Generating Stylistic and Personalized Dialogues for Virtual Agents in Narratives

Weilai Xu
Bournemouth University,
Faculty of Science and Technology
Poole, United Kingdom
wxu@bournemouth.ac.uk

Fred Charles
Bournemouth University,
Faculty of Science and Technology
Poole, United Kingdom
fcharles@bournemouth.ac.uk

Charlie Hargood
Bournemouth University,
Faculty of Science and Technology
Poole, United Kingdom
chargood@bournemouth.ac.uk

ABSTRACT

Virtual agents interact with each other through dialogues in various types of narratives (e.g. films). In this paper, we propose an approach on the basis of DialoGPT pre-trained language model, which explores the impact of dialogue generation with different levels of agents' personalities derived from narrative films based on the Big-Five model, as well as with three different embedding methods. From the experimental results using automatic metrics and human user evaluation, we investigate and analyze the impact of different settings on narrative dialogue generation. We demonstrate that our approach is able to generate dialogues with increased variety that correctly reflect the corresponding target personality.

KEYWORDS

Virtual Agents, Dialogue Generation, Narratives, Deep Learning

ACM Reference Format:
Weilai Xu, Fred Charles, and Charlie Hargood. 2023. Generating Stylistic and Personalized Dialogues for Virtual Agents in Narratives . In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 10 pages.

1 INTRODUCTION

In narratives, characters or virtual agents are accountable for executing narrative actions to progress the storyline as well as to convey author's intentions [4]. Dialogue is one of those actions that agents make use of in some popular types of narratives [24, 25], such as films and video games. With the development of computational narrative systems over the last two decades, and a renewed interest in AI techniques to generate dialogues in narrative context, automated dialogue generation offers a mean to dynamically communicate author-driven story elements. Over the last decade, research in computational narratives has explored several AI-based solutions, such as plan-based narrative structures [30, 33], whilst dialogue-driven narratives are still hindered by the limitations of dialogue generation [8, 9]. 1) Text realization is highly dependent on narrative plans, which means the variations of the narrative discourse are completed in the narrative planning stage. In this way, it is difficult to alter the discourse in text dynamically according to the changes of narrative elements. 2) The lexicalization in text realization stage is based on empirical ontology of templates and operators, which leads to limited semantic and syntactic level representation, and to the lack of story level styles for generated utterances.

Existing works incorporate stylistic information by using additional features, such as speaker profile [16, 49], sentiment or emotions [7, 10], and tense [11], as well as controlling the style of generated sentences by altering these features. Most research only consider local features, i.e. features pertaining to individual sentences affecting expression alteration in the scope of each individual sentence. However, it is necessary to use higher level knowledge for generating narrative-based dialogues, which represent the authorial intent and provide consistency over the story generated.

In this paper, we aim to investigate the potential of enriching agents' dialogues in narratives by incorporating their personalities using advanced AI techniques. Our main contributions can be summarized as follows:

(1) A well parsed, segmented, and labeled dataset from IMSDb[1], which contains dialogues in screenplays, along with the characters, scenes and proposed personality with two levels of granularity.
(2) An approach for generating conditional dialogues by utilizing the Big-Five model based personality traits from screenplays. Our approach based on three embedding methods can generate varied dialogues which are able to reflect the selected target personality traits (Figure 1).
(3) An experiment evaluating the impact of different levels of personality and embedding methods to dialogue generation.

In the first two sections, we present the current research background as well as previous and related works. In Section 3 and 4, we introduce the details of our dataset and approach. We report on the experimental results in Sections 5, along with the analysis and findings. Finally, we discuss our findings and conclusions.

2 RELATED WORKS

Virtual Agent in Narratives

Narrative theorists often represent a narrative as consisting of a series of events or actions done by agents [4, 37]. Therefore, the agents and their relationships with the story are continuously considered in narrative research. Porteous et al. [31] presented a novel notion of a virtual character's point of view to enable a story to be unfolded from the perspectives of different characters. Porteous et al. [32] used the relationship between characters to affect the progress of the narrative as part of an AI-based planning approach. Leong et al. [15] also used the changing relationships of agents through stories to construct the story arcs. Matthews et al.

