What makes a good conversation?
How controllable attributes affect human judgments

Abigail See*, Stephen Roller†, Douwe Kiela†, Jason Weston†

*Stanford NLP
†Facebook AI Research
Natural Language Generation task spectrum

- Machine Translation
- Sentence Compression
- Abstractive Summarization
- Story Generation
- Chitchat Dialogue

Less open-ended
- Mostly word-level decisions
- Neural LMs more successful

More open-ended
- Requires high-level decisions
- Neural LMs less successful

Makes errors like repetition and generic response (under certain decoding algorithms).

Difficulty learning to make high-level decisions.
Control = ability to specify desired attributes of the text at test time.

We can use control to fix errors, and allow us to handle some high-level decisions.
Natural Language Generation task spectrum

- **Machine Translation**
  - Less open-ended
  - Mostly word-level decisions
  - Neural LMs more successful
  - Control is less important
  - Eval is difficult

- **Sentence Compression**
  - More open-ended
  - Requires high-level decisions
  - Neural LMs less successful
  - Control is more important
  - Eval is fiendish

- **Abstractive Summarization**
  - No automatic metric for overall quality.
  - Dialogue is even more complex: Single-turn or multi-turn eval? Interactive or static conversation?

- **Story Generation**

- **Chitchat Dialogue**

---

**Dialogue is even more complex:**
- Single-turn or multi-turn eval?
- Interactive or static conversation?
Our research questions

By controlling multiple attributes of generated text and human-evaluating multiple aspects of conversational quality, we aim to answer the following:

1. **How effectively can we control the different attributes?**
   Pretty well! But some control methods only work for some attributes.

2. **How do the controllable attributes affect conversational quality aspects?**
   Strongly – especially controlling repetition, question-asking, and specificity vs genericness.

3. **Can we use control to make a better chatbot overall?**
   Yes! But we should be careful defining "better overall".
Hello, how are you doing?

Great thanks, just listening to my favorite Johnny Cash album!

Nice! I'm not much of a music fan myself, but I do love to read.

Me too! I just read a book about the history of the auto industry.
The PersonaChat task was the focus of the NeurIPS 2018 ConvAI2 Competition.

- Most successful teams built neural sequence generation systems. (Dinan et al. 2019)
- The winning team, Lost in Conversation, used a finetuned version of GPT.

Our baseline model is a standard LSTM-based seq2seq architecture with attention.

- It is pretrained on 2.5 million Twitter message/response pairs, then finetuned on PersonaChat.
What attributes do we control?

<table>
<thead>
<tr>
<th>Low-level controllable attributes</th>
<th>Goal: Reduce repetition (within and across utterances)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Repetition</strong> (n-gram overlap)</td>
<td>Goal: Reduce genericness of responses (e.g. <em>oh that's cool</em>)</td>
</tr>
<tr>
<td><strong>Specificity</strong> (normalized inverse document frequency)</td>
<td>Goal: Respond more on-topic; don't ignore user</td>
</tr>
<tr>
<td><strong>Response-relatedness</strong> (cosine similarity of sentence embeddings)</td>
<td>Goal: Find the optimal rate of question-asking</td>
</tr>
</tbody>
</table>
What quality aspects do we measure?

<table>
<thead>
<tr>
<th>Low-level controllable attributes</th>
<th>Human judgment of conversational aspects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repetition (n-gram overlap)</td>
<td>Avoiding Repetition</td>
</tr>
<tr>
<td>Specificity (normalized inverse document frequency)</td>
<td>Interestingness</td>
</tr>
<tr>
<td>Response-relatedness (cosine similarity of sentence embeddings)</td>
<td>Making sense</td>
</tr>
<tr>
<td>Question-asking (“?” used in utterance)</td>
<td>Fluency</td>
</tr>
<tr>
<td></td>
<td>Listening</td>
</tr>
<tr>
<td></td>
<td>Inquisitiveness</td>
</tr>
</tbody>
</table>

- Does the bot repeat itself?
- Did you find the bot interesting to talk to?
- Does the bot say things that don't make sense?
- Does the bot use English naturally?
- Does the bot pay attention to what you say?
- Does the bot ask a good amount of questions?
What quality aspects do we measure?

