Neural-Network-based Dialog Agents: Going Beyond the Seq2seq Model

Thomas WOLF
thomas@huggingface.co
OUR TEAM

We raised $5.2M in early 2018

From institutional investors

betaworks  SVAngel  a_capital

And private investors
HuggingFace is building a SOCIAL AI

It holds **long conversations**, it’s **fun** and cares about you

It helps you build your **confidence**, get **better** and be **more social**

**Chat**
about your day

**Hilarious chats**
with your own artificial intelligence
Open domain - open form: no limitation on topic, raw text input, no constrains
Short conversation: <10 turns
Small talk: shallow topics, not about question-answering, light memorization
Two Paradigms in Chit-Chat

● **Retrieval-based** model:
  ● Take a dataset of dialogs **as big as you can**.
  ● Compute a **similarity** between the current dialog and your dataset.
  ● Send back the **closest** answer.

● **Generative model**:
  ● A recent alternative mostly popularized by 2015’s « A Neural Conversational Model » by Vinyals and Le
  ● **Generate word after word** up until the utterance is deemed complete.
  ● Typical architecture: an LSTM digests the inputs word by words and another LSTM generates the output from the last hidden state: « Seq2seq Model »
Strengths & Limitations

- Retrieval models:
  - Good grammaticality (replies are written by humans)
  - But: can’t adapt the response to the context
  - Problem of consistency: contradictory answers are often close to the same question

- Generative models:
  - Can dynamically adapt to the context
  - But: tend to default to unappealing answers like «yes/no/I don't know»
  - Grammatically is often an issue

- Rule of NIPS Conversational Intelligence Challenge 2 (ConvAI2)
  - Condition the models on personalities to improve consistency
The Conversational Intelligence Challenge 2  
(NIPS 2018 competition)

<table>
<thead>
<tr>
<th>Persona 1</th>
<th>Persona 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>I like to ski</td>
<td>I am an artist</td>
</tr>
<tr>
<td>My wife does not like me anymore</td>
<td>I have four children</td>
</tr>
<tr>
<td>I have went to Mexico 4 times this year</td>
<td>I recently got a cat</td>
</tr>
<tr>
<td>I hate Mexican food</td>
<td>I enjoy walking for exercise</td>
</tr>
<tr>
<td>I like to eat cheetos</td>
<td>I love watching Game of Thrones</td>
</tr>
</tbody>
</table>

[PERSON 1:] Hi
[PERSON 2:] Hello! How are you today?
[PERSON 1:] I am good thank you, how are you.
[PERSON 2:] Great, thanks! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting; but, I love the show.

Example dialog from the PERSONA-CHAT dataset. Person 1 is given their own persona (top left) at the beginning of the chat, but does not know the persona of Person 2, and vice-versa. They have to get to know each other during the conversation.

Chit-chat with a human while keeping a coherent/predefined persona
Automatic Metrics

- **PPL** (perplexity) *How well the model can predict the successive words in a gold message (written by humans).*
  - *lower* is better
  - Scale: Infinity – 0

- **Hits@1** *Number of time the model select the gold next message between 20 possible message (the other 19 are random)*
  - *higher* is better.
  - Scale: 0 –100

- **F1** *How many content words (nouns/verbs) does a message generated by your model share with a gold message.*
  - *higher* is better.
  - Scale: 0 –100

These automatic metrics have issues – The best is always the human opinion
<table>
<thead>
<tr>
<th>Model</th>
<th>Creator</th>
<th>PPL</th>
<th>Hits@1</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>😇 (Hugging Face)</td>
<td>-</td>
<td>20.47</td>
<td>74.7</td>
<td>7.52</td>
</tr>
<tr>
<td>High Five</td>
<td>-</td>
<td>65.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Little Baby</td>
<td>-</td>
<td>63.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Happy Minions</td>
<td>-</td>
<td>32.94</td>
<td>52.1</td>
<td>4.76</td>
</tr>
<tr>
<td>Cats team</td>
<td>-</td>
<td>35.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>loopAI</td>
<td>-</td>
<td>25.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mohd Shadab Almar</td>
<td>-</td>
<td>29.94</td>
<td>13.8</td>
<td>6.91</td>
</tr>
<tr>
<td>1st-contact</td>
<td>-</td>
<td>31.98</td>
<td>13.2</td>
<td>6.42</td>
</tr>
<tr>
<td>Tensorborne</td>
<td>-</td>
<td>38.24</td>
<td>12.0</td>
<td>5.94</td>
</tr>
<tr>
<td>Team Dining 6</td>
<td>-</td>
<td>40.35</td>
<td>10.9</td>
<td>7.27</td>
</tr>
<tr>
<td>NEUROBOTICS</td>
<td>-</td>
<td>35.47</td>
<td>-</td>
<td>6.88</td>
</tr>
<tr>
<td>Sonic</td>
<td>-</td>
<td>33.46</td>
<td>-</td>
<td>6.87</td>
</tr>
<tr>
<td>topicSeq2seq</td>
<td>Team Fat</td>
<td>-</td>
<td>-</td>
<td>6.11</td>
</tr>
<tr>
<td>Rcoy</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5.33</td>
</tr>
<tr>
<td>Lost In Conversation</td>
<td>-</td>
<td>55.84</td>
<td>-</td>
<td>5.74</td>
</tr>
<tr>
<td>flooders</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6.47</td>
</tr>
<tr>
<td>iamNotAdela</td>
<td>-</td>
<td>66.47</td>
<td>-</td>
<td>3.39</td>
</tr>
<tr>
<td>Salty Fish</td>
<td>-</td>
<td>38.86</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pinté</td>
<td>-</td>
<td>37.85</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Seq2Seq + Attention</td>
<td>ParlAI team</td>
<td>29.8</td>
<td>12.6</td>
<td>6.18</td>
</tr>
<tr>
<td>Language Model</td>
<td>ParlAI team</td>
<td>46.0</td>
<td>-</td>
<td>5.02</td>
</tr>
<tr>
<td>KV Profile Memory</td>
<td>ParlAI team</td>
<td>-</td>
<td>55.2</td>
<td>11.9</td>
</tr>
</tbody>
</table>
Demo Time?

