except for human evaluation

  **Context of Conversation**
  Speaker A: Hey John, what do you want to do tonight?
  Speaker B: Why don’t we go see a movie?

  **Potential Responses**
  Response 1: Nah, I hate that stuff, let’s do something active.
  Response 2: Oh sure! Heard the film about Turing is out!

  **responses do not share any words**

Word Overlap-Based Metrics

- **BLEU** (Papineni et al., 2002)
  
  \[ \text{BLEU} = \frac{1}{m} \sum_{i=1}^{m} \left( \log \left( \frac{\text{match}(X,Y)}{\text{top-1}(X)} \right) \right) \]

- **METEOR** (Banerjee and Lavie, 2005)

- **ROUGE-L** (Lin, 2004): Longest Common Subsequence
  
  - n-gram f-measure

  \[ R_{lcs} = \frac{\text{LCS}(X,Y)}{m} \quad P_{lcs} = \frac{\text{LCS}(X,Y)}{n} \quad F_{lcs} = \frac{(1 + \beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2P_{lcs}} \]

Embedding-Based Metrics

- **Greedy Matching** (word-level cosine similarity)

- **Embedding Average** (sentence-level cosine similarity)

- **Vector Extrema** (Forgues et al., 2014) (sentence-level)

Evaluations on Dialogue Models

Table 4: Correlation between BLEU metric and human judgements after removing stopwords and punctuation for the Twitter dataset.

<table>
<thead>
<tr>
<th></th>
<th>Spearman</th>
<th>p-value</th>
<th>Pearson</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU-1</td>
<td>0.1580</td>
<td>0.12</td>
<td>0.2074</td>
<td>0.038</td>
</tr>
<tr>
<td>BLEU-2</td>
<td>0.2030</td>
<td>0.043</td>
<td>0.1300</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Table 5: Effect of differences in response length for the Twitter dataset, $\Delta w = \text{absolute difference in \#words between a ground truth response and proposed response}$

<table>
<thead>
<tr>
<th></th>
<th>Mean score</th>
<th></th>
<th></th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Delta w \leq 6$ (n=47)</td>
<td>$\Delta w &gt; 6$ (n=53)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLEU-1</td>
<td>0.1724</td>
<td>0.1009</td>
<td>&lt; 0.01</td>
<td></td>
</tr>
<tr>
<td>BLEU-2</td>
<td>0.0744</td>
<td>0.04176</td>
<td>&lt; 0.01</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.6587</td>
<td>0.6246</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>METEOR</td>
<td>0.2386</td>
<td>0.2073</td>
<td>&lt; 0.01</td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>2.66</td>
<td>2.57</td>
<td>0.73</td>
<td></td>
</tr>
</tbody>
</table>
Evaluations on Dialogue Models

<table>
<thead>
<tr>
<th>Metric</th>
<th>Twitter Spearman</th>
<th>Twitter p-value</th>
<th>Twitter Pearson</th>
<th>Twitter p-value</th>
<th>Ubuntu Spearman</th>
<th>Ubuntu p-value</th>
<th>Ubuntu Pearson</th>
<th>Ubuntu p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greedy</td>
<td>0.2119</td>
<td>0.034</td>
<td>0.1994</td>
<td>0.047</td>
<td>0.05276</td>
<td>0.6</td>
<td>0.02049</td>
<td>0.84</td>
</tr>
<tr>
<td>Average</td>
<td>0.2259</td>
<td>0.024</td>
<td>0.1971</td>
<td>0.049</td>
<td>-0.1387</td>
<td>0.17</td>
<td>-0.1631</td>
<td>0.10</td>
</tr>
<tr>
<td>Extrema</td>
<td>0.2103</td>
<td>0.036</td>
<td>0.1842</td>
<td>0.067</td>
<td>0.09243</td>
<td>0.36</td>
<td>-0.002903</td>
<td>0.98</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.1887</td>
<td>0.06</td>
<td>0.1927</td>
<td>0.055</td>
<td>0.06314</td>
<td>0.53</td>
<td>0.1419</td>
<td>0.16</td>
</tr>
<tr>
<td>BLEU-1</td>
<td>0.1665</td>
<td>0.098</td>
<td>0.1288</td>
<td>0.2</td>
<td>-0.02552</td>
<td>0.8</td>
<td>0.01929</td>
<td>0.85</td>
</tr>
<tr>
<td>BLEU-2</td>
<td>0.3576</td>
<td>&lt; 0.01</td>
<td>0.3874</td>
<td>&lt; 0.01</td>
<td>0.03819</td>
<td>0.71</td>
<td>0.0586</td>
<td>0.56</td>
</tr>
<tr>
<td>BLEU-3</td>
<td>0.3423</td>
<td>&lt; 0.01</td>
<td>0.1443</td>
<td>0.15</td>
<td>0.0878</td>
<td>0.38</td>
<td>0.1116</td>
<td>0.27</td>
</tr>
<tr>
<td>BLEU-4</td>
<td>0.3417</td>
<td>&lt; 0.01</td>
<td>0.1392</td>
<td>0.17</td>
<td>0.1218</td>
<td>0.23</td>
<td>0.1132</td>
<td>0.26</td>
</tr>
<tr>
<td>ROUGE</td>
<td>0.1235</td>
<td>0.22</td>
<td>0.09714</td>
<td>0.34</td>
<td>0.05405</td>
<td>0.5933</td>
<td>0.06401</td>
<td>0.53</td>
</tr>
<tr>
<td>Human</td>
<td>0.9476</td>
<td>&lt; 0.01</td>
<td>1.0</td>
<td>0.0</td>
<td>0.9550</td>
<td>&lt; 0.01</td>
<td>1.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 3: Correlation between each metric and human judgements for each response. Correlations shown in the human row result from randomly dividing human judges into two groups.