[1]The Internet Movie Script Database

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

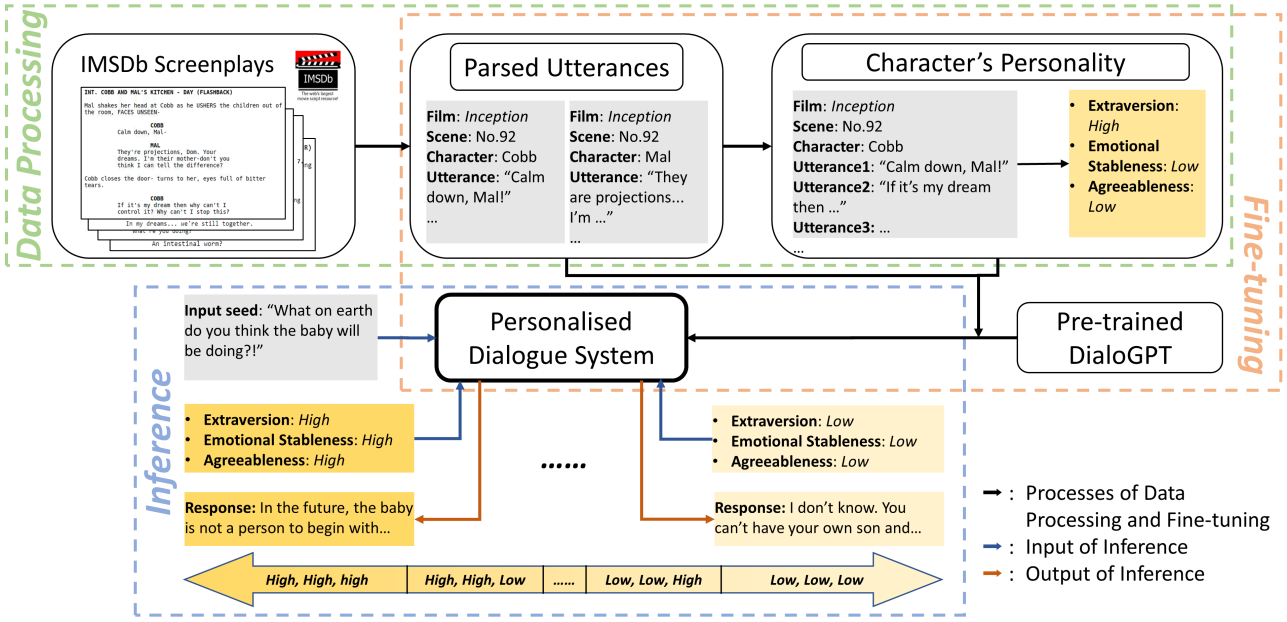

Figure 1: Our approach workflow, including stages of Data Processing, Fine-tuning, and Inference.

[22] introduced an approach called MISE-EN-SCÈNE region (MISER) supporting the dynamic staging of virtual characters' behaviors in narrative scenes.

These research explored various perspectives of agents and narratives, applying structure and event representations, though often not accounting for the discourse level within narratives. According to McKee [24], characters are assigned various characteristics based on their roles throughout a story. Also, Bednarek [1] and Kozloff [14] both agree that narrative dialogues must describe characters, including their personality and relationships, as well as reflect authorial information. Thus, in this paper, we intend to investigate the potential of reflecting virtual characters' personality on text realization as dialogues within the context of narratives.

Character-centered Dialogue Generation

The main factors of speakers behind the style variety can be divided into two categories, *personality* and *feelings*, which is based on whether the factors are individually not varying over time and transient occasionally [8]. For personality, speakers' persona (or speakers' profile, including gender, age, profession, etc.) is the most common feature to be modeled and embedded into dialogue systems whether implicitly [16] or explicitly [34, 40, 47, 49]. There are also several works [19, 21] leveraging the *Big-Five* model [13] to achieve language generation variation. These works systematically explore and analyze the correlations between nearly exhaustive linguistic features and 5 traits of the Big-Five model. Oraby et al. [27] and Xu et al. [44] demonstrated the ability to generate various dialogues with the Big-Five personality.

Other works explore influence of feeling, or emotion, or affects on dialogue generation. Huang et al. [12] trained an LSTM-based emotional classifier for 9 emotions, which were used to generate dialogues expressing corresponding emotions. Colombo et al. [5] use both categorical representation and continuous representation in a VAD space[2] [38] to model six basic emotions[3]. Also, Buechel et al. [2] introduce a methodology for creating almost arbitrarily large emotional lexicons for any target language.

These works either only consider sentence-level features derived from speakers, or focus on conversations and speakers in real life, where the potential of the narrative dialogues with authorised and global personalities of characters are yet to be explored.

Pre-trained Technology

Recent progress in pre-training methods has demonstrated promising results in many tasks [6, 29, 35], benefiting from attention mechanism and transformer structure. Among them, various approaches have been proposed for dialogue generation. Zhang et al. [48] adapted the pre-trained GPT-2 model [36] to DialoGPT for training and generating multi-turn dialogues. Yang et al. [45] adapted DialoGPT using word-level and sentence-level style language model for generating dialogues with ArXiv or Holmes style , i.e. using datasets with text from ArXiv papers and Holmes novels respectively. Shen and Welch [39] proposed a counseling dialogue system based GPT-2 which can generate sample counselor's reflections using dialogue history. Wang et al. [42] built a persona-based chatbot by fine-tuning DialoGPT and Roberta [18]. Zhong et al. [50] proposed a bert-based [6] response selection model with persona to improve empathetic conversations.

[2]Valence, Arousal, and Dominance.
[3]Anger, disgust, fear, joy, sadness, and surprise.

Table 1: Characteristics of the selected and parsed IMSDb dataset used in this paper.

Genre	Dialogue	Turn	Character
Action	103k	317k	7.1k
Drama	245k	785k	14.3k
Romance	81k	263k	4k
Thriller	134k	423k	8.5k

3 DATASET

3.1 Processing

The dialogues are taken from IMSDb similarly to Winer and Young [43] where they demonstrated that rich narrative knowledge can be extracted from screenplays. We create a heuristic-based processing of screenplays to recognize and segment a series of elements, including dialogue, character, transition, staging direction. Consequently, each screenplay is parsed into dialogue sessions based on the transitions stated within the screenplays, along with corresponding character (speaker) for each utterance, and categorized by genre. More specifically, each resulting large dialogue session is split into several dialogues for training, whilst ensuring each dialogue is always taking place between two characters. In this paper, we use dialogues composed of between 2 to 6 turns. All parsed dialogues are generated from 4 major narrative-rich genres: Drama, Romance, Action, and Thriller, which are split into training set and evaluating set with a ratio of 90% and 10% respectively (Table 1 shows the dataset's statistics).

