

Creating Transformer-Based Dialogue Agents with Distinct Personalities using Transfer Learning

Ona Oghogho Igbinedion
Department of Computer Science,
Loyola Marymount University
Los Angeles, United States of America

Claudia Gusti
Department of Computer Science,
Loyola Marymount University
Los Angeles, United States of America

Abstract—This paper intends to explore conversational AI, by using transfer learning to create a persona-based dialogue agent. We used the pre-trained models GPT and GPT-2 from OpenAI, in order to avoid having to train on a large text corpus. We preprocessed Facebook’s PERSONA CHAT dataset and fine-tuned the models on conversational dialogue with defined personas. After training the models, we applied a decoder using top-k and top-p sampling in order to generate the dialogue agent’s responses. We evaluated the models using automated metrics such as perplexity, F1 score and hits@k score and human evaluation metrics such as coherence, relevance, engagement and fluency. We evaluated our fine-tuned GPT and GPT-2 models on ConvAI’s 2018 hidden dialogue data set and found that our models outperformed a Seq2Seq + Attention baseline in terms of F1 and Hits@1 but not perplexity. Additionally, our human evaluations resulted in decently high scores for both models across all 4 evaluation metrics and we were also able to gather valuable feedback from our human evaluators on their conversations with our dialogue agents.

Keywords—*Conversational AI; Dialogue Agent; Chatbot; Transfer Learning; GPT; GPT-2; Transformers; HuggingFace; PERSONA CHAT*

I. INTRODUCTION

A few years ago, creating a dialogue agent -as limited as they were back then- could take months, from designing the rules to actually

writing thousands of answers to cover some of the conversation topics. The most commonly used pretrained NLP model, BERT (which stands for *Bidirectional Encoder Representations from Transformers*), is pretrained on full sentences only and is not able to complete unfinished sentences [1]. In 2018 and 2019, Alec Radford, Jeffrey Wu and their co-workers at OpenAI open-sourced two language models trained on a very large amount of data: GPT and GPT-2 (where *GPT* stands for *Generative Pretrained Transformer*) [2][3].

For our final project we proposed to create a conversational AI, or dialogue agent, and imbue it with a given personality. We use transfer learning via the pretrained GPT and GPT2 models and fine-tuned the models on conversational data from Facebook’s PERSONA CHAT dataset. After training, we then pass the output of our fine-tuned models to the decoder algorithm. We use the top-k and top-p sampling in order to generate the dialogue agent’s responses. After implementing the dialogue agents, we then compared the perplexity, F1 score, and hits@k score of each model. We calculated these scores by creating a script to simulate a chit-chat conversation

between the 2 dialogue agents using the ConvAI hidden dialogue data set [8].

A. Data

As stated previously, our data was sourced from Facebook’s PERSONA CHAT dataset. This dataset contained over ten thousand dialogues between two personas. Each dialogue contained the persona of the first entity, the persona of the second entity, and the labeled conversation that occurred between the two entities. The dialogues came from actors who were instructed to have a chit-chat conversation based on the persona they were given [4].

The dataset was structured as a dictionary consisting of a training set, a validation set, and a hidden set. The training set consisted of 131,438 examples, 17,878 dialogues, and 1,155 different personas. The validation set consisted of 7,801 examples, 1,000 dialogues, and 100 personas. The hidden set consisted of 6,634 examples, 1,015 dialogues and 100 personas [4].

Each dialogue entry consists of a persona, and a list of utterances. Each persona is a list of statements about the given persona to be emulated. Each utterance consists of a set of candidate next statements and a history of what has been said. The candidate statements are structured where the last candidate is the best response, and the other candidates are distractor responses. Both the candidates and the history are stored as a list of statements. Figure 1 depicts an example of a dialogue entry in the dataset.

