Bournemouth University

Stylistic Dialogue Generation Based on Character Personality in Narrative Films

Weilai Xu

Supervisors: Prof. Fred Charles
Dr. Charlie Hargood
Prof. Wen Tang

Faculty of Science and Technology
Bournemouth University

A thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

November 2022

Abstract

Traditional narrative systems consist of two steps of process, story generation and discourse generation. However, many interactive systems make more effort on story generation rather than discourse generation. For discourse generation, dialogue is an important way used to unfold story and reveal characters in stories, and it is reasonable to expand the capability of narrative system by exploring the potential of dialogue generation in narratives. Also, Recent research in conditional dialogue generation is mostly focusing on the context of natural conversation generation with speakers' profile information. While incorporating the styles that relevant to narratives is yet to be widely investigated.

According to the research made, in this document, we propose an approach using a pre-trained language model, in order to explore the potential of generating dialogues with embedded narrative-related features within the context of narrative films. In this approach, three different embedding methods are leveraged to incorporate Big-Five personalities of characters into transformer-based neural networks, training on a new corpus, which is created and well-parsed from screenplays.

We conduct experiments using both automatic metrics and human evaluation to measure the quality of the generated dialogue and personality identification accuracy. All the dialogues for evaluation and analysis are generated with settings of the perspectives of embedding method, personality trait, personality level, and film genre, which is to explore the impact of different setting on dialogue generation with additional narrative-related styles.

According to the automatic experimental results, we demonstrate that our approach is able to generate dialogues with increased variety. Also overall, the generated dialogues are able to correctly reflect the given target personality.

We also conduct three user studies for evaluate dialogues with human judgements. In the first and the second user study, we evaluate the dialogues generated with film-level personality using CTE (Combined Textual Embedding[1]) embedding method. The results show that human participants are inclined to perceive one extreme end of each

[1]See detailed description in Chapter 4.3.2

personality trait. In the third user study, we evaluate generated dialogues with all setting combinations synthetically. Overall, the results show that target personalities can be identified with various degrees of accuracy. Also, a negative correlation between personality identification accuracy and dialogue quality is observed.

In this thesis, we propose a new approach for stylistic dialogue generation and demonstrate its effectiveness. We believe the observations and discoveries could be a start and a tryout to apply deep learning technique and big data to boost narrative dialogue generation. And we also believe that our research can be applied in plenty of potential scenarios, such as helping the authors creating huge amount of conversations between different characters by popping utterance options corresponding to the character settings.

Table of contents

List of figures

List of tables

Acknowledgments

I would like to give my biggest thank to my primary supervisor, Prof. Fred Charles, for his consistent help, support, and guidance throughout my entire PhD. It is such a grateful journey that being supervised by you, cooperating with you, as well as sharing success and progress with you. Merci Beaucoup.

I also want to thank my secondary supervisor, Dr. Charlie Hargood, who always gives me a lot critical advice and helps me build up my scientific thinking.

Besides, sincere appreciation for those who were or still are in company with me throughout my PhD: Colleagues (Daniel Green, Hua Zheng), research administrators (Cansu Kurt Green, Karen Turner, and Naomi Bailey), the chaplain in BU (Tim Peters), and other friends (Festus, Patrice Li, and the "Letletme" WeChat group).

Of course, I would like to thank my family as well, especially my mother, who gives me her love and encouragement, as well as keeps praying for me. And my special and unique thank to my beloved girl, Huihui Zhu, who brings me so much happiness and laughing. It is a grace to have you in this tough PhD journey. Grazie Mille and Xiexie.

Last, thanks be to God, for His mercy endures forever.

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgement, the work presented is entirely my own.

Weilai Xu
November 2022

Chapter 1

Introduction

> *Humans have an endless appetite for stories.*
>
> — David Bordwell, in *Film Art: An Introduction* (Bordwell et al., 2020)

> *You're never going to kill storytelling, because it's built in the human plan. We come with it.*
>
> — Margaret Atwood, the author of *The Handmaid's Tale*

When we talk about stories, what is the first thought that comes across in our minds? We might think of the latest report in newspapers, the bedtime tales read to children every night, the panels organised in the most aesthetically meaningful manner in a comic book, or the sequence of scenes presented in a film. All of these examples are commonly observed in our daily lives and are very likely to be categorised into the scope of "story". However, do they have anything in common that makes them a "story"?

