Conversation Beyond *Yada Yada*: Deep Transformer Models, Persona Embeddings, and Transfer Learning for Sitcom Dialogue Response Generation

Amanda Kuznecov
anr431@nyu.edu

Sophia Tsilerides
smt570@nyu.edu

Ilana Weinstein
igw212@nyu.edu

Kevin Wilson
ksw366@nyu.edu

Abstract

Conversational modeling is an important task in natural language processing (NLP) and machine learning. A successful model of this kind can have many applications, such as allowing people to speak to their favorite characters, creating more life-like AI assistants, or creating virtual alter-egos of ourselves (Nguyen and Morales, 2017). Previously, machine-generated responses sensitive to the context of conversation were trained using persona-based datasets. In this paper, we experiment with fine tuning a pretrained transformer model on individual characters from the T.V. sitcom *Seinfeld*[1] separately, and have dialogue agents converse with each other. This architecture is fully data-driven and can be trained end-to-end without personas. Our results show that generated responses still incorporate aspects of personality and identity.

1 Introduction

Comedy is all about context. Attempting to generate a conversation between two characters in a sitcom directly relates to a common objective of NLP: using language models to emulate human cognitive ability. This is a challenging task for a computer that programmatically has no character, but through NLP can contribute to conversations using an encoder-decoder setup. Encoding personas into character embeddings can even generate realistic dialogue from a specific character or personality.

The challenge with persona embeddings is that they are created manually, and most datasets are not formatted to be automatically accepted by a model that requires them. In selecting the manuscript data from the television sitcom *Seinfeld*, significant data preparation was required to create persona and character embeddings for encoding. Following suboptimal results, we discovered DialoGPT, a large-scale pretrained dialogue response generation model for multiturn conversations (Zhang et al., 2020), and fine-tuned it for our task. After including personas in this architecture, we hypothesized that this model can be trained end-to-end *without* any personas and still effectively produce identity-distinct and context-sensitive responses.

2 Related Work

Encoding personas in distributed embeddings to capture speaker characteristics was our initial inspiration for this project and derived from *A Persona-Based Neural Conversation Model* (Li et al., 2016b). Attempting to generate a conversation between two characters in a sitcom directly relates to a common issue that current neural network models have regarding the generation of conversational responses: the problem of genericness (Li et al., 2016a) (Zhang et al., 2018). By encoding personas into character embeddings (Li et al., 2016a), we were able to generate realistic, character-specific dialogue. We were assured that by encoding interactions between two characters, we would be able to generate conversations between characters from completely unsupervised training data (Sordoni et al., 2015). Personalized dialogue agents were used to approach this by storing personas in a memory-augmented neural network (Zhang et al., 2018). Lastly, we were inspired to seek the use of additional datasets to improve our baseline model because of the availability of large multilingual parallel corpora with tools and interfaces for this task (Tiedemann, 2009).

3 Methods

We present the data and experiments utilized to explore our hypothesis, that a model can be trained

[1] https://en.wikipedia.org/wiki/Seinfeld

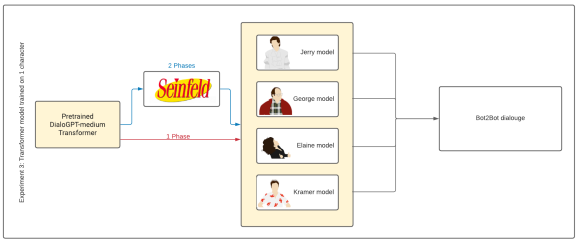

Figure 1: Overview of Experiment 3 of the project which was used to test our hypothesis.

end-to-end without any hand-crafted personas to produce character-specific responses.

3.1 Seinfeld Dataset

For this project, we utilized the *Seinfeld* teleplay corpus obtained from Kaggle.[2] The dataset contained 54,606 lines labeled by speaker and episode. This dataset was chosen because of the relatively small number of characters in the show and the main characters' distinctive personalities.

A framework similar to the PERSONA-CHAT dataset (Zhang et al., 2018) was adopted to incorporate character personas into the model. The original dataset was modified to follow an entire conversation between two characters. Each data instance contains the input of past conversation history between two speakers, as well as the persona of the following speaker. The label for each data instance is the spoken line by the following speaker. These personas were manually created to influence the content and style of responses given by a certain character. Each character was given a persona used for training, as well as a persona used for testing so that the testing persona was previously unseen by the model.