3.2 Personality Definition

According to narrative theories [24], the personality of characters in a single narrative is set by the author and should remain consistent throughout. Therefore, we hypothesize that the personality of a certain character in a film remains consistent. We use the properties of the Big-Five model [13] (a.k.a. the "OCEAN" model, a widely acknowledged alias of the Big-Five model) to define characters' personalities. We specifically select three primary traits (**Extra**version, **Emoti**onal stability, and **Agree**ableness, abbreviated as Extra, Emoti, and Agree) according to the results of principal component analysis (PCA) of the training datasets. Here, we keep

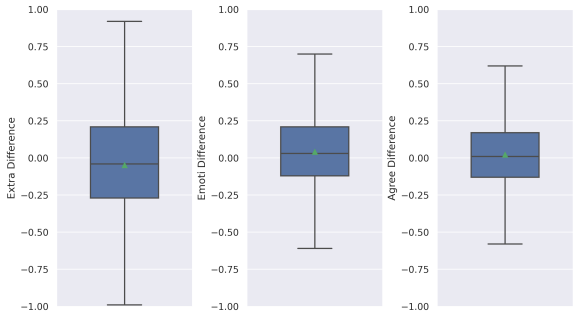

Figure 2: Character personality difference between film-level personality and average scene-level personality.

Table 2: Character Personality Matching Rate between Film and Scene Average.

	3/3	2/3	1/3	0/3
Prop.	37.43%	41.6%	17.92%	3.05%

the name of each trait following [13, 21][4]. Each utterance in the dialogue session is labeled with *High*, *Medium*, or *Low* for each of the 3 traits according to their character's personality score using the personality recognizer from [21]. This score is here calculated from all the utterances spoken by a single character for a complete screenplay, representing a *film-level overall* personality score for this character. This score is defined in the range 1 to 7, which is then divided into 3 sub-ranges: *Low* in the range lower than 3.8, *Medium* in 3.8 - 4.2, and *High* in the range greater than 4.2. For instance, for the extraversion trait, the label *Low* denotes more introvert and the label *High* denotes more extravert.

The personality score for a given screenplay changes over the course of the dialogues progressing, quite significantly sometimes. Thus, we calculated personality scores and labeled work at the more granular scene level (i.e. to calculate the trait scores from all the utterances spoken by a single character within each scene) as a finer-grained *scene-level* personality. To compare differences between film-level personality and scene-level personality of each character, we first calculated a weighted normalized average of each character's scene-level score depending on the word count of utterances for each scene and label characters' personalities with these scores. Then, we compared the matching rate on the trait labels between film-level personalities and scene-level personality average. Table 2 shows that there are 79% of characters who have at least 2 out 3 (2/3) trait labels matching. For each trait, there are over 50% of characters whose difference between average scene-level personality and film-level personality is less than ±0.25 (Figure 2).

4 APPROACH

4.1 Problem Formalization

Our aim is to generate an utterance response that corresponds to a given dialogue context and a representation of composite target personality traits.

We trained our model based on GPT-2 [35] and DialoGPT [48] architectures, which are both adopted from the generic transformer language model [41]. GPT-2 uses a series of masked multi-head self-attention layers to train on a huge amount of web data and is able to be fine-tuned for multiple downstream NLP tasks. Following GPT-2, DialoGPT models a multi-turn dialogue session as a long text and frames the generation task as language modeling on an enormous dataset collected from Reddit. The performance of these language models has demonstrated that self-attention based transformer language model has capacity to illustrate the distributions of natural language.

Therefore, we first use the standard language model as our backbone model. Here, we denote all dialogue turns (utterances) in a

[4]Particularly, the trait "Emotional Stableness" in the Big-Five model refers to trait "Neuroticism" in OCEAN model. While "Extraversion" and "Agreeableness" are labeled the same.

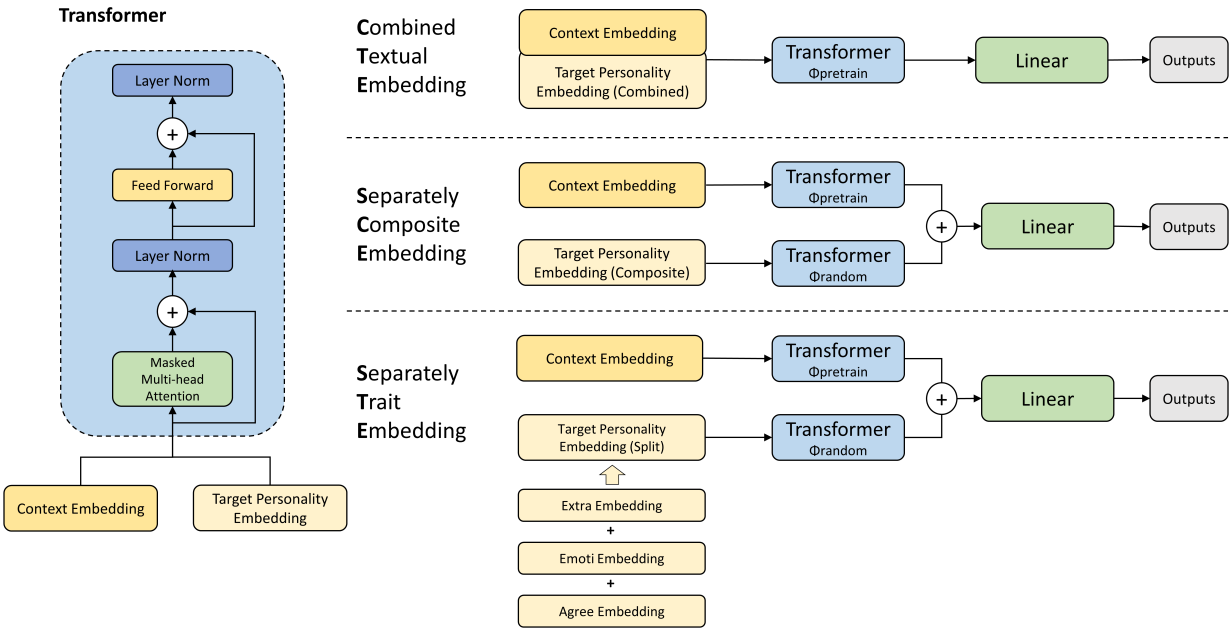

Figure 3: Our framework. The basic Transformer inherited from prior works on the *left*. The 3 methods for embedding characters' personalities are presented on the *right*. Each transformer has the same configuration and uses initialized parameters from pre-trained DialoGPT or randomization as noted.