Normally in narrative theory, a "*story*" consists of a sequence of abstracted events, plus other elements (e.g. characters, time, location, etc. which are within *existent* concepts, as reported by narrative theorists (Bordwell et al., 2020; Chatman, 1978; Rimmon-Kenan, 1983; Toolan, 2013; Young, Ware, Cassell and Robertson, 2013), i.e. the content, or more plainly, ***what*** is told. Therefore, the reason of the examples above

are likely to be *story* is they contain almost all of these essential narrative elements and event chain, regardless the means by which they are expressed and conveyed. Particularly here, structuralists in narratology call these means as the "*discourse*", i.e. ***how*** the content in *story* is transmitted. In a common understanding, the *story* and the *discourse* are two constituent components of a *narrative* (Chatman, 1978; Rimmon-Kenan, 1983; Toolan, 2013; Young, Ware, Cassell and Robertson, 2013), in which they play the roles of content and style/form/expression respectively, following Herman et al. (2010).

According to Rimmon-Kenan (1983), a story in narrative sense is the narrated events and participants in abstraction from the text. While Chatman (1978) regards event as a change of state, or narrative action, which is *brought about by an agent or one that affects a patient.* From these definitions, it is not difficult to notice that the key and fundamental element in a story is event (Propp, 1968), which is an action carried out by some individuals, or *characters* in narrated domain (Herman et al., 2010). From a more intuitive perspective of linguistic manifestation, it could be described as a combination of pronouns and predicates. For example, "John gets up, and he says he is hungry". Also, from the perspective of temporal sequentiality, a story is widely treated as a series of events ordered chronologically, while a *plot* is these events with different orders and combinations of causation between events by authors (Chatman, 1978; Genette, 1983; Propp, 1968). In this document, story, along with its elements, are our primary considerations rather than plot.

In a story, everything that a character performs can be an action, which can be categorised into three primary types according to McKee (2016): 1) Non-verbal physical actions, such as body languages; 2) Mental actions: The thoughts in the characters' minds, which may affect their other types of actions; 3) Verbal actions: simply the speech or the dialogue[1]. Among all these types of narrative actions, we

[1]Apart from these three types, Chatman (1978) also mentioned perceptions and sensations, which might be summarised into the type of the general feelings.

found speech is most interesting one and common one, as it is the one involved in more than one characters. Like the novelist Elizabeth Bowen commented, "*Dialogue is what characters do to one another.*" **In the following of this thesis, we will focus on the verbal action, more specifically, dialogue, as well as investigate dialogue generation in narrative stories. By briefly introducing the concepts of story, event, and action, we establish a general context and theoretical foundation for our research.**

As a common type of action in narrating events in stories, dialogue can take place in both real world as well as in fictional world, which is created by artists in general sense, such as painter or author. And dialogue features similarly in terms of forms of representation and in these two worlds. This is to say that dialogue is an action for exchanging information verbally between more than one individual. However, in academic views of narratives (McKee, 1997), computational linguistics (Serban et al., 2018), and film industry (Berliner, 1999), there exist acknowledgement that in the fictional world, dialogue differs crucially from that in the real world, in the aspect of how it is produced in two worlds. According to them, the one in real world is taken place naturally and real-time between people, while the other is created intentionally to mimic the former one and has narrative directions. Therefore, in this thesis, we specifically call the one spoken by the real human being in the real world "*conversation*", and call the one created manually by human hands "*dialogue*", as McKee (1997) pointed out, *"**Dialogue is not conversation.**"*.

Apart from action, character is an another key element in story, no matter from which direction these two elements are related to each other, where Propp (1968) considered the character is subordinated to action, whereas Greimas (1966) proposed an inverse view. Because as Chatman (1978) mentioned, an action is supposed to be executed by or to affect character(s), which can also be viewed as a non-verbal or pre-verbal abstraction (Rimmon-Kenan, 1983), or a counterpart of action. In a story, a character is not just a name symbol, but a personified individual entails human-

like properties, such as enduring personality traits, relationships, emotions, and so on (Herman et al., 2010; McKee, 1997, 2016; Young, Ware, Cassell and Robertson, 2013). **In this thesis, on the basis of dialogue generation in narrative stories, we investigate the impacts that the characters with different attributes make on the dialogue generation in the fictional world.**

When a dialogue action[2] is realised and conducted in the form of the dialogue in the creative fictional world, it is actually transmitted from an abstracted narrative action in the *story* scope to a representational expression in the *discourse* scope by a certain mean. More specifically, a dialogue action could be realised textually as part of a novel, or verbally in a theatre play, or a film. This transmission process can be called medium of narrative (Ryan et al., 2004) according to – Webster's Dictionary –, which refers to "*A channel or system of communication, information, or entertainment*" and "*Material or technical means of artistic expression*". Also, Rimmon-Kenan (1983) pointed out that the *discourse* is a medium-dependent process. Referring to the taxonomy of Ryan et al. (2004), a story can be transmitted into different discourses through singular or multiple media channels, which could be linguistic, acoustic, and visual from the perspective of sensory dimensions. We also consider that the category of linguistic could be further split into textual and verbal. In this document, two (narrative-based) multiple media channels are selected as our specific research context: films and video games. This is due to the fact that dialogues play an important role in unfolding the story and revealing characters (Nelmes, 2011) under the writer's point of view.

