The Problem(s) with Neural Chatbots

Ryan Lowe
McGill University, OpenAI
Dialogue systems
Why work on dialogue systems?

• Many commercial applications

• Creating a ‘general-purpose communicating agent’
  • An agent that can communicate with humans on many topics, to exchange knowledge and complete a variety of tasks in its environment.

• Language is a natural communication interface between humans and machines
Modular dialogue systems

• Traditional system consists of modules

• Each module optimized with separate objective function

• Achieves fairly good performance with small amounts of data

Problem: require supervised data for each module – does not scale well to general domains!
End-to-end dialogue systems

• A single model trained directly on conversational data

• Uses a single objective function, usually maximum likelihood on next response

• Significant recent work using neural networks to predict the next response. (Ritter et al., 2011; Sordoni et al., 2015; Shang et al., 2015)
Why might this work?

Advantages of end-to-end systems:

1) Does not require feature engineering (only architecture engineering).

2) Can be transferred to different domains.

3) Does not require supervised data for each module! (collecting this data is difficult at large scales)
Early results – Vinyals & Le (2015)

Machine: hi
Human: hello
Machine: could you please let me know what are the operating systems you are using?
Human: i am using linux
Machine: ok it was in the ssl vpn on which you stated it was broken
Machine: please check if you are logged in to corp network
Machine: what is the error that you are running please
Human: i am seeing an error related to vpn
Machine: what is the error message that you are getting when connecting to vpn using network connect?
Human: connection refused or something like that
Machine: may i know the version of network connect you connect ?
Human: i am not sure i know that
Machine: is the network connect prompting for any user certificate
Machine: ?
Human: yes
Machine: are you at home or at work at the moment?
Machine: office ?
Human: i am at home
Machine: try this
Machine: goto <URL> page and enter your name and password
Human: done
Machine: try logging in with <NAME> and check if its working fine now
Human: yes, now it works !
Machine: great
<first_speaker> what do you mean?
<second_speaker> what are you talking about?

What went wrong?
Problem #1: Data
Dialogue datasets

• Building general-purpose dialogue systems requires lots of data

• The best datasets are proprietary

• We need large (>500k dialogues), open-source datasets to make progress
Ubuntu Dialogue Corpus

• Large dataset of ~1 million tech support dialogues
• Scraped from Ubuntu IRC channel
• 2-person dialogues extracted from chat stream

Ubuntu Dialogue Corpus

Pros:
• Hard
• Large
• Open-source
• Related to many real-world technical problems

Cons:
• Too hard?
• Not perfectly disentangled
• Requires external knowledge to solve
• Ideally suited for task-oriented setting, but no reward signal in dataset
Large-scale dialogue datasets

- Ubuntu Dialogue Corpus (Lowe et al., 2015)
- Twitter Corpus (Ritter et al., 2011)
- Movie Dialog Dataset (Dodge et al. 2016)
- Reddit
- ...

Survey paper covering existing datasets:
Problem #2: Model Architecture
Recurrent neural networks

- Augment neural networks with self-loops
- Leads to the formation of a *hidden state* $s_t$ that evolves over time:
  \[
  h_t = f(W_{hh}h_{t-1} + W_{ih}x_t)
  \]
- Used to model sequences (e.g. natural language)

Source: colah.github.io
Sequence-to-sequence learning

- Use an RNN **encoder** to map an input sequence to a fixed-length vector
- Use an RNN **decoder** (with different parameters) to map the vector to the target sequence (Cho et al., 2014; Sustkever et al., 2014)
Main goal

Build models with right inductive biases to effectively represent dialogue data

Judge model quality by quality of generated responses
Some problems: generic responses

• Most models trained to predict most likely next utterance given context

• But some utterances are likely given any context!

• Neural models often generate “I don’t know”, or “I’m not sure” to most contexts

(Li et al., 2016)
More problems

• Strong constraint on generation process: only source of variation is at the output

• When the model lacks capacity, it is encouraged to mostly capture short-term dependencies

• Want to explicitly model variations at ‘higher level’ representations (e.g. topic, tone, sentiment, etc.)
Variational encoder-decoder (VHRED)

- Augment HRED with Gaussian latent variable $z$
- $z$ can capture high-level utterance features (e.g. topic, tone)
- When generating first sample latent variable, then use it to condition generation

Variational encoder-decoder (VHRED)

- Inspired by VAE (Kingma & Welling, 2014; Rezende et al., 2014): train model with backprop using reparameterization trick

- Prior mean and variance are learned conditioned on previous utterance representation. Posterior mean and variance also conditioned on representation of target utterance.