session $Dialogue = \{T_1, ..., T_{i-1}, T_i, ..., T_n\}$, where the first $i-1$ turns are set as context(C), and the next turn T_i as the response(R). So the conditional probability of $P(R|C)$ is the product of a series of conditional probabilities:

$$P(R|C) = \prod_{i=2}^{n} P(T_i|T_1, ..., T_{i-1}) \tag{1}$$

As prior work about conditional dialogue generation [5, 12, 45, 49] introduced, in our work, we extend the standard language model by incorporating our target personality Psn, which specifically leads to $P(R|C')$, where $C' = \{C, Psn\}$. And we set our personality-based objective using the negative log-likelihood loss following DialoGPT:

$$L_{NLL} = -logP(R|C') \tag{2}$$

4.2 Personality Incorporation

We incorporate target personality Psn into our dialogue system using three different methods.

First, use a naive method which is to treat the target personality as another turn of dialogue in text, referred to *combined textual embedding (CTE)*. For this representation, we explicitly set the personality with the labels of the three specific traits in the following order *Extra, Emoti, Agree*. Then, the personality is concatenated with context as the input sequence, which is then fed in a transformer initialized with pre-trained parameters (Figure 3).

The second method is to embed context and target personality separately, referring to *separately composite embedding (SCE)*. For this representation, we treat each personality with 3 traits as a composite, and label all personalities as $Psn_i (i = 0, 1, ..., 27(3^3))$. Instead of feeding context embedding and target personality embedding independently into on single transformer as Zheng et al. [49], we use these two embeddings as input into two same transformers respectively like Mazare et al. [23], but with different initial parameters as Figure 3 shows.

Final and third method is to represent target personality from a finer-grained aspect, which sets an embedding for every single trait, then builds the target personality embedding that is a sum of three trait embeddings following Zheng et al. [49], named as *separately trait embedding (STE)*. The neural network modules are set in the same configuration as in the second method (Figure 3).

5 EXPERIMENT

5.1 Experiment details

Three different scales of pre-trained DialoGPT models are provided with total parameters of 117M, 345M, and 762M respectively. Here we use the 117M configuration to fine-tune and evaluate our approach. We conduct fine-tuning process from the original DialoGPT on our datasets for 2 epochs following parameters setup[5] with manually optimized hyperparameter values, and select the models with the lowest values of evaluating loss and perplexity during fine-tuning.

[5]https://github.com/microsoft/DialoGPT

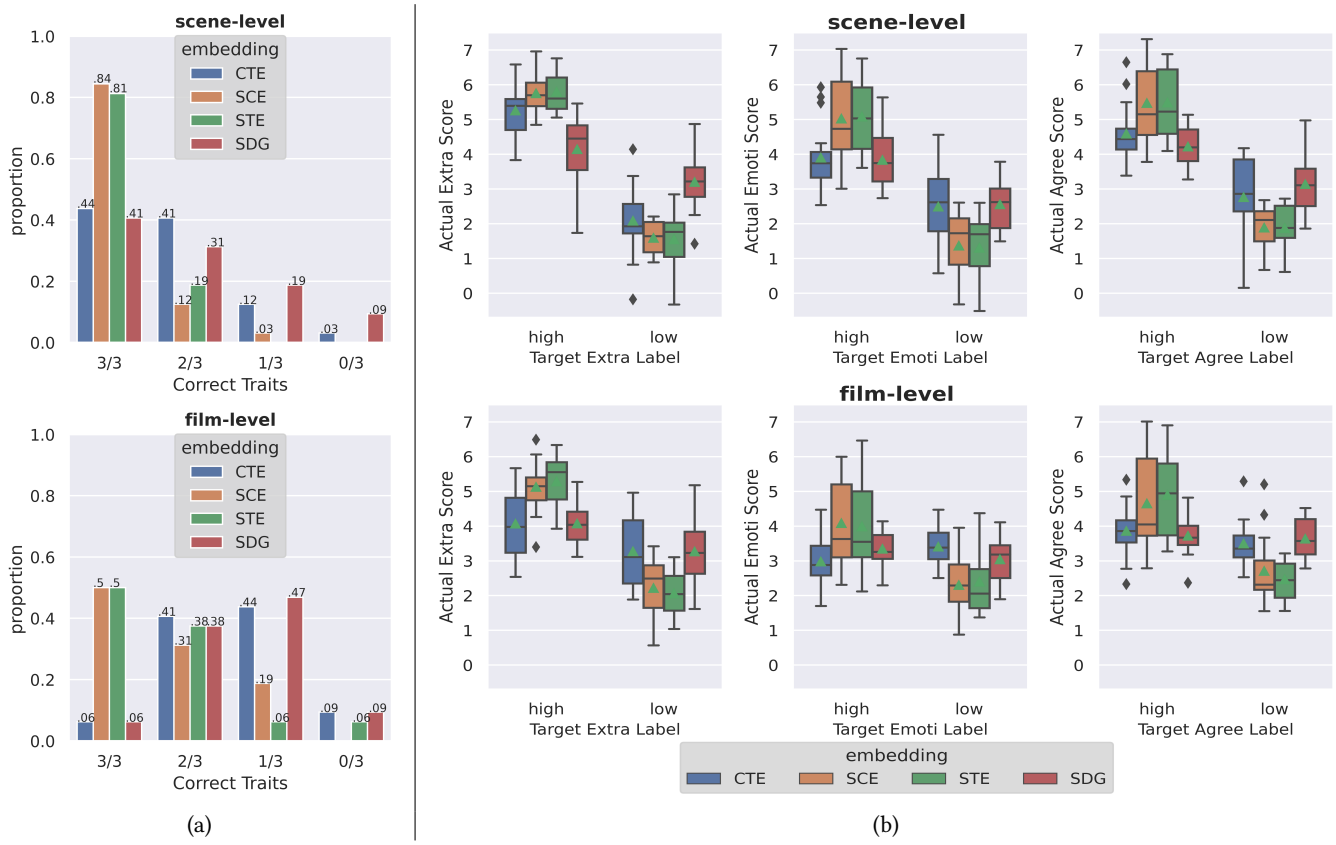

Figure 4: (a) Results of personality identification for generated dialogues on the overall aspect with trait matching accuracy, for both film and scene levels. (b) Trait aspects with identified scores. For each aspect, the first row shows the results with scene-level personality, and the second row shows the results with film-level personality. All results are grouped by embedding methods.