Since the original dataset did not contain any markers for scene changes, dialogue was judiciously separated into conversation by following the dialogue between two speakers until a new speaker appeared; at which time, a new conversation was declared, and the previous conversation history was not appended. Lines from the original dataset were extracted only if one of the main characters delivered the next line. During data processing, it was important to separate each episode so that conversation from the end of one episode did not carry over into the next. Additionally, several episodes began with a monologue delivered by Jerry; these were entirely removed.

[2] https://www.kaggle.com/thec03u5/seinfeld-chronicles

3.2 Experiments

To compare text generative models without encoded personas, we conducted three experiments: two incorporating manually crafted personas and one novel method, without personas. Our baseline was a simple SEQ2SEQ model with persona. Each experiment utilizes the *Seinfeld* dataset.

3.2.1 SEQ2SEQ

A sequence-to-sequence (SEQ2SEQ) framework was used to model dialogue, because of its ability to take one sequence and convert it into another sequence (Sutskever et al., 2014). To make the model output sound like a certain character, we incorporated persona embeddings for each character, with the hope that these embeddings would be able to encode information about the characters and their style of speech, in addition to generating personalized responses. Our model was developed to take as input the word embeddings of the previous conversation history between two characters, as well as the encoded persona of the respondent. Targets were the word embeddings of the spoken response. The model consists of two layers of encoder and decoder components which take the input sequence as a hidden state vector and output a response sequence. A GRU (Gated Recurrent Unit) RNN was used for this task as it is an enhancement to standard neural networks, due to its ability to retain information for a long time (Chung et al., 2014). This was necessary for generating good responses based on the context of the conversation. Other model parameters included hidden size of 512, embedding size of 256, and dropout of 0.3.

SEQ2SEQ modeling with attention and the encoder-decoder architecture was performed two ways. Task-specific modeling was performed as a baseline by passing in only the *Seinfeld* persona dataset. It was evident that with a small training set, it was not large enough to even model coherent sentences. Next, a phased modeling approach was conducted by first training the model on the PERSONA-CHAT and *Seinfeld* persona datasets, and then fine tuning using only the *Seinfeld* data. Model performance was clearly better because generated sentences were coherent and generally stayed on topic. However, they did not seem to fully embody any of the *Seinfeld* characters as most responses were generic.

Historically, the encoder-decoder architecture tends to produce generic responses that do not add anything meaningful to conversation. For future modeling, the objective function could be replaced with maximum mutual information, which would prevent favoring responses with the highest probability, and rather favor responses that are specific to the input (Li et al., 2016a). Although a GRU was used to combat the issue of encoder-decoders not retaining long conversation history, it seemed that a different model was required to produce meaningful results.

3.2.2 DIALOGPT with Personas

We determined that by identifying an existing pretrained model and fine-tuning for our task, we could generate more relevant, fluent, and context-aware dialogue. For this we chose DialoGPT, a large-scale pretrained response generation model released by Microsoft (Zhang et al., 2020). The model is built on the HuggingFace `transformers` library and based on OpenAI's state-of-the-art GPT-2 (Radford et al., 2019). Importantly, and distinctly from GPT-2, DialoGPT was trained on 147 million conversation-like exchanges (about 1.8 billion words) extracted from Reddit comment threads from 2005-2017. While sitcom dialogue differs from Reddit discussions in many ways, we expected that using this conversation-based pretrained model would effectively capture the casual, friendly nature of sitcom scenes better than text-generated from other commonly used state-of-the-art pretrained models such as BERT or GPT-2, which are trained on more formal corpora such as Wikipedia. (Devlin et al., 2019)

Like other transformer-based language models, DialoGPT uses a multi-layer self-attention mechanism to enable fully connected cross-attention to the full context. Having been trained on a large dataset, it effectively captures long-term contextual dependencies and produces text with high quality response relevance and diversity. For our task, we selected `DialoGPT-medium`, a smaller version of the full pretrained model with 345M parameters that would allow us to produce results and iterate more timely. We fine-tuned for 3 epochs using an AdamW optimizer with learning rate of 5e-5, epsilon of 1e-8, max sequence length of 128, and batch size of 1 (to avoid out-of-memory issues encountered when training on Google Colab GPUs).