In narrative films[3], dialogue is an essential element that serves 9 purposes from the perspectives of communication and aesthetics (Kozloff, 2000). Character revelation is an important one among these purposes, as Crothers (2016) says, "*It (Dialogue) conveys so much in a few words ... with it can make the audience know the depths of*

[2] An abstraction consists of speech act and dialogue states. See more in Chapter 3.

[3] Compared with narrative films, a significant difference of narrative video games is that they contain interactive module for users to alter elements in stories, e.g. the order of events. However, it is still reasonable to treat them as our context similarly as they basically share stories, characters, and manifestations.

his (the character's) being". For example, in a dialogue excerpt shown below of the film *Inception*[4], protagonist Cobb is furious with an uncontrolled situation where his teammate Arthur has not perfectly done his research job, which is likely to cause an unexpected attack of enemies.

ARTHUR: *Calm down.*

COBB: *Don't tell me to calm down - you were meant to check Fischer's background thoroughly. You can't make this kind of mistake - we're not prepared for this kind of violence-*

The excerpt above contains a series of negative statements that reveal Cobb's high expectation, anger and disappointment towards his teammate, as well as his concern about the probable attack. **Therefore, we set narrative film, an appropriate narrative medium, as the more specific context of this research, where human creators represent stories, as well as reveal characters' attributes using dialogues.**

Traditionally, human people create and express stories through various media as previously mentioned, which is accomplished with human intelligence. While this procedure can also be ranged from computer-aided to theoretically fully computerised. According to Herman (1998), one of the goals that computational approaches to textual narrative pursue is the modelling of narrative intelligence, which focuses on the processes by which narratives are generated. *Narrative systems* are a tool supporting the generation of narratives from narrative representations using computational mechanisms/algorithms/solutions. A narrative comprises of a story and its discourse. A narrative system usually includes two major procedural steps for generating a narrative: *story generation*, automatically creating narrative elements and plans, and *text realisation*, transforming narrative structures into actual text (e.g. from dialogue action to each utterance in a dialogue).

[4] *Inception*, 2010, directed by Christopher Nolan.

However, Gatt and Krahmer (2018) and Gervás (2010) imply that narrative systems are inclined to put more emphasis on story generation, than on (text) realisation. Considering existing narrative systems (Cavazza et al., 2002; Mateas and Stern, 2003; Matthews et al., 2017; Porteous et al., 2010*b*, 2013), it is not difficult to find that these systems make a heavy effort to generate consistent and reasonable narrative structures, but have less capability to convert the stories into natural language for human readers. Also, even for the narrative systems with the ability to generate discourse, the process of realising the planned plot with narrative events and other elements into text is achieved by exhaustive pre-defined semantic lexical grammars (Callaway and Lester, 2002; Cavazza and Charles, 2005)(e.g. Tree-adjoining grammars (Joshi and Schabes, 1997)), or by an ontology of templates and operators (Cheong and Young, 2015; Gervás et al., 2004; Pizzi et al., 2007). Such an approach suffers from the problems of scalability and efficiency, as well as the "authoring bottleneck" (Lin, 2016; Mateas, 2007). These limitations could, to some extents, explain the fact that less consideration has been put into text realisation in narratives, even though it is an important step in narrative generation. These are likely to be linked to the limitations in natural language generation. In the very recent years, with the rapidly improving performance of computer hardware, large scale stochastic computation, or deep learning approaches using non-linear neural network architectures have revived its popularity by succeeding as new solutions in broad natural language processing: e.g. machine translation (Cho et al., 2014; Sutskever et al., 2014), language modelling (Devlin et al., 2019; Mikolov, Chen, Corrado and Dean, 2013; Mikolov et al., 2011; Pennington et al., 2014; Radford et al., 2018), response generation (Sordoni et al., 2015; Wen et al., 2015; Yan et al., 2017; Zhang et al., 2020), etc.. They have also allowed to replace the previously dominating linear model based machine learning approaches.