We conduct experiments to evaluate our proposed models, from the perspective of personality identification and variety. We selected 50 utterances from various genres of films which are not included in our dataset as input seeds for the generation process. Based on each seed, 3 more successive utterances were generated with 1 of 8 target personalities (to simplify the process of evaluation, we set the target personality with *High* and *Low* labels for each trait, removing *Medium*, i.e. 8 combinations of target personalities in total) to generate a 4-turn short dialogue. This process of generation is repeated 15 times for each seed across all 8 target personalities using 4 different embedding methods (*CTE*, *SCE*, and *STE* introduced in the previous section plus *SDG* for **S**tyle**DG**PT [45]). Therefore, 6,000 ($15 \times 50 \times 8$) sets of dialogues are generated for each personality embedding method, per level of personality and per genre. We present a selection of generated samples by personality combinations in Section 5.5.

5.2 Personality Identification

We first evaluate whether the generated dialogues reflect the "correct" personality. All the sentences generated with the same target personality are assembled and evaluated by the same tool as per for labeling.

5.2.1 Trait Matching Accuracy. To evaluate the extent of the model's ability to generate dialogues with the matching personality, we use the same scale to label the calculated personality for generated dialogues and then compare these labels with the given target personality labels, trait by trait. For example, in Figure 4 (a), "3/3" denotes all 3 trait labels match between the target personality and the identified personality from generated dialogues, while "0/3" denotes that none of them match.

In Figure 4 (a), we show an overall personality identification accuracy with two levels of personality and three embedding methods. From the perspective of scene-level personality, we notice that using both *SCE* and *STE*, over 80% of target personalities across 4 genres can be correctly identified, with all three correct identified traits (3/3). And almost all target personalities with at least two traits are correctly identified. The identification accuracy of *SCE* and *STE* is significantly higher than *CTE* and *SDG*, which has more than 40% personalities with all three traits correct and at least two traits correct. From the perspective of film-level personality, we can

Table 3: Statistics of generated dialogues grouped by personality level and embedding method. (↑ denotes the expectation of greater numbers, and ↓ that of lower numbers.)

Personality Level	Embedding Method	Sent Count	Word Count	Word-Sent Ratio	Bleu 12 (dialogue)↑	Bleu 12 (utterance)↑	Edit Distance↑	Semantic Similarity↓
film	CTE	1.391	10.695	7.755	5.596	2.386	0.764	0.209
	SCE	1.354	10.130	7.515	5.40	2.317	0.764	0.207
	STE	1.343	9.910	7.395	5.303	2.227	0.766	0.204
	SDG	1.299	8.498	6.622	5.161	2.121	0.768	0.197
scene	CTE	1.371	10.430	7.672	5.310	2.287	0.764	0.204
	SCE	1.359	10.131	7.469	5.125	2.191	0.764	0.205
	STE	1.336	9.623	7.185	4.932	2.074	0.766	0.200
	SDG	1.291	8.493	6.665	5.122	2.075	0.768	0.193
original DialoGPT (117M)		1.275	12.346	10.021	4.988	2.154	0.753	0.236
written screenplay		1.730	13.202	7.504	N/A	N/A	N/A	N/A

also obtain the similar observation that *SCE* and *STE* have better performance than *CTE* and *SDG*.

Comparing the performance between scene-level and film-level personality (Figure 4 (a)), it is easy to be aware that scene-level personality contributes more positive impact on personality identification rather than film-level one, where more personalities with all three traits can be correctly identified.

5.2.2 Trait Comparison. Figure 4 (b) shows the scores of identified personalities on a finer-grained trait aspect. With the target labels for each trait, the differences in the scores between the *high* and *low* are the focuses of interest, where a more significant difference means the dialogues can be identified more correctly, i.e. the identified scores with *high* label are expected to be higher than the ones with *low* label.

From trait perspective, we observe that the trait extroversion is the most correctly identified trait among all three traits with the great difference of score distribution between two labels and a small box (smaller box and shorter whisker denote the trait can be identified more steadily), especially with scene-level personality. While the other two traits have similar score distributions.

5.3 Analysis on Variety

Our findings for the generated dialogues are presented in Table 3.

For each generated turn of 4-turn generated dialogue (excluding the first turn, i.e. the seed), we count the number of sentences and the number of words, as well as calculate the word-sentence ratio. We compare the results of our approach with different settings, the generated dialogues from the original DialoGPT, and randomly collected dialogues from our dataset (written screenplays). We observe the sentence counts of our approach as well as *SDG* are more than dialogues from the original DialoGPT, but much less than the written screenplay. And the word counts per turn are around 10 words, which is less than DialoGPT and written screenplay. This observation probably indicates that attempting to correctly control the target personality of dialogue tends to cause losing some of the abilities of the generation.