What's wrong with the evaluation metrics? (on Open-Domain Dialogue)

- Automatic Metrics:
  - correlate very weakly with human judgement
  - incapable of considering the semantic similarity between responses

- Human Evaluations:
  - too expensive
  - time-consuming for every model specification

Lowe et al., "Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses", ACL 2017
Learning to Evaluate Dialogue Responses

- Motivations:
  - train an automatic dialogue evaluation model (ADEM) to predict human scores and can:
    - capture semantic similarity beyond word overlap statistics
    - exploit both the context and reference responses

---

<table>
<thead>
<tr>
<th>Context of Conversation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker A: Hey, what do you want to do tonight?</td>
</tr>
<tr>
<td>Speaker B: Why don’t we go see a movie?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nah, let’s do something active.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeah, the film about Turing looks great!</td>
</tr>
</tbody>
</table>

Figure 1: Example where word-overlap scores fail for dialogue evaluation; although the model response is reasonable, it has no words in common with the reference response, and thus would be given low scores by metrics such as BLEU.
Automatic Dialogue Evaluation Model (ADEM)

- $M, N \in \mathbb{R}^n$: linear projection (without activation)

Figure 2: The ADEM model, which uses a hierarchical encoder to produce the context embedding $c$.

\[
\text{score}(c, r, \hat{r}) = \frac{(c^T M \hat{r} + r^T N \hat{r} - \alpha)}{\beta}
\]

\[
L = \sum_{i=1:K} \left[ \text{score}(c_i, r_i, \hat{r}_i) - \text{human}_i \right]^2 + \gamma \|\theta\|_2
\]

Lowe et al., "Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses", ACL 2017
### Utterance-Level Correlations

<table>
<thead>
<tr>
<th>Metric</th>
<th>Full dataset</th>
<th></th>
<th>Test set</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spearman</td>
<td>Pearson</td>
<td>Spearman</td>
<td>Pearson</td>
</tr>
<tr>
<td>BLEU-2</td>
<td>0.039 (0.013)</td>
<td>0.081 (&lt;0.001)</td>
<td>0.051 (0.254)</td>
<td>0.120 (&lt;0.001)</td>
</tr>
<tr>
<td>BLEU-4</td>
<td>0.051 (0.001)</td>
<td>0.025 (0.113)</td>
<td>0.063 (0.156)</td>
<td>0.073 (0.103)</td>
</tr>
<tr>
<td>ROUGE</td>
<td>0.062 (&lt;0.001)</td>
<td>0.114 (&lt;0.001)</td>
<td>0.096 (0.031)</td>
<td>0.147 (&lt;0.001)</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.021 (0.189)</td>
<td>0.022 (0.165)</td>
<td>0.013 (0.745)</td>
<td>0.021 (0.601)</td>
</tr>
<tr>
<td>T2V</td>
<td>0.140 (&lt;0.001)</td>
<td>0.141 (&lt;0.001)</td>
<td>0.140 (&lt;0.001)</td>
<td>0.141 (&lt;0.001)</td>
</tr>
<tr>
<td>VHRED</td>
<td>-0.035 (0.062)</td>
<td>-0.030 (0.106)</td>
<td>-0.091 (0.023)</td>
<td>-0.010 (0.805)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Validation set</th>
<th>Test set</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spearman</td>
<td>Pearson</td>
<td>Spearman</td>
<td>Pearson</td>
</tr>
<tr>
<td>C-ADEM</td>
<td>0.338 (&lt;0.001)</td>
<td>0.355 (&lt;0.001)</td>
<td>0.366 (&lt;0.001)</td>
<td>0.363 (&lt;0.001)</td>
</tr>
<tr>
<td>R-ADEM</td>
<td>0.404 (&lt;0.001)</td>
<td>0.404 (&lt;0.001)</td>
<td>0.352 (&lt;0.001)</td>
<td>0.360 (&lt;0.001)</td>
</tr>
<tr>
<td>ADEM (T2V)</td>
<td>0.252 (&lt;0.001)</td>
<td>0.265 (&lt;0.001)</td>
<td>0.280 (&lt;0.001)</td>
<td>0.287 (&lt;0.001)</td>
</tr>
<tr>
<td>ADEM</td>
<td><strong>0.410 (&lt;0.001)</strong></td>
<td><strong>0.418 (&lt;0.001)</strong></td>
<td><strong>0.428 (&lt;0.001)</strong></td>
<td><strong>0.436 (&lt;0.001)</strong></td>
</tr>
</tbody>
</table>

Lowe et al., "Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses", ACL 2017
Utterance-Level Correlations

Figure 4: Scatter plot showing model against human scores, for BLEU-2 and ROUGE on the full dataset, and ADEM on the test set. We add Gaussian noise drawn from $\mathcal{N}(0, 0.3)$ to the integer human scores to better visualize the density of points, at the expense of appearing less correlated.
Summary
Summary

• Dialogue systems remain a challenging topic
  • diversity problem
  • context aware generation
  • higher level human-like generation

• Open-domain dialogue evaluation remains an open problem
  • extended reference
  • adversarial evaluation