Among these sub-topics in natural language processing (NLP), natural language generation (NLG), or more specifically dialogue generation, is the core problem in

traditional dialogue system[5] (Serban et al., 2018). While for the neural network based *end-to-end dialogue system*, which has a vague component division, NLG even can be treated as the only component of this kind of dialogue system. **Regarding the insufficiency of ability of the text generation in narrative systems, in this thesis, we consider to leverage the technique of end-to-end dialogue generation based on neural networks to improve this situation.**

The *Encoder-Decoder framework* (Sutskever et al., 2014) is an influential architectures mostly leveraged for dialogue generation, which is also a sequence-to-sequence (seq2seq) task like machine translation. Based upon the Encoder-Decoder framework, various improvements with some optimistic methods such as attention mechanism (Vaswani et al., 2017) are witnessed in dialogue generation. However, only to generate a grammatically correct and short-term semantically consistent response in dialogue is not sufficient for imitating real utterances spoken by human beings. In the real world, people express similar information by different linguistic choices, with different emotions or sentiments, as well as in different styles. Therefore, how to make the generated dialogues with more variety, with more consistency, and with more style become the research points. Conditional dialogue generation is an advanced field in dialogue generation, which focuses on generating stylistic dialogues by incorporating various additional information along with normal textual information (e.g. dialogue history). To achieve this purpose, most existing works incorporate speaker related additional information, such as speaker profile (Li et al., 2016; Zheng et al., 2020), sentiment or emotions (Ficler and Goldberg, 2017; Ghosh et al., 2017), and tense (Hu et al., 2017). By controlling this additional information, the generated dialogues are supposed to reflect or present a certain state linguistically which corresponds to the target additional condition(s), as well as to improve the linguistic performance on some aspects, such as semantic consistency.

[5] A traditional dialogue system incorporates the following components normally: Automatic Speech Recognition, Natural Language Understanding, Dialogue State Tracking, Dialogue Response Action Selection, Natural Language Generation, and Speech Synthesis (See Chapter 3).

During investigating this additional information which is incorporated usually in conditional dialogue generation, we observe that most research only consider local features (Colombo et al., 2019; Li et al., 2016), 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.

1.1 Research Questions

Thus, in this thesis, we intend to investigate the potential and impact of stylistic conditional dialogue generation in the context of narrative stories based on character personality which derived from narrative films, by leveraging the technique of deep learning. It is feasible to obtain rich structured discourse information in narrative-based film screenplays rather than other narrative text (Jhala, 2008; Winer and Young, 2017). Also, the advanced pre-trained transformer-based language models are capable to provide the ability for generating sensible grammatically correct text. Considering all the investigation aforementioned, we conclude our motivations based on current limitations as follows:

1. Narrative system is limited to realise a story from structured plots to text in natural language.
2. Deep learning is an advanced and efficient technique to generate conditional dialogue by adding various additional information. However, the influence on the generated dialogues of additional information from the authorial aspect has not been sufficiently explored.

Our research uses a neural language model along with features extracted from screenplays to answer the following research questions:

1. How to reflect authorial intentions on characters' personalities from narratives and how to incorporate them into deep neural networks?
2. What are the influences on dialogue generation by adding different characters' personalities derived from narratives using deep learning techniques?
3. What are the differences on dialogue generation of the influence by using different embedding methods and datasets for characters' personalities?

We believe our research could contribute to both narrative community and natural language generation community from the following aspects:

1. An approach for generating conditional dialogues by utilising Big-Five model based personality traits from film screenplays. Our approach based on three embedding methods can generate varied dialogues which are able to reflect selected target personality traits.
2. Experiments and detailed analysis of the impact of personality combinations, levels of personality, and embedding methods on the performance of the dialogue generation.
3. A well parsed, segmented, and labelled dataset from IMSDb[6], which contains dialogues in screenplays, characters, scenes and corresponding personalities.

1.2 Thesis Structure

The structure of this thesis document is as follows. Chapter 2 discusses related work on computational narrative, narrative system, and visual narrative productions (e.g. film). In Chapter 3, we discuss related work on dialogue generation, neural dialogue generation, and language model. In Chapter 4, we introduce the details of our approach, including the selection of a proper personality model, the representation of characters' personalities based on he selected personality model, and the ways to incorporate

[6]The **I**nternet **M**ovie **S**cript **D**ata**b**ase

characters' personalities into neural networks as well as the adapted language model. We also report results of automatic evaluation and analysis of the observations from the results in Chapter 5. And the results and analysis of human evaluations with user studies are presented in Chapter 6 and Chapter 7. In the final chapter, we make conclusions regarding all the results and observations, as well as discuss the limitations and future envisions.

Chapter 2

Literature Review of Narratives

In the next two chapters, we will explore research works which are related to both narratives and dialogue generation. Each topic is intrinsically intertwined, but they need to be presented separately as well, or at least from a different emphasis, which is the reason for the two main literature review chapters. In this chapter, where narratives are the focus, we will explore two sub-topics of *computational narratives* and *visual narrative media*.