We also evaluate the variety of dialogues generated by our approach by calculating the edit distance and semantic similarity. More precisely, for all generated dialogues with same seed and same target personality (For DialoGPT, only the seed controlled), we calculate the edit distance (normalized by text length with range 0 to 1) using Levenshiten distance, and semantic cosine similarity (with range -1 to 1) using Universal Sentence Encoder [3] for these dialogue pair-wisely. We observe that the dialogues generated with personality control are able to provide higher edit distance, as well as lower semantic similarity. This observation indicate that adding personality is able to generate dialogues with more variety given the same seeds from surface text perspective and semantic perspective.

Furthermore, we use Bleu [28] (n=1,2) as a representative of word-overlap metrics, to evaluate generated dialogues against the sole reference from original screenplay. We apply Bleu metric on dialogue level and utterance level evaluations. As the results shown in Table 3, from personality level aspect, the Bleu scores for film-level personality are higher than the ones for scene-level personality. Also from embedding method aspect, *CTE* reaches highest Bleu scores in either personality level. Normally higher Bleu scores indicate that the generated samples and the reference(s) share more overlapped n-gram words. However, in our case, we noticed that the settings achieve higher Bleu scores (film-level personality, *CTE*) have lower personality identification accuracy (see Figure 4 (a)). This observation is understandable, because generations with different personality combinations are more similar to the sole reference also denotes they are more similar to each other. Therefore, the difference of personalities is less likely to be reflected accordingly. Considering this analysis, we argue that such word-overlap based metrics are less applicable in our case. We also consider that for open-domain creative content generation, to chase higher scores by such metrics would to some extents frame the diversity of generations [17], especially in the case where there is a lack of golden references.

5.4 Human Evaluation

We conducted a human evaluation to score dialogue quality and personality identification accuracy. We recruited native English participants ($N = 13$). We selected two input sentences per embedding method per personality level and per personality combination (to simplify the process, two extreme combinations were selected in evaluation) for generation with personality, thus $2 \times 4 \times 2 \times 2 = 32$

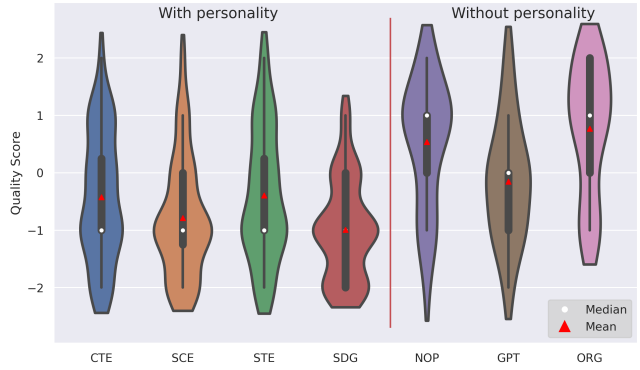

Figure 5: The score distribution of dialogue quality with different settings. They are divided into two categories: generations with (left) or without personality (right).

Table 4: 1-tail T-test results for personality identification evaluation by human. For each embedding method, the T-test results and p-values are calculated with both scene-level and film-level personality. Each group contains 52 ($13 \times 2 \times 2$) scores. The digits in bold denote significance (<0.05).

embed method	setting #1 (nos.) vs #2 (nos.)	setting #1 mean (std)	setting #2 mean (std)	pairwise t-stats	pairwise p-value
CTE	extra (52) vs intro (52)	0.48(0.85)	0.39(0.89)	0.56	0.287
	emoti (52) vs neuro (52)	-0.35(1.14)	-0.69(0.88)	1.74	**0.042**
	agree (52) vs disag (52)	-0.25(1.08)	-0.73(0.97)	2.48	**0.007**
SCE	extra (52) vs intro (52)	0.5(0.87)	0.39(0.91)	0.66	0.256
	emoti (52) vs neuro (52)	-0.14(1.22)	-0.65(1.08)	2.29	**0.012**
	agree (52) vs disag (52)	0.40(1.11)	-0.33(1.08)	3.41	**0.0005**
STE	extra (52) vs intro (52)	0.37(0.93)	0.69(0.78)	-1.94	0.973
	emoti (52) vs neuro (52)	-0.15(1.07)	-0.40(0.96)	1.26	0.106
	agree (52) vs disag (52)	-0.08(1.06)	-0.27(0.87)	1.01	0.157
SDG	extra (52) vs intro (52)	0.33(0.68)	0.19(0.84)	0.90	0.186
	emoti (52) vs neuro (52)	-0.08(0.97)	-0.58(0.83)	2.84	**0.003**
	agree (52) vs disag (52)	-0.27(0.99)	-0.46(0.98)	0.99	0.161

dialogues generated. All participants were asked to score each dialogue with 4 questions regarding dialogue quality and characters' personalities (e.g. from bad to good for quality, and from introvert to extravert for personality trait) on the 5-Likert scale. For instance, we assigned option "Strongly Extravert" as score 2, while option "Strongly Introvert" as score -2. Particularly, we set some groups without embedded personality for comparison, where *NOP* denotes dialogues generated using fine-tuned DialoGPT on our corpus with no personality, *GPT* denotes dialogues generated using plain DialoGPT, and *ORG* denotes dialogues collected in original screenplays. Each of these groups contains 2 dialogues, which are started with the same input sentences as generations with personality.

The results for dialogue quality evaluation are presented in Figure 5. Overall, we notice that the groups with personality (left) received lower quality scores than those without personality (right). Within generations with personality, *CTE* and *STE* have higher scores than *SCE* and *SDG*. On the other side, it is observed that the dialogues collected from original screenplays (*ORG*) receive the highest median and mean scores. While *NOP* has significantly higher median and mean scores than *GPT*, which indicates that fine-tuning DialoGPT on our corpus is able to increase the quality of generation.

We also present personality identification results in Table 4. We notice that *CTE* and *SCE* have 2 trait comparisons with significant p-values (differences), and *SDG* has 1 significant p-value. Overall almost all comparisons (except *STE* extra vs. intro) have p-values lower than 0.3, which tends to indicate human participants are able to identify personality traits towards a correct direction.