FIGURE 1

```
EXAMPLE_DIALOG = {
  "personality": ["i like to remodel homes .", "i like to go hunting .",
    "i like to shoot a bow .", "my favorite holiday is halloween ."],
  "utterances": [
    ("candidates": ["hi there . how are you today ?",
      "i totally understand how stressful that can be .",
      "yeah sometimes you do not know what you are actually watching",
      "you must be very fast . hunting is one of my favorite hobbies ."],
    "history": [
      "hi . how are you doing ? i'm getting ready to do some cheetah chasing to stay in shape .",
    ]
  ),
  ("candidates": ["hello i am doing well how are you ?",
    "they tire me so . i probably sleep about 10 hrs a day "
    "because of this .",
    "i also remodel homes when i am not out bow hunting ."],
    "history": [
      "hi . how are you doing ? i'm getting ready to do some cheetah chasing to stay in shape .",
      "you must be very fast . hunting is one of my favorite hobbies .",
      "i am ! for my hobby i like to do canning or some whittling .",
    ]
  )
]
```

Fig. 1. Sample data entry

We then tokenized this data, and added 5 additional tokens to the dataset. We added a token that indicated the start of a sequence, a token that indicated the end of a sequence, a token that indicated the first speaker, a token that indicated the second speaker, and a token that was used as padding.

II. MODEL DETAILS

We selected the transformer models GPT and GPT2 because we wanted to use a model pre-trained on a large corpus of data. Figure 2 below shows the underlying architecture of the GPT and GPT2 transformers [2][3]. The primary difference between the two models is that GPT-2 is larger, containing 1.5 billion parameters, 48 layers and 1600 dimensions, while GPT contains 110 million parameters, 12 layers and 768 dimensions. We decided that we wanted to see if changing the pre-trained model would affect the performance of the dialogue agent [2][3].

FIGURE 2

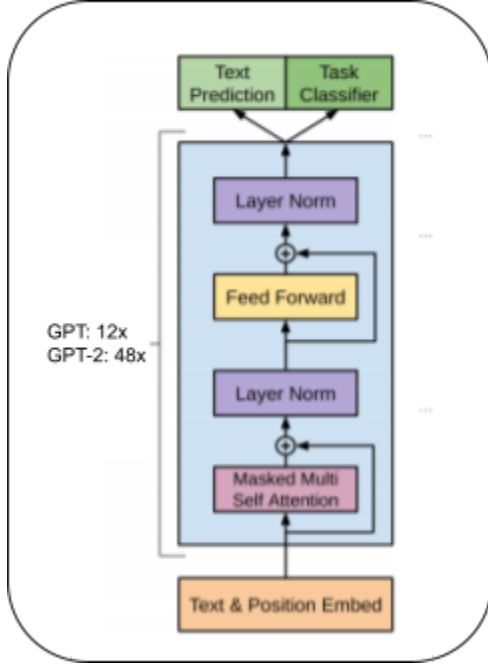


Fig. 2. GPT and GPT-2 Model Architecture [2][3]

We used the OpenAI GPT Double Heads model and the GPT-2 Double Heads model. We used the Double Heads model to simultaneously compute the language model predictions while predicting the next sentence classification labels. Because of this, we used a multi-task loss in order to optimize the model. The total loss was equal to the weighted sum of the language model loss (LML) in addition to the next sentence prediction loss (NSP) [3].

For each model, we used the corresponding tokenizer to tokenize the data before sending them to the model. The OpenAI GPT Double Heads model used the OpenAI GPT Tokenizer, while the GPT-2 Double Heads model used the GPT-2 Tokenizer. We trained each model for 3 epochs and 6 epochs, giving us 4 fine-tuned model checkpoints to evaluate.

III. METHODS

To train the models, we first get the data from PERSONA CHAT. We then preprocess the data by adding our special tokens to the dialogue set. We

tokenize the data and create an embedding layer before sending that to the models to train. Once the model has finished training, we pass it to the decoder, and the dialogue agent is ready to be used.

A. Embeddings

The embeddings are composed of three different embeddings, the word embeddings, the segment embeddings, and the positional embeddings. The word embeddings are just standard word embeddings, but with the inclusion of the special tokens. Because the model has not been pre trained on those tokens, we need to add them to the embeddings. The position embeddings indicate where each token was located within the text. The segment embeddings indicate in which segment the token was found in. The three embedding layers are created and then concatenated to be sent to each model.