In the first sub-section, we will discuss the research works that relate to computational narrative generation, within which plan-based narrative systems and narrative systems using neural networks will be discussed. For the different types of narrative system, we intend to discuss these related works by the two stages of procedure in narrative generation, which are story generation and text realisation.

In the second sub-section, we will discuss visual narrative media, including narrative films and narrative video games. Particularly, the relation between character and these two media, as well as dialogues in narrative media. The investigation on this sub-topic provides the research context of this document.

Investigating and discussing these sub-topics, we are able to provide a better theoretical foundation on the use of characters in narratives, as an essential narrative element, as well as be able to observe the gaps that exist in computational narratives.

2.1 Computational Narrative Generation

We structure this section by the category of the existing computational narrative systems according to their techniques: non-neural narrative systems and neural narrative systems. Within each category, we will introduce them following the stages of narrative generation: story generation and text realisation. **The goal of this section is to identify the limitations regarding existing narrative systems, which leads to the motivations.**

2.1.1 Non-Neural Narrative System

Story Generation

The planning system in Artificial Intelligence (AI) is used to produce plans to solve problems. Planning problems are composed of an initial state, a goal state, and a set of potential actions which can be executed. Planning algorithms are designed to construct a solution plan, which is a sequence of actions which allow to resolve the defined planning problem, achieving the goal state from the given initial state.

As narrative theorists (Bal and Van Boheemen, 2009; Chatman, 1978; Rimmon-Kenan, 1983) formalised a story as a sequence of narrative events linked with each other causally and temporally, this conceptual formal nature of the progress of stories was considered as a foundation, which is to apply AI planning method to generate narratives computationally (Young, 2000) Since them, researchers started to view the generation of narrative as a planning problem or process (Lebowitz, 1985; Porteous et al., 2010*a*; Riedl and Young, 2010). By applying planning algorithm in narrative generation, storytellers are free to add a action to a story plan for providing more automation (Kybartas and Bidarra, 2016), as well as reasoning about multiple narrative goals in a story rather than a single goal (Young, Ware, Cassell and Robertson, 2013), as

traditional template-based (Concepción et al., 2018) and grammar-based (Pemberton, 1989) story generation approaches do.

One essential advantage of planning approach from which story generation can benefit is that it provides potential and feasibility to create highly branching story with little effort (Young, Ware, Cassell and Robertson, 2013). Branching stories are represented as directed graphs, within which the nodes represent the events or scenes and arcs denote decisions (Riedl and Young, 2006). Therefore, this allows participants to interact with the story by making different event decisions and unfold different story branches.

To achieve this goal of story generation, Hierarchical Task Networks (HTN) (Nau et al., 1998) is a planning strategy that decomposes a high-level task or goal into a sequence of multiple smaller sub-tasks (actions) until primitive ones. By executing these primitive actions directly, the original story goal can be accomplished. There are many plan-based narrative systems apply HTN planning strategy (Cavazza et al., 2002; Skorupski et al., 2007; Skorupski and Mateas, 2010).

Cavazza et al. (2002) propose a character-based system using HTN planning to generate the branching stories. Once the states of objects in the story world are changed by user participants or other characters (e.g. a box of chocolate or other items are picked up by a character), the planner is able to do re-planning by searching for other event branches due to the change of the world. With this interaction between the user and the story, different events occur and story branches are supposed to be progressed by re-planning.

Apart from generating stories with variety by interacting with branching events, we also notice that plan-based systems are able to provide story diversity through re-planning by altering the attributes of some narrative elements in the sense of existent, particularly character.

Porteous et al. (2010*b*) present an interactive narrative system in the context of Shakespeare's *Merchant of Venice*. In this system, the planner is able to conduct plot

re-planning whenever a character's different point of view (PoV) is switched by user interaction. For example, when the PoV is changed to a victim perspective from a neutral one, the event with more emotional accusation would be selected to match this character's victim identity.

Another work introduced by Porteous et al. (2013) uses a classification of character social relationships with three major categories (affective, romantic, and default - each of them contain different sub-levels of relationships.) to affect the story planning process. Before the story planning process, changes in these relationships between certain two characters can be made by user interaction as an initial planning state. Then these changes impact the selection of narrative actions, leading to different story goals that correspond to social network configurations and previous actions.

Pizzi et al. (2007) introduce a narrative system based on a standard planner called Heuristic Search Planner (HSP). Their system is integrated with a natural language input for users to update characters' beliefs and emotional states. In this way, the selection of the narrative action can be influenced during story progress. On top of this system, Peinado et al. (2008) apply a cognitive additional layer for modelling intelligent characters using Belief-Desire-Intention (BDI) theory, by which the narrative causality can be reinforced.

