Using adaptive knowledge graphs in neural dialogue generation

Patrik Purgai
Chatbots

Domain specific
- Rasa
- Luis
- Dialogflow

Open domain
- Blender
- Cleverbot
Chatbots

- Rule-based
- Neural Network based
  - Retrieval based
  - Generative
Rule-based

- Pattern matching - ELIZA
- AIML (Artificial Intelligence Markup Language) - Mitsuku
- Chatscript - SUZETTE
<category>
  <pattern>MY NAME IS *</pattern>
  <template>Hi there, my name is Andrew.</template>
</category>

<category>
  <pattern>GUESS A NUMBER</pattern>
  <template>
    <random>
      <li>1</li>
      <li>6</li>
      <li>227</li>
    </random>
  </template>
</category>
Neural Network based

- RNN (LSTM)
- Transformer
- BERT
- GPT-2
RNN
LSTM
Hochreiter & Schmidhuber (1997)
Transformer
Vaswani et al. (2017)
Attention Mechanism

The animal didn't cross the street because it was too tired. The animal didn't cross the street because it was too wide.
BERT
Radford et al. (2019)
BERT vs GPT-2
Retrieval-based

- Speaker-Aware BERT - Gu et al. (2020)
- Poly-encoders - Humeau et al. (2019)
I am fine thanks!
My name is Peter.
I am 12 years old.
I live in Mexico.
Cross-Encoder

Classifier($u_1$)

Cross Encoder

- How are you? [SEP] I live in Mexico.
- How are you? [SEP] I am fine thanks!
- How are you? [SEP] I am 12 years old.
Speaker Aware BERT

[SP1] Hi [SP2] Hello [SP1] How are you? [SEP] [SP2] I am fine thanks!
I am fine thanks!
My name is Peter.
I am 12 years old.
I live in Mexico.
Neural Network based - Generative

- A Neural Conversational Model - Vinyals et al. (2015)
- DialoGPT - Zhang et al. (2019)
- Towards a Human-like Open-Domain Chatbot (Meena) - Adiwardana et al. (2020)
- Recipes for building an open-domain chatbot - Roller et al. (2020)
Can you please come here?
Seq2Seq
Sutskever et al. (2015)
Neural Conversational Model
Vinyals et al. (2015)
DialoGPT
Zhang et al. (2019)

- Reddit conversations
  - 146.8M training instances
  - 3.87B source tokens
  - 2.14B target tokens

- Different Models
  - 117M
  - 345M + Reverse
  - 762M
Does money buy happiness?
Depends how much money you spend on it.
What is the best way to buy happiness?
You just have to be a millionaire by your early 20s, then you can be happy.
This is so difficult!
You have no idea how hard it is to be a millionaire and happy. There is a reason the rich have a lot of money.
Meena
Adiwardana et al. (2020)

User: How are things?

Meena: They're good. How about you?

User: I'm trying to decide what to watch

Meena: What are some of your favorite shows?

User: I like the Good Place

Meena: I haven't watched that. My favorite shows are Supernatural, and Star Trek: TNG.

User: What is TNG?

Meena: The Next Generation
Blender
Roller et al. (2020)

- **Blending skills**
  - Engaging use of personality (PersonaChat)
  - Engaging use of knowledge (Wizard of Wikipedia)
  - Display of empathy (Empathetic Dialogues)
  - Ability to blend all three seamlessly (BST)

- **Retrieve and refine**
Generative Decoding

- **Deterministic**
  - Greedy Decoding
  - Beam Search

- **Sampling**
  - Top-k
  - Nucleus

- **Similarity search (Continuous output)**
Hyperparameter: K (beam width)

Greedy decoding K = 1
Token Probabilities for "I don't know." Repeated 200 times
Holtzman et al. (2020)

Beam Search Text is Less Surprising
Enhancing Beam search

- **Response length**
  - Hard coded minimum
  - Response length prediction
    (Blender) 4 class classifier (<10, <20, <30, >30)

- **Subsequence blocking**
  Not allowing existing n-grams
Top-k decoding

Hyperparameter: $K$ (sample size)
Broad Distribution

She said, "I never thought, knew, had, saw, did, said, wanted, told, liked, got."