- At training time, sample from posterior. At test time, sample from prior.

- Developed concurrently with Bowman et al. (2016)
  - Use word-dropping and KL annealing tricks
Quantitative results

Table 1: Wins, losses and ties (in %) of VHRED against baselines based on the human study (mean preferences ± 90% confidence intervals, where * indicates significant differences at 90% confidence)

<table>
<thead>
<tr>
<th>Opponent</th>
<th>Wins</th>
<th>Losses</th>
<th>Ties</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Short Contexts</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VHRED vs LSTM</td>
<td>32.3 ±2.4</td>
<td>42.5 ±2.6*</td>
<td>25.2 ±2.3</td>
</tr>
<tr>
<td>VHRED vs HRED</td>
<td>42.0 ±2.8*</td>
<td>31.9 ±2.6</td>
<td>26.2 ±2.5</td>
</tr>
<tr>
<td>VHRED vs TF-IDF</td>
<td>51.6 ±3.3*</td>
<td>17.9 ±2.5</td>
<td>30.4 ±3.0</td>
</tr>
<tr>
<td><strong>Long Contexts</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VHRED vs LSTM</td>
<td>41.9 ±2.2*</td>
<td>36.8 ±2.2</td>
<td>21.3 ±1.9</td>
</tr>
<tr>
<td>VHRED vs HRED</td>
<td>41.5 ±2.8*</td>
<td>29.4 ±2.6</td>
<td>29.1 ±2.6</td>
</tr>
<tr>
<td>VHRED vs TF-IDF</td>
<td>47.9 ±3.4*</td>
<td>11.7 ±2.2</td>
<td>40.3 ±3.4</td>
</tr>
</tbody>
</table>

Figure 2: Human evaluator preferences for VHRED vs LSTM by context length excluding ties. For short contexts humans prefer the generic responses generated by LSTM, while for long contexts humans prefer the semantically richer responses generated by VHRED.
### Table 2: Twitter examples for the neural network models. The → token indicates a change of turn.

<table>
<thead>
<tr>
<th>Context</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>jamaica blood!! &lt;url&gt; → lol. yes [name]! [name]! → what’s poppin in toronto homie</td>
<td>VHRED: nothing much. just chillin. what’s good with you? &lt;br&gt; LSTM: something!!!!!!! &lt;br&gt; HRED: something!!!!!!!</td>
</tr>
<tr>
<td>reaches for [name]! → - approaches him warily - hey there buddy. - swings him into the air - how you doin? → tilts his head and your wariness and reaches for your face [name].</td>
<td>VHRED: - looks up at you and smiles - you’re so cute. - kisses your cheek - &lt;br&gt; LSTM: - smiles - i’m fine thank you. how are you? &lt;br&gt; HRED: - grins. looking at you as i look at you -</td>
</tr>
<tr>
<td>my math teacher over here talkn bout nickels nd i’m tryna make big money i need to be in calculus → lol</td>
<td>VHRED: i need to go to the gym and get my hair done lol &lt;br&gt; LSTM: lol &lt;br&gt; HRED: how u been</td>
</tr>
<tr>
<td>i think i want an ipad2. but i already have an ibook. want vs. need. → if you have an ibook. then you deserve an ipad!</td>
<td>VHRED: that’s what i’m thinking. i’m going to buy a new one. &lt;br&gt; LSTM: i want one!!!! &lt;br&gt; HRED: i don’t know what to do with it. i don’t know what to do.</td>
</tr>
</tbody>
</table>
| ah echt? dan antwoord ik volgende keer via twitter als je me smst. → haha doe maar p. goed bezig hoor myrthe d → goed bezig? die snap ik even niet p | VHRED: haha, ik heb er geen zin in de something, maar ik heb er geen something meer something <br> LSTM: haha something something something <br> HRED: something something something something something something
Future work

• Many interesting areas to be investigated:
  • Modifying the loss function
  • Adversarial training
  • Reinforcement learning
  • Learning from human interaction
  • ...

Problem #3: Evaluation
• Hard to know if we’re making progress in building dialogue models

• Important to define – wrong metrics can lead to spurious research

• Human evaluation is effective, but slow and expensive – want to have an automatic evaluation metric

• Lack of reliable metrics means researchers only compare to their own previously implemented models
Comparison of ground-truth utterance

Context
Hey, want to go to the movies tonight?