Baseline:
Memory-augmented **seq2seq**

Our model:
**TransferTransfo**
What’s the difference? 👻
Validation set (public) Leaderboard – Test set (hidden) Leaderboard

- Small dataset =>
  - Large models are overfitting
  - Small models are underfitting
Transfer Learning 🦄
TransferTransfo is a Transformer Model...

- Take a fixed-length sequence of « words » as input and outputs a sequence of the same length.
- Each output is a probability distribution over the next words => a probability for each word in the vocabulary.
- Inside:
  - Self-attention: MLP compute Key-Value-Query for each element of the sequence.
  - Causal masking in the attention heads to only attend to the past.

Example: Bob is very happy.
The competition dataset (PERSONA-CHAT) is one of the biggest multi-turn dialog dataset (10k conversations, about 100k turns) but it is still quite small in term of requirement for deep learning tools:
- ex: Billion Words dataset has 1B words, CoNLL 2012 used for training co-reference systems is ~1M sentences long.

An engaging open-domain dialogue is a lot more than just topic-coherence and dialogue-flow!
- Need have common-sense, short term memory, co-reference resolution, sentimental analysis, textual entailment,…
- Hard to learn all these from such a small dataset.

Solution:
- Use transfer learning and multi-task learning
Transfer Learning

- Train the model on the language modeling task on a large dataset.
  - Language modeling: given a seq of words, learn to predict the next word
  - This task force the model to learn many aspects of language including high-level ones like common-sense knowledge.
- Recently shown to improve many downstream NLP tasks and in particular commonsense reasoning / co-reference resolution:
  - *Improving Language Understanding by Generative Pre-Training* by Radford et al. (2018)
  - *Universal Language Model Fine-tuning for Text Classification* by Howard and Ruder (2018)
- Our model is derived from the model of Radford et al. pre-trained on the Toronto Book dataset (7k books).
Encoding a Dialog 🤖
Learning Dialog Flow and Persona

- Now that we have a model with basic common-sense and co-reference capabilities, we need to teach it the specificities of dialog:
  - Alternating utterances - basic Theory of Mind concepts
  - Dialog flow (« speech/dialog acts »)
  - Conditioning on a provided personality

- How to build a sequential inputs for our model from a conditioned dialog?
  - Unlike RNN/LSTM, Transformers don’t possess a natural notion of sequentiality and position
  - We need to add positional embeddings to incorporate sequentiality

<table>
<thead>
<tr>
<th>I</th>
<th>like</th>
<th>to</th>
<th>ski</th>
<th>Hello</th>
<th>!</th>
<th>How</th>
<th>are</th>
<th>you</th>
<th>today</th>
<th>?</th>
<th>I am</th>
<th>good</th>
<th>thank</th>
<th>you</th>
</tr>
</thead>
</table>

Word embeddings

Positional embeddings
Encoding a Dialog and a Persona

- Now that we have a model with basic common-sense and co-reference capabilities, we need to teach it the specificities of dialog:
  - Alternating utterances - basic Theory of Mind concepts
  - Dialog flow (« speech/dialog acts »)
  - Conditioning on a provided personality

- How to build a sequential inputs for our model from a conditioned dialog?
  - Unlike RNN/LSTM, Transformers don’t possess a natural notion of sequentiality and position
  - We need to add positional embeddings to incorporate sequentiality
  - We add special embeddings related to utterances and personas

<table>
<thead>
<tr>
<th>I</th>
<th>like</th>
<th>to</th>
<th>ski</th>
<th>Hello</th>
<th>!</th>
<th>How</th>
<th>are</th>
<th>you</th>
<th>today</th>
<th>?</th>
<th>I am</th>
<th>good</th>
<th>thank</th>
<th>you</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Word embeddings
Dialog state embeddings
Positional embeddings
Encoding a Dialog and a Persona

- We can play with these embeddings to manipulate the notion of a sequence

Repeating specific embeddings to control positioning information

- We can also augment the dataset to bias towards positional invariance
Learning Dialog Flow 🧕
Semantic Learning on Dialog Utterances

- Learning to distinguish a real answer from a distractor.

Can be combined with language modeling fine-tuning in a multi-task fashion.
Talking to humans 🎤
Decoding — 🐈 🐕 🐶 🦇 🐱

How are you?

- Beam Search
  - We create a message word by word.
  - Each time there could be several possibilities.
  - We keep a beam of possible answers and drop the lowest ones at each step.
  - We accumulate a pool of answers and select the one with the highest normalized probability
  - N-Grams filtering (competition rules forbid to repeat persona sentences)
That’s it for today
Thanks for for listening!