5.5 Examples

We present several examples of dialogues generated with the setting that reach the best performance (scene-level, and *SCE*), given the same input seed (statement) and on 8 personalities (Table 5). And in Table 6 we present examples with another input seed (question) on 2 extreme personalities. All examples are selected from the dialogues generated by the model trained on the Drama dataset, and they are expected to take place between two characters as the dialogue structure in dataset.

6 DISCUSSION

According to our results, we observe that the accuracy of scene-level personality is improved compared to film-level personality, with the possible reason that the overall labeled personality could not match a finer-grained utterance perfectly. However, based on narrative theories, a character's personality is supposed to remain consistent across the story duration. Currently, we only leverage the dialogues, characters and transitions to define the personality, discarding some other elements of screenplays, such as staging or directions. These elements also contain information reflecting authorial intentions, which could be used to assign the scenario context of dialogues and affect story progression specifically at a higher-level of representation of the narrative plot. These knowledge could be modelled either as an embedding or a classifier, ensuring to leverage updated larger language models (e.g. ChatGPT[26]) for the generation process.

We also note that linguistic features of personality are indirectly reflected in speaking or writing [8]. This might also be the reason of an observation that readers vary significantly in their judgments of personality in text [20]. In fact, we notice that there exist some deviations of the results between automatic evaluation and human evaluation. For example, in automatic evaluation, *CTE* and *SDG* do not have an expected high accuracy of personality identification compared with *STE*, while they have more significant differences than *STE* in human evaluation.

7 CONCLUSIONS

We propose a corpus and an approach for generating dialogues using pre-trained neural networks in the context of narrative films, with target personalities derived from the characters in screenplays. We use two levels of personality and 3 embedding methods to reach these results. From the presented results, we demonstrate that our approach is able to generate dialogues with the correct target personality, by separately embedding the personality *(SCE)* and using half pre-trained transformer framework. Also, by fine-tuning pre-trained language model on our corpus, dialogues with increased

Table 5: Examples of generated dialogues from the full range of personality trait combinations (8 separate ones).

(Extra, Emoti, Agree)	**Turn1(Seed)**: All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars.
(high, high, high)	Turn2: Well, that's a lot to ask. And it's not that far out. And the food there is the same. Turn3: It is for a friend, but thank you, I 'll get a couple of them, and a couple of the bags for the kids. Turn4: The kids. What the hell are you doing?
(high, high, low)	Turn2: What's that for? You can 't do that to my bag. Turn3: You know what the bag is worth to you at five thousand dollars a year and you can 't do that to that one anymore. Turn4: I can 't do that.
(high, low, high)	Turn2: Twenty five, what? It's just a little something, I think? That I got a couple other guys on this. And we're going out to lunch – I think, you know, if you really want to get some coffee, that I could come out here and work out some food... Turn3: You think I'm serious? Turn4: No. I'm just saying...
(high, low, low)	Turn2: Twenty-five dollars? Turn3: To make it to New York, then back to New England. That's it. I want to go back. And I want to go in a city... I'm gon na find myself somewhere in a city. Where are we going? Turn4: The city.
(low, high, high)	Turn2: Yeah. That's a good idea. Turn3: You think they can get out? Turn4: Yeah.
(low, high, low)	Turn2: You're kidding. You're making it up. Turn3: You're so crazy. You can 't f** believe it. Turn4: I can 't believe it either. You 'd think that I 'd be drunk. I'm just a kid.
(low, low, high)	Turn2: I'm trying. Turn3: You're not trying, son. You can 't just walk around the city with no gas on your hands. Turn4: I can 't.
(low, low, low)	Turn2: I don 't think it has to be. Turn3: I'm telling you, it does. Turn4: No, no. There's too many of them, too. You've got one... and it sucks. It's no good.

Table 6: Generated dialogue examples from the two extreme personality trait combinations.

(Extra, Emoti, Agree)	**Turn1(Seed)**: What on earth do you think the baby will be doing?!
(high, high, high)	Turn2: It's all right, man, I mean he's a little old one. He 'll grow outa that, it's just a baby. Turn3: I'm going. You 'll have to stay for the baby. Turn4: It's all right, man, we 'll get it over with.
(high, high, high)	Turn2: In the future, the baby is not a person to begin with, and so no one will recognize it until, so that the child can be the best person we can think of now. Turn3: I believe you. In twenty years the child will be as intelligent as the mother, the father, the child's mother, the child's father. Turn4: Then he will be a son of God.
(low, low, low)	Turn2: I don 't know, it was just a thought... Turn3: What you have to do is ask Mommy what the baby's name is. Turn4: I can 't do that! I don 't know it.
(low, low, low)	Turn2: I don 't know. You can 't have your own son and your son will never see the world. Turn3: What about the baby? What about him? Turn4: I don 't know if he was born yet.

variety can be generated on surface-text level and semantic-level compared with the original DialoGPT.

We acknowledge that research in conditional dialogue generation is a developing topic with enormous challenges to effectively represent and convey the desired conditions. Our research demonstrates the potential of such approach to investigate stylistic dialogue generation based on attributes of agents within the context of narratives. It could also be used to investigate the relation between the presentations of these attributes from narrative perspective and the embedding methods from technical perspective. As a character is a personified individual that entails human-like properties [24, 46], we believe that to enrich the representations of an agent's attributes in natural language could benefit products of narratives. For example, a virtual agent in a narrative-based video game could utter with more believable responses if they is incorporated rich attributes. Moreover, this can be a process working in real-time through leveraging deep learning techniques.

REFERENCES

[1] Monika Bednarek. 2017. The role of dialogue in fiction. *Pragmatics of fiction* (2017), 129–158.