Narrow Distribution

I ate the pizza while it was still hot.
Nucleus decoding

Hyperparameter: $p$ (probability mass)

\[
\sum_{w \in V_{\text{top-}p}} P(w \mid \text{“The”}) = 0.94
\]

\[
\sum_{w \in V_{\text{top-}p}} P(w \mid \text{“The”, “car”}) = 0.97
\]
MMI Decoding

Forward + Backward language model

1. The forward model generates the output and computes $P_{\text{forward}}$
2. The backward model computes $P_{\text{backward}}$

Output is then scored by $P_{\text{backward}} + P_{\text{forward}}$
Similarity Search
(Continuous output)

Method:
- Model output is not projected onto the vocabulary
- Objective is to minimize $\text{Sim}(v_{\text{output}}, v_{\text{target}})$
- $V_{\text{target}}$ is drawn from pre-trained embeddings

Advantages:
- Faster training
- Interpretable output
Issue with end-to-end solutions

● Bad interpretability
  ○ Knowledge is stored as parameter values
  ○ We can’t inspect the cause of a bad answer

● Information hallucination
  Example: "Cats have three legs"
Knowledge graphs
Where is the Mona Lisa located?

It is located in Paris.

(Mona Lisa, is_in, Louvre)

(Louvre, is_located_in, Paris)
Architecture

- **Word Embedding**
  - Fasttext
- **Graph Embedding**
  - TransE
- **Graph Decoder**
  - LSTM
- **Response Generator**
  - Transformer
  - Continuous output
- **Entity Linker**
  - BERT
Dataset

OpenDialKG
Moon et al. (2019)

91031 messages
100813 entities
1358 relations
1190658 triplets

[{
  "Message": "Do you like Iron Man",
  "Metadata": [
    
    [ "Iron Man",
      "starred_actors",
      "Robert Downey Jr."
    ]
  
  ]
}]

Word embeddings
(Fasttext)
Graph embeddings

- **DeepWalk** - Perozzi et al. (2014)
- **TransE** - Bordes et al. (2013)
- **TransR** - Lin et al. (2015)
TransE
Bordes et al. (2013)

Knowledge triplet:
Mona Lisa (head), is_in (relation), Louvre (tail)

\[ V_{\text{head}} + V_{\text{relation}} = V_{\text{tail}} \]
TransR
Lin et al. (2014)

Each relation has its own transformation matrix: $M_r$

$$M_r v_h + v_r = M_r v_t$$
Entity Linker

Limited information in OpenDialKG

Labels are learned with BCE objective
Where is the **Mona Lisa** located?
Response Generator

- **Transformer encoder-decoder architecture**
  - additional cross-attention layer for the entity node representations

- **Continuous outputs with dot-product loss + negative sampling**
  - (Future work) Decoding entity nodes directly
Response Generator

Where is the *Mona Lisa* located?

It is located in Paris.
Baselines

- GPT-2 (345M)
- XLNet (110M)
- Vanilla Transformer
- Continuous Transformer
Automatic Metrics

- **Length**
  Generated data: Average length of model outputs
  Test data: Delta length compared to gold responses

- **BLEU**
  Modified precision score of n-gram overlaps
  Common metric in NMT
  Low correlation with human judgement