Notably, DialoGPT was *not* trained with personas, but in our initial experiment with this method we thought it necessary to incorporate character-specific information in training and generation in order to capture character-specific personalities, emotions, and identities. As such, when fine-tuning the model, we appended our manually crafted personas for each target character in the conversation history as context along with conversation cues. The same personas were also provided to the model when generating dialogue.

3.2.3 DIALOGPT Character-Specific Models

Since DialoGPT was not initially trained with persona embeddings, we found the resulting generated conversations of the above experiment frequently included words or phrases from the manually-crafted personas of each character. To overcome this, and as part of our principal hypothesis, we used the same DialoGPT pretrained model as before, but this time fine-tuned *separate* models for each target character {Jerry, George, Elaine, and Kramer}. It was therefore necessary to set apart the lines for each of these target characters, with the preceding same-conversation lines as context for training.

This architecture has many advantages, the most salient of which is that no manually crafted personas are necessary at either training or generation time. Our expectation with this approach was that the relatively large (for a television sitcom) dialogue corpus for these four target characters *incorporates everything fundamentally important about these individual characters' personalities, emotions, and identities*, and in fact accomplishes this better than with manually crafted personas, which are inherently biased and limited.

From this architecture, two characters are selected to converse and the output of one character's model becomes the input of the other for conversation generation. Refer to Appendix for examples.

There are a number of drawbacks to this approach, which we acknowledge. Firstly, it would be unlikely to obtain reasonable results on smaller corpora; realistic and context-consistent machine-generated dialogue produced by fine-tuning a large-scale pretrained model can only be expected with sufficient character-specific data (lines and associated conversation history) in order to effectively learn personality and identity information. In addition, this method requires character-specific fine-tuning, which is more computationally difficult than the persona-based methods discussed above. We found that training required approximately 90 minutes per character on Google Colab GPUs.

Model	**BLEU**	**Perplexity**
Seq2Seq+Persona	**0.56**	63.15
DialoGPT+Persona	0.54	186.55
DialoGPT+Character_Specific_Models	0.55	46.11
DialoGPT+Intermediate_Phase+Character_Specific_Models	0.54	**21.26**

Table 1: Results for baseline and different approaches to text generation for *Seinfeld* characters. BLEU is generated by the average BLEU for each character to character interaction in the test set. Perplexity is the overall perplexity on the test set.

3.2.4 DIALOGPT with Intermediate Phase

Lastly, we attempted to improve this procedure by incorporating an intermediate phase of fine-tuning: fine-tuning first on the entire *Seinfeld* corpus, agnostic of character, to capture the context of the entire show, improves the generalizability of the model, and reduces the character-specific fine-tuning time. After this intermediate phase is trained, we then use the same method as before to fine-tine individual character-specific models.

Several variations of this approach phase were attempted, with a goal of avoiding overfitting. We determined a single epoch of intermediate-layer fine-tuning was appropriate, and tested using 100%, 50%, 25% of the *Seinfeld* corpus in this phase. These variations produced validation perplexities of 36.12, 30.54 and 25.22, respectively. We additionally trained the 100% intermediate phase on 2 epochs which exploded perplexity to 113.02. Attempts to train longer were discarded to avoid over-fitting. With these results we chose to continue with our intermediate phase using 25% of the manuscripts: our analysis revealed that the number of lines for each core character ranged between 11-27% of the corpus; as such we found 25% performed the best. We believe that this strategy helped our model generalize and create a more robust conversational model for this task.

4 Evaluation and Analysis

To evaluate the discussed model architectures, we use BLEU and perplexity as they are commonly used to evaluate chat-bots and translation models.

As seen in Table 1, DialoGPT with one and two phases lead to performance gains when evaluating the test set on perplexity. However the persona agnostic models do not provide much improvement on BLEU score as our baseline SEQ2SEQ performs the best. This result is unexpected as the baseline output was less fluent than others.

To further understand the performance of the models we also calculated BLEU for each core character across models. As seen in Figure 6, Elaine and Kramer did best with our novel approaches whereas Jerry and George did not.

		Human Judgement	
		Real	Machine
True Class	Machine	41%	59%
	Real	70%	30%

Table 2: Survey Results

Chatbots are generally evaluated by humans since word-overlap metrics like BLEU turn out to correlate very poorly with human judgments and don't evaluate on the context of the scene nor the performance of the personality (Liu et al., 2016). Thus, we conducted a survey geared towards those familiar with the show to provide us with a better understanding of large conversation performance. There were 47 participants in the survey and they were asked to classify if an example scene is from the show or machine-generated (DIALOGPT Character-Specific). Table 2 shows that out of the machine questions, 41% of those respondents were not able to detect machine-generated dialogue, and out of the real conversations, 30% thought they were machine-generated. Some participants may have recognized scenes from the show, however a great deal of them couldn't discern the real from machine generated conversations.