The research above demonstrate that AI planning can be used to generate branching stories with variety in terms of making narrative event decisions and altering attributes of narrative elements with interaction. In the next sub-section, we will introduce existing approaches for realising these stories and then discuss the limitations of these systems.

There are also some works that use case-based reasoning (CBR) approach for story generation (Gervás et al., 2004; PÉrez and Sharples, 2001; Swanson and Gordon, 2012). For example, in the CBR system of Gervás et al. (2004), the information of user queries (e.g. characters, places, Propp functions) is used to retrieve the cases depending on Propp morphology and available characters from a selection of stories from Afanasiev

compilation used by Propp. Then plot-unit template in each retrieved case will be partially complemented with the information of the query, an ontology of explicitly declared relevant knowledge, and other cases. By such CBR process, believable plot plans can be generated from different perspectives with empirical knowledge.

Text Realisation

Normally plan-based narrative systems (Cavazza et al., 2002, 2009; Mateas and Stern, 2003) include multiple media (e.g. visual, textual) to realise and represent planned stories. In this and following sections, text realisation in plan-based narrative systems is to be discussed particularly as the context of this thesis set.

We observe that many plan-based narrative systems make more effort on story generation rather than text realisation, which is the process through which the planned story can be converted into text in natural language. This observation is also implied in the works of Gatt and Krahmer (2018), Gervás (2010). Therefore, we intend to investigate some NLG approaches used in narrative generation systems.

Some narrative systems apply pre-defined semantic lexical grammars to achieve this process. For example, Joshi and Schabes (1997) defined a grammar formalism called Tree-Adjoining Grammar (TAG), in which each node of a tree structure is associated with a lexical item[1]. These structural lexical items then can be realised into actual text by operations, such as substitution . For example, in a simple tree structure of TAG shown in 2.1 for sentence "John likes Mary" in natural language, NP, VP, and V denote grammatical categories Noun phrase, Verb phrase, and Verb respectively. And an operation of substitution is to replace a non-terminal leaf with another tree that has the same grammatical label, i.e. to replace current NP ("Mary") with another NP ("the girl", where "Det" denotes Determiners)

Cavazza and Charles (2005) leverage this lexicalised grammar to realise the generation of semantic contents into utterances in dialogue with different affinities and

[1] also called *Part of Speech* (POS) in NLP.

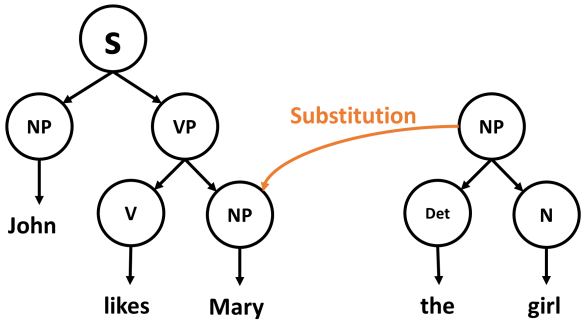


Fig. 2.1 An example of the substitution operation in TAG.

modalities of expression. Callaway and Lester (2002) used another grammar called FUF[2]/SURGE[3] (Elhadad, 1993) to convert sentence structures into surface sentence string.

Some other narrative systems follow NLG process based on an ontology of templates and operators rather than a full grammar (Cheong and Young, 2015; Gervás et al., 2004; Pizzi et al., 2007). Gervás et al. (2004) use templates to provide verbs (e.g. from Propp (1968)'s character functions) as well as other parts, resulting in final surface text from CBR plot plans.

Besides, some other works alter discourses with different linguistic phenomena on surface text level. Bowden et al. (2016) present a Monolog-to-Dialog (M2D) generation architecture to convert a deep representation of story into different versions of a two-speaker dialogue using a parameterisable framework. This paper uses various parameters of linguistic phenomena (e.g. length of sentence, pronominal, repetition) to achieve the dialogue styles for different speaker voices. Rishes et al. (2013) present a working model of reproducing different tellings (dialogue or monologue) of a story from its representation. They also propose an automatic method for converting a representation called STORY INTENTION GRAPH (SIG) (Elson and McKeown, 2009) to another called deep syntactic structures (DsyntS) (Lavoie and Rainbow, 1997). And the DsyntS representation can be realised in to surface sentences through PERSON-

[2]Functional Unification Formalism Interpreter.
[3]A Syntactic Realization Grammar for Text Generation.

AGE (Mairesse and Walker, 2007), a NLG engine integrated with over 50 linguistic parameters within speakers' personality dimensions.

These approaches used in non-neural narrative systems are able to represent story plans or structures in text precisely, as well as easy to understand. However, they are mostly based on the empirical ontology of templates and operators, which leads to limited semantic and syntactic level representation, i.e. less ability of generation in terms of variety.