- **WMD**
WMD
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Decoding</th>
<th>Length</th>
<th>WMD</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Std</td>
<td>Mean</td>
</tr>
<tr>
<td>GPT-2 (small)</td>
<td>Greedy decoding</td>
<td>10.2</td>
<td>3.5</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>Beam Search $k = 3$</td>
<td>12.1</td>
<td>3.6</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>Beam Search $k = 5$</td>
<td>12.0</td>
<td>3.5</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>Top-k Sampling $k = 100$</td>
<td>14.2</td>
<td>3.8</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>Nucleus decoding $p = 0.1$</td>
<td>12.1</td>
<td>3.9</td>
<td>1.7</td>
</tr>
<tr>
<td>XLNet (base)</td>
<td>Greedy decoding</td>
<td>9.4</td>
<td>3.4</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>Beam Search $k = 3$</td>
<td>11.2</td>
<td>3.2</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>Beam Search $k = 5$</td>
<td>12.9</td>
<td>3.3</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>Top-k Sampling $k = 100$</td>
<td>14.3</td>
<td>4.3</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>Nucleus decoding $p = 0.1$</td>
<td>14.1</td>
<td>4.5</td>
<td>1.8</td>
</tr>
<tr>
<td>Transformer</td>
<td>Greedy decoding</td>
<td>7.1</td>
<td>2.2</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>Beam Search $k = 3$</td>
<td>8.5</td>
<td>2.0</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>Beam Search $k = 5$</td>
<td>9.4</td>
<td>2.8</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>Top-k Sampling $k = 100$</td>
<td>9.4</td>
<td>4.1</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>Nucleus decoding $p = 0.1$</td>
<td>9.6</td>
<td>4.7</td>
<td>2.3</td>
</tr>
<tr>
<td>Continuous Transformer</td>
<td>Similarity Search</td>
<td>9.8</td>
<td>3.3</td>
<td>2.1</td>
</tr>
<tr>
<td>KG-Enhanced Transformer</td>
<td>Similarity Search</td>
<td>12.8</td>
<td>3.4</td>
<td>1.6</td>
</tr>
</tbody>
</table>
Generated dialogues
<table>
<thead>
<tr>
<th>Input</th>
<th>Method</th>
<th>Model Output</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>I like the movie 'The Martian', could you recommend a movie similar to it?</td>
<td>GPT-2 (small) Beam Search $k = 5$</td>
<td>You might enjoy Cloverfield with Drew Barrymore or The Martian. They are both Sci-Fi films.</td>
<td>Do you like The Martian because it stars Matt Damon?</td>
</tr>
<tr>
<td></td>
<td>GPT-2 (small) Nucleus decoding $p = 0.1$</td>
<td>Sure, do you prefer Sci-Fi or Adventure movies?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>XLNet (base) Beam Search $k = 5$</td>
<td>It is starring Michael Fassbender, who also starred in X-Men: First Class and X-Man 2. Do you like movies that he stars in?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>XLNet (base) Nucleus decoding $p = 0.1$</td>
<td>I do! I love Sci Fi movies as well. Have you seen either of those?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Transformer Nucleus decoding $p = 0.1$</td>
<td>This is a good one</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Continuous Transformer</td>
<td>Have you seen The Running Man?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KG-Enhanced Transformer</td>
<td>It is written by James Patterson, starring Matt Damon and also stars Tracie Thoms. Do you like them?</td>
<td></td>
</tr>
</tbody>
</table>
KG-enhanced transformer

- [Human] Hi
- [Model] Hello, have you heard of the movie The Hobbit: Desolation of Smaug?
- [Human] No, can you tell me more about it?
- [Model] It is a fantasy adventure film starring James Nesbitt. Do you like this actor or would you like to know other movies he has starred in?
- [Human] List some of his movies please.
- [Model] Sure! He also starred in The Way and The Amateurs. Have you seen either of those?
- [Human] What is 'The Way' about?
- [Model] It’s a teen film, but it was directed by Marc Forster who also wrote the movie The Valley of Elah which stars Tom Hanks.
- [Human] Oh I see. Can you recommend me a Sci-Fi movie?
- [Model] Sure! White Television and Undertow are good ones as well. Have you seen either of those?
GPT-2

- [Human] hi
- [Model] greetings, i’m gerald.
- [Human] how are you ?
- [Model] good, thanks. getting ready for another hike, i love hiking!
- [Human] oh realy? where are you going?
- [Model] pikes peak. i am curently exploring famous mountains when i have time.
- [Human] that is cool are you going alone?
- [Model] my brother is in the us, we go hiking with him every month.
- [Human] how old are you by the way?
- [Model] a little under 46 years old
- [Human] are you in a relationship?
- [Model] yes i am married, 3 boys, i am pregnant again. i am too busy to focus
Conclusion

- Lack of automatic metrics for checking factual accuracy
- Lack of datasets for this purpose
- Unsatisfactory results from the proposed architecture

+ I paved the way for future work
+ I created a framework for training chatbots from pre-trained transformers:

https://github.com/bme-chatbotts/dialogue-generation
Future work

- Implementing pre-trained transformer as the response generator
- Improving the OpenDialKG dataset
- Finding better automatic metric
References

- arxiv.org/abs/1706.03762 (Transformer)
- jalammarr.github.io/illustrated-gpt2/ (GPT-2)
- yashuseth.blog/2019/10/08/ (Knowledge graph)