B. Fine Tuning on PERSONA CHAT

We trained for 6 epochs on the PERSONA CHAT dataset using 131,438 training examples with a mini batch size of 4. We set the remembered history to be the last 2 dialogues that occurred. We set the gradient accumulation steps to be 8, the learning rate to be $6.25e-5$. We used a learning model coefficient and a MC coefficient of 1. We set the max normalization value to 1 and used 1 personality permutation while training.

C. Decoder

We implemented a decoder that will build full sequences from the next token predictions of our model. The two most common decoders for language generation are greedy-decoding and beam-search. With greedy-decoding, the algorithm selects the most likely next token according to the model until we reach end-of-sequence tokens [7][8]. However, one risk with greedy decoding is that a highly probable token may be hiding after a low-probability token. Beam search mitigates this

issue by maintaining a beam of several possible sequences that we construct word-by-word [7][8].

However, recently, there have been at least two influential papers published where greedy/beam-search decoding was shown to be less effective and replaced by sampling from the next token distribution at each time step. These papers used a variant of sampling called top-k sampling in which the decoder samples only from the top-k most probable tokens (k is a hyper-parameter). In the study published by Ari Holtzman et al. distributions of words in the text generated by Greedy decoding and beam-search were also shown to be less human-like and top-k sampling [6].

We decided to use the two most promising candidates to choose as our decoding algorithm which are top-k and nucleus (or top-p) sampling to generate text sequences. The general principle of these two methods is to sample from the next-token distribution after having filtered this distribution to keep only the top k tokens or the top tokens (top-k) with a cumulative probability just above a threshold (nucleus/top-p) [6].

IV. EVALUATION

To test our models, we referenced the evaluation schema and script used by the ConvAI 2018 competition to evaluate our models. The ConvAI competition used both automatic metrics and human judgment to assess the performance of the conversational agents developed by participants. Automated metrics include F1, perplexity and Hits@1/20 [8]. The human evaluation schema consisted of ‘Live’ chat using 10 volunteers to chat with our fine-tuned models. The framework comprises several metrics that measure different aspects of the agent’s performance in open-domain conversation, such as the relevance, coherence, and engagingness of the agent’s responses. Volunteers were instructed to chat with our model through at least 10 turns and asked to score criteria with a

score between 0 and 5, with 5 being the ‘best’ and highest score for each evaluation criteria. We assigned a final score to each evaluation criteria by averaging the scores from 5 volunteers. We used the hidden test set provided by ConvAI to evaluate our agent based on automated and human metrics. We compared results to our baseline, an LSTM-based attentive Seq2Seq model (downloaded from keras) trained and tested on ConvAI’s hidden test set [8].

A. Automated evaluation metrics

- *Perplexity*: a metric of text fluency which is computed as $\frac{1}{m} \sum_{i=1}^m \log p(w_i)$ for sentence $w = w_1, w_2, \dots, w_m$ [8].
- *F1-score*: In the context of dialogue, precision is the fraction of words in the predicted response of the model that are contained in the gold response. Recall is the fraction of words in the gold response that were in the predicted response [8].
- *Hits@1/20*: hits@1/N is the accuracy of the next dialogue utterance when choosing between the gold response and N-1 distractor responses (here, N=19). Distractor responses are random responses taken from the hidden dataset [8].

B. Human evaluation metrics

- *Fluency*: This category evaluated the conversational agents’ ability to produce grammatically correct and natural-sounding responses [8].
- *Coherence*: This category evaluated the conversational agents’ ability to maintain a coherent conversation that followed a logical flow. The evaluators assessed whether the agents’ responses were relevant to the previous turns in the conversation and whether the agents’ responses showed an understanding of the topic being discussed [8].