[2] Sven Buechel, Susanna Rücker, and Udo Hahn. 2020. Learning and Evaluating Emotion Lexicons for 91 Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1202–1217.

[3] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 169–174.

[4] Seymour Chatman. 1978. *Story and Discourse: Narrative Structure in Fiction and Film*. Cornell University Press. https://doi.org/doi:10.1515/9781501741616

[5] Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-Driven Dialog Generation. 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)*. 3734–3743.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Jessica Ficler and Yoav Goldberg. 2017. Controlling Linguistic Style Aspects in Neural Language Generation. In *Proceedings of the Workshop on Stylistic Variation*. 94–104.

[8] Albert Gatt and Emiel Krahmer. 2018. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61 (2018), 65–170.

[9] Pablo Gervás. 2010. Engineering linguistic creativity: Bird flight and jet planes. In *Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity*. Association for Computational Linguistics, 23–30.

[10] Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A Neural Language Model for Customizable Affective Text Generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 634–642.

[11] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1587–1596.

[12] Chenyang Huang, Osmar R Zaiane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 49–54.

[13] O.P. John and S. Srivastava. 1999. The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of personality: Theory and research*. Guilford Press, 102–138.

[14] Sarah Kozloff. 2000. *Overhearing film dialogue*. Univ of California Press.

[15] Wilkins Leong, Julie Porteous, and John Thangarajah. 2022. Automated Story Sifting Using Story Arcs. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 1669–1671.

[16] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 994–1003.

[17] Chia-Wei Liu, Ryan Lowe, Iulian Vlad 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*. 2122–2132.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).

[19] François Mairesse and Marilyn Walker. 2007. PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. 496–503.

[20] François Mairesse and Marilyn A Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics* 37, 3 (2011), 455–488.

[21] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research* 30 (2007), 457–500.

[22] Jamie Matthews, Fred Charles, Julie Porteous, and Alexandra Mendes. 2017. MISER: Mise-en-scène region support for staging narrative actions in interactive storytelling. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 782–790.

[23] Pierre-Emmanuel Mazare, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training Millions of Personalized Dialogue Agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2775–2779.

[24] Robert McKee. 1997. *Story: Style, Structure, Substance, and the Principles of Screenwriting*. New York: HarperCollins.

[25] Robert McKee. 2016. *Dialogue: The art of verbal action for page, stage, and screen*. Hachette UK.

[26] OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt/. Accessed on Feburary 23, 2023.

[27] Shereen Oraby, Lena Reed, Shubhangi Tandon, TS Sharath, Stephanie Lukin, and Marilyn Walker. 2018. Controlling Personality-Based Stylistic Variation with Neural Natural Language Generators. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. 180–190.

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.

[29] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*. 2227–2237.

[30] Mihai Polceanu, Julie Porteous, Alan Lindsay, and Marc Cavazza. 2020. Narrative Plan Generation with Self-Supervised Learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*. AAAI Press.

[31] Julie Porteous, Marc Cavazza, and Fred Charles. 2010. Narrative generation through characters' point of view. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 1297–1304.

[32] Julie Porteous, Fred Charles, and Marc Cavazza. 2013. NetworkING: using character relationships for interactive narrative generation. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 595–602.

[33] Julie Porteous, Fred Charles, and Marc Cavazza. 2016. Plan-based narrative generation with coordinated subplots. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*. IOS Press, 846–854.

[34] Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning Personality/Profile to a Chatting Machine for Coherent Conversation Generation.. In *IJCAI*. 4279–4285.

[35] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/languageunsupervised/languageunderstandingpaper.pdf.

[36] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. 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)*.

[37] S. Rimmon-Kenan. 1983. *Narrative Fiction: Contemporary Poetics*. Routledge. https://books.google.co.uk/books?id=PqzWemM8C3cC

[38] James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110, 1 (2003), 145.

[39] Siqi Shen and Charles Welch. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

[40] Haoyu Song, Weinan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting Persona Information for Diverse Generation of Conversational Responses. *CoRR* abs/1905.12188 (2019). arXiv:1905.12188 http://arxiv.org/abs/1905.12188

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[42] Weixuan Wang, Xiaoling Cai, Chong Hsuan Huang, Haoran Wang, Haonan Lu, Ximing Liu, and Wei Peng. 2021. Emily: Developing An Emotion-affective Open-Domain Chatbot with Knowledge Graph-based Persona. *arXiv preprint arXiv:2109.08875* (2021).

[43] David R Winer and R Michael Young. 2017. Automated Screenplay Annotation for Extracting Storytelling Knowledge. In *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference*.

[44] Weilai Xu, Fred Charles, Charlie Hargood, Feng Tian, and Wen Tang. 2020. Influence of Personality-Based Features for Dialogue Generation in Computational Narratives. In *ECAI 2020 - 24th European Conference on Artificial Intelligence (Frontiers in Artificial Intelligence and Applications, Vol. 325)*. IOS Press, 2945–2946. https://doi.org/10.3233/FAIA200466

[45] Ze Yang, Wei Wu, Can Xu, Xinnian Liang, Jiaqi Bai, Liran Wang, Wei Wang, and Zhoujun Li. 2020. StyleDGPT: Stylized Response Generation with Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 1548–1559.

[46] R Michael Young, Stephen G Ware, Brad A Cassell, and Justus Robertson. 2013. Plans and planning in narrative generation: a review of plan-based approaches to the generation of story, discourse and interactivity in narratives. *Sprache und*

Datenverarbeitung, Special Issue on Formal and Computational Models of Narrative 37, 1-2 (2013), 41–64.

[47] 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?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2204–2213.

[48] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 270–278.

[49] Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9693–9700.

[50] Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards Persona-Based Empathetic Conversational Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6556–6566.