Generated Response
Nah, let’s do something active.

Reference response
Yeah, the film about Turing looks great!
Comparison of ground-truth utterance

• Word-overlap metrics:
  • BLEU, METEOR, ROUGE

• Look at the number of overlapping n-grams between the generated and reference responses

• Correlate poorly with humans in dialogue

Generated Response
Yes, let’s go see that movie about Turing!

Reference response
Nah, I’d rather stay at home, thanks.

SCORE
Correlation study

- Created 100 questions each for Twitter and Ubuntu datasets (20 contexts with responses from 5 ‘diverse models’)
- 25 volunteers from CS department at McGill
- Asked to judge response quality on a scale from 1 to 5
- Compared human ratings with ratings from automatic evaluation metrics
Models for response variety

1) Randomly selected response

2) Retrieval models:
   • Response with smallest TF-IDF cosine distance
   • Response selected by Dual Encoder (DE) model

3) Generative models:
   • Hierarchical recurrent encoder-decoder (HRED)

4) Human-written response (not ground truth)
Goal (inter-annotator)

Figure 3: Scatter plots showing the correlation between two randomly chosen groups of human volunteers on the Twitter corpus (left) and Ubuntu Dialogue Corpus (right).
Reality (BLEU)
Reality (ROUGE & METEOR)

(a) ROUGE

(b) METEOR
Correlation Results

Original paper (Liu et al., 2016):  

<table>
<thead>
<tr>
<th>Metric</th>
<th>Spearman</th>
<th>p-value</th>
<th>Spearman</th>
<th>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>
</tr>
<tr>
<td>Average</td>
<td>0.2259</td>
<td>0.024</td>
<td>0.1971</td>
<td>0.049</td>
</tr>
<tr>
<td>Extrema</td>
<td>0.2103</td>
<td>0.036</td>
<td>0.1842</td>
<td>0.067</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.1887</td>
<td>0.06</td>
<td>0.1927</td>
<td>0.055</td>
</tr>
<tr>
<td>BLEU-1</td>
<td>0.1665</td>
<td>0.098</td>
<td>0.1288</td>
<td>0.2</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>
</tr>
<tr>
<td>BLEU-3</td>
<td>0.3423</td>
<td>&lt; 0.01</td>
<td>0.1443</td>
<td>0.15</td>
</tr>
<tr>
<td>BLEU-4</td>
<td>0.3417</td>
<td>&lt; 0.01</td>
<td>0.1392</td>
<td>0.17</td>
</tr>
<tr>
<td>ROUGE</td>
<td>0.1235</td>
<td>0.22</td>
<td>0.09714</td>
<td>0.34</td>
</tr>
<tr>
<td>Human</td>
<td>0.9476</td>
<td>&lt; 0.01</td>
<td>1.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

After removing pre-processing artifacts (<speaker> token):  

<table>
<thead>
<tr>
<th>Metric</th>
<th>Spearman</th>
<th>Pearson</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU-1</td>
<td>-0.026 (0.80)</td>
<td>0.016 (0.87)</td>
</tr>
<tr>
<td>BLEU-2</td>
<td>0.065 (0.52)</td>
<td>0.080 (0.43)</td>
</tr>
<tr>
<td>BLEU-3</td>
<td>0.139 (0.17)</td>
<td>0.088 (0.39)</td>
</tr>
<tr>
<td>BLEU-4</td>
<td>0.139 (0.17)</td>
<td>0.092 (0.36)</td>
</tr>
<tr>
<td>ROUGE</td>
<td>-0.083 (0.41)</td>
<td>-0.010 (0.92)</td>
</tr>
</tbody>
</table>

Word-overlap metrics are **poor substitute for human evaluations**
A dialogue response is probably good if it is rated highly by humans.

- Collect a labelled dataset of human scores of responses

- Build a model that learns to predict human scores of response quality (ADEM)

- Condition response score on the reference response and the context
Hey, want to go to the movies tonight?

Nah, let’s do something active.

Yeah, the film about Turing looks great!
Hey, want to go to the movies tonight?

Seen any good movies recently?

Nah, let’s do something active.

Yeah, the film about Turing looks great!