5 Conclusion

We have explored various techniques for generating fluent, identity-aware, context-consistent sitcom dialogue, both with and without manually crafted personas. Our results demonstrate that a completely data-driven framework—fine-tuning a state-of-the-art pretrained transformer model for our task —effectively captures elements of personality and identity into sitcom dialogue generation without the need for persona embeddings.

References

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *NIPS 2014 Deep Learning and Representation Learning Workshop*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *Proceedings of the 2019 Conference of the North*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Huyen Nguyen and David Morales. 2017. A neural chatbot with personality.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

J. Tiedemann. 2009. News from opus — a collection of multilingual parallel corpora with tools and interfaces.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT: Large-scale generative pre-training for conversational response generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Appendix

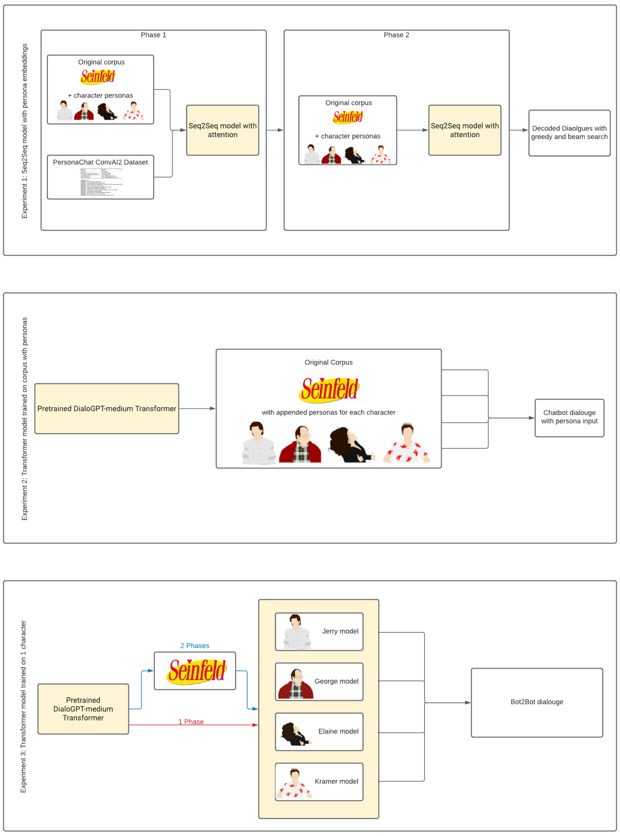

Figure 2: An overview of project experiments. Our first experiment was a seq2seq model trained on the Seinfeld teleplay corpus with added character personas. Our second experiment used a transformer fine tuned on the Seinfeld corpus with adding persona information. Lastly, we fine-tune 4 separate character-specific models.

	character	target line	conversation	conversation/0	conversation/1	conversation/2	conversation/3	conversation/4	conversation/5
539	ELAINE	Hmmm , I dont think so .	Doesnt matter . Whats the young mans name ? I would like to meet him .	No , we just met ;	No , thats great ! Thats terrific !	What ? No , its not a big deal .	Oh really ? Elaine Marie Benes ;	No . A general guy .	Uh - huh . Yeah . So , anybody specific ?
541	ELAINE	Wall street .	Well , what does he do ? Is he an artisan , a craftsman , a labourer of some sort ?	Hmmm , I dont think so .	Doesnt matter . Whats the young mans name ? I would like to meet him .	No , we just met ;	No , thats great ! Thats terrific !	What ? No , its not a big deal .	Oh really ? Elaine Marie Benes ;
543	ELAINE	And he happens to be pretty good lookin .	Ah , high finance . Bulls , Bears , people from Conneticut .	Wall street .	Well , what does he do ? Is he an artisan , a craftsman , a labourer of some sort ?	Hmmm , I dont think so .	Doesnt matter . Whats the young mans name ? I would like to meet him .	No , we just met ;	No , thats great ! Thats terrific !
545	ELAINE	And ; hes hilarious .	All right , sir .	And he happens to be pretty good lookin .	Ah , high finance . Bulls , Bears , people from Conneticut .	Wall street .	Well , what does he do ? Is he an artisan , a craftsman , a labourer of some sort ?	Hmmm , I dont think so .	Doesnt matter . Whats the young mans name ? I would like to meet him .
547	ELAINE	I staked out his health club .	Now thats not fair ! So where did you meet this guy ?	And ; hes hilarious .	All right , sir .	And he happens to be pretty good lookin .	Ah , high finance . Bulls , Bears , people from Conneticut .	Wall street .	Well , what does he do ? Is he an artisan , a craftsman , a labourer of some sort ?