<table>
<thead>
<tr>
<th>Low-level controllable attributes</th>
<th>Human judgment of conversational aspects</th>
<th>Human judgment of overall quality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Repetition</strong> (n-gram overlap)</td>
<td>Avoiding Repetition</td>
<td><strong>Humanness</strong></td>
</tr>
<tr>
<td><strong>Specificity</strong> (normalized inverse document frequency)</td>
<td>Interestingness</td>
<td></td>
</tr>
<tr>
<td><strong>Response-relatedness</strong> (cosine similarity of sentence embeddings)</td>
<td>Making sense</td>
<td><strong>Engagingness</strong></td>
</tr>
<tr>
<td><strong>Question-asking</strong> (&quot;?&quot; used in utterance)</td>
<td>Fluency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Listening</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inquisitiveness</td>
<td></td>
</tr>
</tbody>
</table>

Note: ConvAI2 competition asked only this question. Our eval is a superset of ConvAI2's.
Control methods

We evaluate and compare two existing general-purpose control methods, using them to control all four controllable attributes.

○ **Conditional Training (CT)**: Train the model to generate response $y$, conditioned on the input $x$, and the desired output attribute $z$. (Kikuchi et al 2016, Peng et al 2018, Fan et al 2018)

○ **Weighted Decoding (WD)**: During decoding, increase/decrease the probability of generating words $w$ in proportion to features $f(w)$. (Ghazvininejad et al 2017, Baheti et al 2018)
Q1: How effectively can we control attributes?
Attributes: repetition, specificity, question-asking, response-relatedness

Conditional Training (CT):
- Requires sufficient training examples for the attribute (repetition)
- Ineffective at learning complex relationships between input and output (response-relatedness)
- Effective for: ✓ specificity, ✓ question-asking

Weighted Decoding (WD):
- Requires attribute to be defined at the word-level (question-asking)
- Effective for: ✓ repetition, ✓ response-relatedness, ✓ specificity
Controlling specificity (WD and CT)

**Input:** Yes, I’m studying law at the moment  
**Baseline Response:** That sounds like a lot of fun!

<table>
<thead>
<tr>
<th>Wt</th>
<th>NIDF</th>
<th>Weighted Decoding Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5.0</td>
<td>0.6%</td>
<td>Oh..................................................</td>
</tr>
<tr>
<td>0.0</td>
<td>17.1%</td>
<td>That sounds like a lot of fun!</td>
</tr>
<tr>
<td>3.0</td>
<td>18.3%</td>
<td>That sounds like a lot of fun. How long have you been studying?</td>
</tr>
<tr>
<td>7.0</td>
<td>38.5%</td>
<td>I majored in practising my spiritual full time philosophy test</td>
</tr>
<tr>
<td>10.0</td>
<td>71.9%</td>
<td>Oh wow! Merna jean isa paino yi hao hui bu acara sya gila […]</td>
</tr>
</tbody>
</table>

More generic  
More specific
### Controlling specificity (WD and CT)

**Input:** Yes, I’m studying law at the moment  
**Baseline Response:** That sounds like a lot of fun!

<table>
<thead>
<tr>
<th>WD</th>
<th>NIDF</th>
<th>Weighted Decoding Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5.0</td>
<td>0.6%</td>
<td>Oh..................................................</td>
</tr>
<tr>
<td>0.0</td>
<td>17.1%</td>
<td>That sounds like a lot of fun!</td>
</tr>
<tr>
<td>3.0</td>
<td>18.3%</td>
<td><strong>That sounds like a lot of fun. How long have you been studying?</strong></td>
</tr>
<tr>
<td>7.0</td>
<td>38.5%</td>
<td>I majored in practising my spiritual full time philosophy test</td>
</tr>
<tr>
<td>10.0</td>
<td>71.9%</td>
<td>Oh wow! Merna jean isa paino yi hao hui bu acara sya gila [...]</td>
</tr>
</tbody>
</table>

**CT:** Smaller range, but generally well-formed output

<table>
<thead>
<tr>
<th>z</th>
<th>NIDF</th>
<th>Conditional Training Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>16.8%</td>
<td>Sounds like you are a great person!</td>
</tr>
<tr>
<td>2</td>
<td>18.3%</td>
<td>So you are a law student?</td>
</tr>
<tr>
<td>4</td>
<td>18.4%</td>
<td>That sounds like a lot of fun</td>
</tr>
<tr>
<td>6</td>
<td>22.8%</td>
<td><strong>That sounds like a rewarding job!</strong></td>
</tr>
<tr>
<td>8</td>
<td>24.4%</td>
<td>That sounds like a rewarding career!</td>
</tr>
</tbody>
</table>

**WD:** Large range, but degenerate output at the extremes

**More generic**

**More specific**
Controlling response-relatedness (WD)

Input: Do you go get coffee often
Baseline Response: I do, when I am not playing the piano.