- *Relevance*: This category evaluated the conversational agents' ability to provide relevant and informative responses. The evaluators assessed whether the agents' responses were factually correct and whether they provided helpful information to the user [8].
- *Engagement*: This category evaluated the conversational agents' ability to engage the user and maintain their interest throughout the conversation. The evaluators assessed whether the agents' responses were interesting, entertaining, or emotionally engaging [8].

V. RESULTS

Our results were as follows:

TABLE 1: AUTOMATED METRICS RESULTS

Model	Hits@1	Perplexity	F1
OpenAI GPT	80	20.25	17.00
GPT2	79	16.00	17.76
Seq2Seq + Attention Baseline	29.8	12.6	16.18

Table data from [5]

Table 1. Above are the average automated metrics for the models using Seq2Seq + Attention as our baseline. The baseline results were defined in the ConvAI 2018 competition results [8].

TABLE 2: HUMAN EVALUATION METRICS RESULTS

Model	Fluency (0-5)	Coherence (0-5)	Relevance (0-5)	Engagement (0-5)
OpenAI GPT	5	3.75	4	4.25
GPT2	5	4.7	4.25	4.25

Table data from [5]

Table 2. Above are the average human evaluation metrics for the models on a scale of 0 to 5, where 0 is the worst and 5 is the best.

A. Analysis

From table 1, we can see that OpenAI's models GPT and GPT2 significantly outperforms our baseline for Hits@1, slightly outperforms our baseline on F1 but underperforms in perplexity. However, while automated metrics are useful for evaluating generative dialogue models on a common, quantifiable criteria, we feel that automated metrics like hits@1/20, f1 and perplexity may be the best metrics to capture the full range of skills and abilities needed for success in open-domain conversations. This is why we decided to conduct human evaluations that focus on metrics like relevance, coherence and engagement for the following reasons

- F1 score only measures the accuracy of the bot's responses based on a set of predefined responses for a specific task, and may not capture the quality of the bot's responses in generating new and engaging conversation.
- Perplexity is a measure of how well a language model can predict a sequence of words. While it can be useful in evaluating how well a dialogue agent can predict the next sequence of words/sentences/tokens, it may not capture the complexity and richness of human language, which can involve multiple layers of meaning, metaphorical expressions and cultural references. For example, an agent with a low perplexity score may not necessarily be engaging in a natural, human-like way in the context of an open domain conversation
- Hits@1/20 only evaluates whether the bot's response is identical to the human's/gold response, which means that it may not capture the diversity of possible responses that the bot could generate. In open-domain conversations, there may be multiple valid responses to a given prompt. Therefore a bot

may still obtain a low Hits@1/20 score even though it is able to generate a diverse range of responses and is more engaging to users.

For our human evaluation, we can see from Table 2 that the GPT-2 model outperforms OpenAI GPT model in coherence and relevance and is on par in terms of fluency and engagement.

Furthermore, we recorded conversations and gathered additional feedback from human evaluators on their experience with conversing with our dialogue agents. One positive feedback is that our dialogue agents are very fluent, which means that they are always grammatically correct during conversations. However, one recurring feedback is that agents are only able to generate surface-level conversations because they struggle with generating specific details about their personalities. For instance, in Diagram 1, when the agent was asked ‘Why do you work in the Navy’, the dialogue agent was only able to respond with a vague reply, ‘I love the navy. I love reading’. A more realistic - human like answer would be more detailed and coherent such as ‘I love working for the Navy because I love serving my country’. Another recurring criticism was that our dialogue agents were only capable of giving short responses, which is very different from a human’s capacity to vary the length of their responses as the conversation goes. For instance, in both diagrams 1 and 2, we can see that the dialogue agent is only able to generate utterances of length of 5 - 10 words. This can make the conversation feel very abrupt and disjointed because friendly chit-chat requires a delicate balance of question and answering.

FIGURE 3

Selected personality: I work for the navy. I am blue and tall. I like to read. My favorite show is the voice.