Figure 3: Sample training data for Elaine character model with target line and conversation history as context.

```
JERRY: you know , kramer , i rented out half of my space to you .
KRAMER: yeah, you did.
JERRY: kramer, why did you give him my tv?
KRAMER: because he's a friend of yours.
JERRY: i'm not his friend.
KRAMER: all right, then i'm his friend!
```

```
KRAMER: george , that ' s what he has and i have to give him a call . now we gotta came clean .
GEORGE: i can't do it.
KRAMER: what can've he do?
GEORGE: what's he good for?
KRAMER: well, he's got that big new format. computers generate'em.
GEORGE: oh yeah.
```

```
ELAINE: i mean i lost my job , i can ' t go to africa . i was gonna meet the bush - men of the kalahari
GEORGE: but africa is such a vast place. how do you get all that wood in there?
ELAINE: i cut down a few trees.
GEORGE: you cut down some trees?
ELAINE: yes.
GEORGE: what did you do with the stump?
```

Figure 4: Three sample conversations generated by fine-tuned DialoGPT character-specific models ”talking to each other.”

SEQ2SEQ with Seinfeld persona dataset and PERSONA-CHAT	DIALOGPT with personas	DIALOGPT Character-Specific Models

Input:
• *George's persona:* `I am vulnerable and slightly neurotic. I am bald. I am short, stocky, and slow-witted. I am dishonest. I am a real estate broker.`
 	
Conversation History:
`JERRY: What? What is it?`
 	
Output:
`GEORGE: I don't know what to say`
	Input:
	• *Elaine's persona:* `I am intelligent. I am funny. I am assertive and confident. I am edgy and superficial.`
	
	Conversation History:
	`JERRY: I can't believe your a friend of mine.`
	
	Output:
	`ELAINE: That's a little too much.`
		Input:
		• JERRY Model
		• KRAMER Model
		
		Conversation History:
		`JERRY: You know, kramer, I rented out half of my space to you.`
		
		Output:
		`KRAMER: Yeah , you did .`
		`JERRY: Kramer, why did you give him my tv ?`
		`KRAMER: Because he ' s a friend of yours .`
		`JERRY: I'm not his friend.`
		`KRAMER: All right, then I'm his friend!`

Figure 5: Sample conversations generated by all experiments.

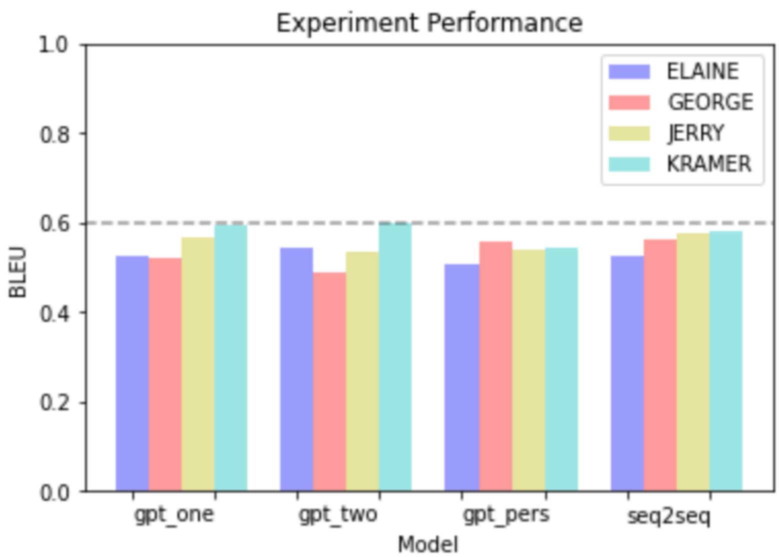

Figure 6: Model performance per character