2.1.2 Neural Narrative System

Story Generation

In NLP, deep learning technique was first widely applied to solve machine translation and chatbot tasks. Progressively, more research work using deep learning for narrative generation have reached completion in the last few years.

In machine translation and chatbot tasks, a sentence can be represented as a sequence of word vectors. The preceding vectors are used to calculate the possibilities of the following words and the one with the highest possibility is to be selected. Similarly, a story also consists of a sequence of events, where the following events are supposed to be decided by the preceding ones. Therefore, it occurs a potential to see a story generation process as a neural language generation process, in which an event can be viewed as an equivalent of a word vector. If this stands, then how to represent narrative events as vectors which can be calculated by neural networks become the key of neural story generation.

Chambers and Jurafsky (2008) developed a representation that took note of the event/verb that occurred and the type of dependency that connected the event to the protagonist. This could be seen as one of the examples (Chambers and Jurafsky, 2008, 2009) of early event abstraction applied in learning approaches. Follow these works, Pichotta and Mooney (2014) proposed a 4-tuple event abstraction in the form

verb (subject, object, prepositional), where three entities are all relation to the verb. By encoding the pairwise entity relationships between events, the accuracy of inferred components of the event increased. Pichotta and Mooney (2016) also present a similar story abstraction in 5-tuple by adding preposition and use LSTM[4]-based neural network to model and predict information about event abstraction.

Martin et al. (2018) also treat a sentence as a 4-tuple event representation, containing a verb, the subject of the verb, the object of the verb, and a modifier for the other linguistic constituents. Then they train and generate LSTM seq2seq neural network with this representation. These works can also be viewed as a trial of combining the planning and surface text realisation for controlling the dialogue generation using neural networks. Also, Ammanabrolu et al. (2020), Tambwekar et al. (2019) use similar N-tuple semantic abstractions to represent events, which are then passed into seq2seq model to predict next event.

As mentioned before, the character is one of the most important elements in a story. Therefore, it is natural to consider achieving a story generation task from the perspective of the character. Liu et al. (2020) proposed a character-centric neural storytelling model, which explicitly encodes characters in distributed embeddings to guide the story generation. In this model, the story is generated sentence by sentence controlled by the context environment and the character embedding, which is trained on the corpus of movie plot summaries extracted from Wikipedia.

Some other story generation approaches model the story by structuring the sentence text from the perspective of linguistic. They design representations for stories by abstracting over linguistic constituents for each sentence. Fan et al. (2019) decompose a story into a sequence of action plans (verbs) and placeholder entities, which are supposed to be replaced by actual references during generation.

[4]**L**ong **S**hort-**T**erm **M**emory (Hochreiter and Schmidhuber, 1997)

Text Realisation

Since deep learning technique started flourishing, dialogue system is developing towards the end-to-end direction, which is to omit some components like dialogue state tracking, response action selection but keep NLG. Like dialogue systems, narrative systems using deep learning technique are in the similar progressing trend. Because of this trend, the "story" in story generation using deep learning is more likely denoted the actual textual sentences in stories rather than just the representations of event and existent according to traditional narrative theories. Therefore, in this sub-section, we will not just introduce ones inspired by the narrative process in traditional plan-based narrative systems, but also consider others that generate textual story more directly.

The works inspired by traditional plan-based narrative system follow the its process, which is generating event sequence first and then converting each event into sentence(s). As we introduced in the last sub-section, those approaches, which generate events based on preceding n-tuple event abstraction using neural networks, translate these generated event into actual sentences afterwards (Ammanabrolu et al., 2020; Martin et al., 2018; Tambwekar et al., 2019). For example, Martin et al. (2018) trained a neural network called event2sentence on a parallel corpora of sentences from a story corpus and the corresponding events.

Jain et al. (2017) use a compromised end-to-end strategy to generate story text using neural networks. they use several a sequence of independent short textual descriptions (sentences) describing a scene or event as the input for a LSTM-basd seq2seq model, and then the outputs are comprehensive story-like summaries.

The promising textually generative ability of large-scale pre-trained language models (See more in Section 3.2.2) attract researcher to apply them for neural story generation. See et al. (2019) conducted a study to show that GPT-2 (Radford et al., 2019) has better performance than the state-of-the-art neural story generation model (Fan et al., 2018) at that time. Guan et al. (2020) proposed to utilise commonsense knowledge

from external knowledge base to generate stories using pre-trained language model. By doing this, they improved the issues suffered with the original GPT-2, which are repetition, logic conflicts, and lack of long-range coherence.