>>> *How are you today? [human]*

I'm good. just got off work. [agent]
 >>> *What work did you do? [human]*
I work in the navy. [agent]
 >>> *Why do you work in the Navy? [human]*
I love the navy. I love reading. [agent]
 >>> *what are you doing for the rest of the day? [human]*
I am reading a book about the voice. [agent]
 >>> *Nice! [human]*
do you have a favorite show? [agent]
 >>> *Yes, I do. I love the show 'How I Met Your Mother' [human]*
me too. I love reading too. [agent]
 >>> *what do you like to read? [human]*
I love the voices. I read them on tv. [agent]
 >>> *what is that? [human]*
they're my favorite show. [agent]
 >>> *what is the show about? [human]*
it is about the voices. [agent]
 >>> *what is the voices? [human]*
they're like a show about the navy. [agent]
 >>> *Why do you like it? [human]*
I like reading and they're the best. [agent]
 >>> *Nice to talk to you! [human]*
I'm a navy brat! [agent]

Fig. 3. Example dialogue from our human evaluations receiving an average score of 5 on Fluency, Engagement, Coherence and Relevance

FIGURE 4

Selected personality: I like to drink. I live alone. I shop online, too much. I don't believe in love.

>>> *Hi, how are you today? [human]*
I am well, thanks for asking. how are you? [agent]
 >>> *I am well [human]*

that is good. what do you do for fun?
 >>> I like to crochet. How about you? [human]
 i like to shop online. i also love to shop online. [agent]
 >>> why do you like to shop online? [human]
 i love to go to the bookstore [agent]
 >>> What do you do at the bookstore? [human]
 i buy stuff online. i buy things online. [agent]
 >>> Ok. [human]
 i need to go to the grocery store. [agent]
 >>> Nice! What are you going to buy at the grocery store? [human]
 i am going to buy some alcohol. [agent]
 >>> What kind of alchohol are you going to buy? [human]
 i like filet mignon, it is really good [agent]

Fig. 4. Example dialogue receiving an average score of 5 on fluency, 3 engagement, and relevance and 2 on coherence

B. Conclusion

Our results show that it is possible to train a dialogue agent to exhibit certain aspects of personality and identity in its responses using a transfer learning approach.

While our dialogue agents can speak reasonably fluently, they fail to utter anything other than short, surface level conversations, which is why they still lack the human-like quality to generate deep and detailed responses as well as the ability to vary the length of utterances as the conversation progresses. Future work in this direction may include enriching our current data set with more detailed descriptions of each persona that go beyond the surface-level (e.g. including adding personal stories and anecdotes into our training data so that when asked questions like ‘why do you want to be a runner?’

our dialogue agents may generate detailed answers like ‘I would like to be runner because I look up to my dad and he was a runner as well’ instead of ‘I like to run’).

ACKNOWLEDGMENTS

We would like to thank Dr. Korpusik for her knowledge throughout the duration of the semester.

REFERENCES AND RELATED WORKS

- [1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186). Association for Computational Linguistics.
- [2] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding with Unsupervised Learning. Retrieved from https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
- [3] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. Retrieved from <https://d4mucfpgksyww.cloudfront.net/better-language-models/language-models.pdf>

- [4] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too?. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 2204-2213). Association for Computational Linguistics.
- [5] Hierarchical Neural Story Generation by Angela Fan, Mike Lewis, Yann Dauphin
(<https://arxiv.org/abs/1805.04833>)
- [6] *The Curious Case of Neural Text Degeneration* by Ari Holtzman, Jan Buys, Maxwell Forbes, Yejin Choi (<https://arxiv.org/abs/1904.09751>)
- [7] Breaking the Beam Search Curse: A Study of (Re-)Scoring Methods and Stopping Criteria for Neural Machine Translation by Yilin Yang, Liang Huang, Mingbo Ma
(<https://arxiv.org/abs/1808.09582>)
- [8] Dinan, E., Logacheva, V., Malykh, V., Miller, A., Shusterer, A., Urbanek, J., Kiela, D., Szlam, A., Serban, I. V., Lowe, R., et al. (2019). The second conversational intelligence challenge (ConvAI2). In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3601-3610).