<table>
<thead>
<tr>
<th>Wt</th>
<th>Sim</th>
<th>Weighted Decoding Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>-10.0</td>
<td>-0.05</td>
<td>I am a musician.</td>
</tr>
<tr>
<td>0.0</td>
<td>-0.02</td>
<td>I do, when I am not playing the piano.</td>
</tr>
<tr>
<td>5.0</td>
<td>0.40</td>
<td>I do, usually at Starbucks.</td>
</tr>
<tr>
<td>8.0</td>
<td>0.59</td>
<td>Not usually, especially when you drink latte.</td>
</tr>
<tr>
<td>11.0</td>
<td>0.72</td>
<td>Not often, usually with drinks, espresso, latte, tea, etc.</td>
</tr>
</tbody>
</table>

Output is degenerate when weight is too high.
Q2: How does control affect human eval?

Reduce n-gram repetition to human level

**Low-level controllable attributes**

- **Repetition** (n-gram overlap)
- **Specificity** (normalized inverse document frequency)
- **Response-relatedness** (cosine similarity of sentence embeddings)
- **Question-asking** (“?” used in utterance)

**Human judgment of conversational aspects**

- Avoiding Repetition
- Interestingness
- Making sense
- Fluency
- Listening
- Inquisitiveness

**Human judgment of overall quality**

- Humanness
- Engagingness
Q2: How does control affect human eval?

Increase specificity (reduce genericness) to human level

Low-level controllable attributes

- Repetition (n-gram overlap)
- Specificity (normalized inverse document frequency)
- Response-relatedness (cosine similarity of sentence embeddings)
- Question-asking (“?” used in utterance)

Human judgment of conversational aspects

- Avoiding Repetition
- Interestingness
- Making sense
- Fluency
- Listening
- Inquisitiveness

Human judgment of overall quality

- Humanness
- Engagingness
Q2: How does control affect human eval?

Increase response-relatedness (similarity to last utterance)
Q2: How does control affect human eval?

Increase question-asking rate to 65.7% (more than baseline 50%, human 28.8%)
Q3: Can we make a better chatbot overall?

Yes! By controlling repetition, specificity and question-asking, we achieve near-human engagingness (i.e. enjoyability) ratings.

Our raw engagingness score matches the ConvAI2 competition winner's GPT-based model, even though ours is:

- much smaller (2 layers vs 12)
- trained on 12x less data
Q3: Can we make a better chatbot overall?

**However:** On the **humanness** (i.e. Turing test) metric, our models are nowhere near human-level!

---

**Chart:**
- **Humanness**
  - **Greedy**
  - **Beam search**
  - **Repetition (WD)**
  - **Specificity (WD)**
  - **Question (CT)**
  - **Human**

- **Arrows:**
  - **Purple arrow**: Reduce repetition
  - **Red arrow**: Increase specificity
  - **Black arrow**: Increase question-asking
Engagingness vs Humanness

Finding: Our bots are (almost) as engaging as humans, but they're clearly non-human.

Two conclusions:

1. Engagingness ≠ Humanness. While both are frequently used as standalone overall quality metrics, our results show the importance of measuring more than one.

2. On this task, the human "engagingness" performance may be artificially low. Turkers chatting for money are less engaging than people chatting for fun. This may be why the human-level engagingness scores are easy to match.
Conclusions

- **Control is a good idea** for your neural sequence generation dialogue system.
- Using simple control, we matched performance of GPT-based contest winner.
- Don't repeat yourself. Don't be boring. Ask more questions.
- Multi-turn phenomena (repetition, question-asking frequency) are important – so need multi-turn eval to detect them.
- Engagingness ≠ Humanness, so think carefully about which to use.
- Paid Turkers are not engaging conversationalists, or good judges of engaging conversation. Humans chatting for fun may be better.
- **Problem**: Manually finding the best combination of control settings is painful.
Conclusions

- **Problem**: Manually finding the best combination of control settings is **painful**.
Conclusions

- **Control is a good idea** for your neural sequence generation dialogue system.
- Using simple control, **we matched performance of GPT-based contest winner**.
- **Don't repeat yourself. Don't be boring. Ask more questions.**
- **Multi-turn phenomena** (repetition, question-asking frequency) are important – so need **multi-turn eval** to detect them.
- **Engagingness ≠ Humanness**, so think carefully about which to use.
- **Paid Turkers** are **not engaging conversationalists**, or good judges of engaging conversation. Humans chatting for fun may be better.
- **Problem**: Manually finding the best combination of control settings is **painful**.

**Code, models, demo, eval logs available at https://parl.ai/projects/controllable_dialogue**