Dialogue response score should also depend on context!
Evaluation dataset

Conducted 2 rounds of AMT studies to get evaluation on Twitter

**Study 1:** ask workers to generate next sentence of a conversation

**Study 2:** ask workers to evaluate responses from various models (human, TFIDF, HRED, DE)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># Examples</td>
<td>4104</td>
</tr>
<tr>
<td># Contexts</td>
<td>1026</td>
</tr>
<tr>
<td># Training examples</td>
<td>2,872</td>
</tr>
<tr>
<td># Validation examples</td>
<td>616</td>
</tr>
<tr>
<td># Test examples</td>
<td>616</td>
</tr>
<tr>
<td>$\kappa$ score (inter-annotator correlation)</td>
<td>0.63</td>
</tr>
</tbody>
</table>
Our simplifying assumption is that dialogue response quality measured by ‘appropriateness’

In our experiments, other measures (‘topicality’, ‘informativeness’, etc.) either had little inter-annotator agreement, or correlated strongly with ‘appropriateness’

Table 1: Median $\kappa$ inter-annotator agreement scores for various questions asked in the survey.
ADEM

• Given: context $c$, model response $r$, reference response $\hat{r}$ (with embeddings $c$, $r$, $\hat{r}$), compute score as:

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

where $M$, $N$ are parameter matrices, $\alpha$, $\beta$ are constants.

• Trained to minimize squared error:

$$\mathcal{L} = \sum_{i=1:K} [\text{score}(c_i, r_i, \hat{r}_i) - \text{human-score}_i]^2 + \gamma \|\theta\|_1$$
Figure 2: The ADEM model, which uses a hierarchical encoder to produce the context embedding $c$. 

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

• Want model that can learn from limited data (since collection is expensive)

• Pre-train RNN encoder of ADEM using VHRED

Figure 5: The VHRED model used for pre-training. The hierarchical structure of the RNN encoder is shown in the red box around the bottom half of the figure.
Length correlation

**Problem:** humans favour shorter responses, and ADEM can trivially use this for better performance (length gets 0.27 correlation with human score)

**Solution:** bin training set examples by length, re-weight samples such that each length bin has same average score
Utterance-level results

<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>Validation set</th>
<th></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>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>
System-level results

Figure 4: Scatterplots depicting the system-level correlation results for ADEM, BLEU-2, BLEU-4, and ROUGE on the test set. Each point represents the average scores for the responses from a dialogue model (TFIDF, DE, HRED, human). Human scores are shown on the horizontal axis, with normalized metric scores on the vertical axis. The ideal metric has a perfectly linear relationship.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Pearson</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU-1</td>
<td>-0.079 (0.921)</td>
</tr>
<tr>
<td>BLEU-2</td>
<td>0.308 (0.692)</td>
</tr>
<tr>
<td>BLEU-3</td>
<td>-0.537 (0.463)</td>
</tr>
<tr>
<td>BLEU-4</td>
<td>-0.536 (0.464)</td>
</tr>
<tr>
<td>ROUGE</td>
<td>0.268 (0.732)</td>
</tr>
<tr>
<td>ADEM</td>
<td><strong>0.954 (0.046)</strong></td>
</tr>
</tbody>
</table>
Table 4: Correlation for ADEM when various model responses are removed from the training set. The left two columns show performance on the entire test set, and the right two columns show performance on responses only from the dialogue model not seen during training. The last row (25% at random) corresponds to the ADEM model trained on all model responses, but with the same amount of training data as the model above (i.e. 25% less data than the full training set).
How useful is this?

• Moderately. Need to collect more data for better generalization

• Only considers single utterances, rather than a whole dialogue

• What about other aspects of dialogue quality?
Adversarial evaluation

• Rather than imitating human scores, train a model to **distinguish between real and generated responses** (Kannan et al., 2016; Li et al., 2017)

• Similar to discriminator in a GAN

• Combines well with ADEM – want dialogue responses that are appropriate, and similar to human responses

Table 6: Performance of the CAE model in terms of accuracy of predicting $y$. 

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRED</td>
<td>99.28</td>
</tr>
<tr>
<td>VHRED</td>
<td>97.87</td>
</tr>
<tr>
<td>Reference</td>
<td>97.27</td>
</tr>
<tr>
<td>Average</td>
<td><strong>98.14</strong></td>
</tr>
</tbody>
</table>
Problem #4: Entire Premise?
Learning from static datasets

• Will training solely from static datasets lead to a ‘general-purpose communicating agent’?

• Probably not. In this setting, we are primarily learning the statistical structure of language

• But we also want to learn the function of language, and ground the learned language in the agent’s observations

• An alternative approach: have simulated agents in physical environments learn to communicate to solve tasks in that environment (Gauthier & Mordatch, 2016)
Multi-agent language learning
Primary Collaborators

Joelle Pineau Mcgill
Iulian V. Serban U. Montreal
Mike Noseworthy McGill
Chia-Wei Liu McGill
Laurent Charlin HEC Montreal
Nicolas Angelard-Gontier McGill
Nissan Pow McGill
Aaron Courville U. Montreal
Yoshua Bengio U. Montreal
References


Thank you!
Quantitative VHRED results

Table 4: Response information content on 1-turn generation as measured by average utterance length $|U|$, word entropy $H_w = -\sum_{w \in U} p(w) \log p(w)$ and utterance entropy $H_U$ with respect to the maximum-likelihood unigram distribution of the training corpus $p$.

| Model   | $|U|$ | $H_w$ | $H_U$ | $|U|$ | $H_w$ | $H_U$ |
|---------|------|-------|-------|------|-------|-------|
| LSTM    | 11.21| 6.75  | 75.61 | 4.27 | 6.50  | 27.77 |
| HRED    | 11.64| 6.73  | 78.35 | 11.05| 7.53  | 83.16 |
| VHRED   | **12.29** | **6.88** | **84.56** | 9.22 | **7.70** | 71.00 |
| Human   | 20.57| 8.10  | 166.57| 18.30| 8.90  | 162.88|
VHRED results

A

hey zake how's you ? xo

how are you ?

how are you sweetheart

B

thank you

osmir estas de vacaciones nermosa good for you aja

thank you ! i really appreciate your input

high praise . thank you .

ooouuu okay . thank you

that's what it do !

thank you ya di!

love you bb ! ! !

C

a orden rt sinceramente me he matado de la risa con todos tus tweets jajajaja buenisimos ! !

cuando ese nombre del autor , tendre que buscarlo , debe ser super interesante . es una novela ?
do , he oido " jugadas " comentadas entre ellos . asik cuando lo veas me acordate de sus madres .

sili lo he visto ! jaja mira esa es la residencia !

paso ! es + pa q veas a soy buena pera . i dov mi face para co volver a perder contacto con you .

iyyi es como holas soy terno & all day como holas soy sensual e irresistible jsakjsakj xd

pues lo he estado pensando , no creas . buendia primero quiero hacerlo para po .

nou , dat gevoel krij gk . wat ik ga met jari , roben en romulo heen o
Length bias of word overlap metrics

<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)</td>
<td>(\Delta w \geq 6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n=47)</td>
<td>(n=53)</td>
<td></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>

Table 5: Effect of differences in response length for the Twitter dataset, \(\Delta w\) = absolute difference in #words between a ground truth response and proposed response
Where does ADEM do better?

<table>
<thead>
<tr>
<th>Context</th>
<th>Reference response</th>
<th>Model response</th>
<th>Human score</th>
<th>BLEU-2 score</th>
<th>ROUGE score</th>
<th>ADEM score</th>
</tr>
</thead>
<tbody>
<tr>
<td>i’d recommend &lt;url&gt; - or build buy an htpc and put &lt;url&gt; on it. → you’re the some nd person this week that’s recommended roku to me.</td>
<td>an htpc with xmbc is what i run. but i ’ve decked out my setup. i ’ve got &lt;number&gt; tb of data on my home server</td>
<td>because it’s brilliant</td>
<td>5</td>
<td>1.0</td>
<td>1.0</td>
<td>4.726</td>
</tr>
<tr>
<td>imma be an auntie this weekend. i guess i have to go albany. herewego → u supposed to been here → i come off nd on. → never tell me smh</td>
<td>lol you someting</td>
<td>haha, anyway, how’re you?</td>
<td>5</td>
<td>1.0</td>
<td>1.0</td>
<td>4.201</td>
</tr>
<tr>
<td>my son thinks she is plain. and the girl that plays her sister. seekhlp4him? → send him this. he’ll thank you. &lt;url&gt;</td>
<td>you are too kind for words.</td>
<td>i will do</td>
<td>5</td>
<td>1.0</td>
<td>1.0</td>
<td>5.0</td>
</tr>
</tbody>
</table>

Table 8: Examples where both human and ADEM score the model response highly, while BLEU-2 and ROUGE do not. These examples are drawn randomly (i.e. no cherry-picking) from the examples where ADEM outperforms BLEU-2 and ROUGE (as defined in the text). ADEM is able to correctly assign high scores to short responses that have no word-overlap with the reference response. The bars around |metric| indicate that the metric scores have been normalized.