2.1.3 Discussions

Kybartas and Bidarra (2016) follow the narrative theory provided by Chatman (1978) and other structuralists to divide story into *plot* and *space*[5]. Referred to this structural division, they introduce a typology of non-neural narrative systems in terms of the degree of both plot and space automation, which are with five degrees respectively, from manual to automated. According to their typology table, we notice that most non-neural narrative systems intend to investigate automated plot generation rather than automated space generation. And they pointed out that there is "no intent" to creatively modify the existing story world. This indicates that 1) the lack of variation of the narrative existents (e.g. characters) in non-neural narrative generation.

Some works (Gatt and Krahmer, 2018; Gervás, 2010) also implied, compared to the plot generation, the text realisation is not the main consideration for non-neural narrative systems. Cavazza and Charles (2005) also note that "*most interactive storytelling systems are still missing the ability to generate dialogues between characters*". Therefore, there are two more limitations of text realisation that occur in plan-based narrative generation works: 2) The text realisation stage is highly dependent on the generated plans, which means the variations of the narrative discourse mainly 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. 3) The lexication in the text realisation stage is based on the empirical ontology of templates and operators, which leads to limited semantic and syntactic level representation, i.e. less ability of generation.

[5]In Chatman's theory, these two concepts are called *event* and *existent* as mentioned in Chapter 1.

Deep learning techniques support neural narrative systems to generate story plots and realise them into text without templates and operators, which provides variety of the generations and reduces human expertise requirements. However, like non-neural narrative systems, so far neural narrative systems also tend to focus on plot or story structure generation, making effort to improve causality between events. And their processes of text realisation are based on the event abstractions on the plot aspect, while the functions of existents are less incorporated yet, and the ability of generation of advanced NLG techniques using pre-trained language models in neural narrative systems has yet been sufficiently investigated.

2.2 Visual Narratives Media

In this section, we intend to introduce some visual narrative media, particularly narrative films and narrative video games, as well as the character and dialogue in them. **The goal of this section is to introduce the context and research object of this thesis research.**

2.2.1 A Brief History

Telling stories is one of the instincts of human people, which happens over time in human history, and in daily life every moment. Enormous works of *narrative content* are created across different cultures and the world. Comparing with the oldest ancient narrative art on rocks dating back 30,000 years in Africa (The British Museum, 2019) or 44,000 years ago in Indonesia (The Washington Post, 2019) , nowadays people are able to create stories and present them through much more various high-tech media, such as 3D films, video games, and virtual reality.

According to Dehejia (Dehejia, 1990) and other literature (Damiano and Lieto, 2013; Meister and Schernus, 2011; Small, 1999), a consensus exists that some distinct modes[6] (or types) of visual narration categorised by the characteristics of the ways where artists choose to represent the space and shape time within their narrative artworks.

Due to the limitations of technology and/or the underdevelopment of language, most ancient or traditional visual narrative contents were created with a single certain mode or a single medium (e.g. pictorial, sculptural, literal, oral). Before the literacy era, many rock paintings only contain simple graphical elements, such as lines and dots, or multiple abstract symbols for presenting stories, which caused difficulties to determine the sequential order among these overlapped story events. This narrative mode is called *simultaneous narrative* discussed by Petersen (2010) regarding to its features. Also, without the aid of writing or recording, the cultural information and interpretation of narrative art have to be passed on from one generation and the next verbally in preliterate societies.

The emergence of literacy, as well as rapidly advancing new technologies, have significantly expanded the development of the visual narrative. Filming and cutting technology allows storytellers to narrative a series of events with various structures (e.g. linear, flashback, montage) in a single frame continuously, rather than merely present events in a certain narrative mode (e.g. sequential mode, continuous mode) with a fixed order as most static graphic narrative contents, like fresco and comic books. Apart from the impact of the development of technology, literacy, or language more specifically, also provides another dimension of information, either in text or in voice, for storytellers to establish more complex visual narrative content. By the means of these two developments, artists and authors are able to make more creative and various visual narrative contents, which are also more concrete but less relied on the interpretation of the audience.

[6]Considering the goal of this section is not to discuss the features of these various narrative modes, here I only list these modes as following: monoscenic, continuous, synoptic, sequential/linear, panoramic/narrative network, progressive, and simultaneous mode.

2.2.2 Narrative Films

> *Film is a visual medium.*
>
> — Alexander Steele, in *Writing Movies: The Practical Guide to Creating Stellar Screenplays* (Steele et al., 2006)

> *The stories we love best do live in us forever. So whether you come back by page or by the big screen, Hogwarts will always be there to welcome you home.*
>
> — J.K. Rowling, the author of *Harry Potter* Series

Since Lumière brothers screened the mostly acknowledged first film [7] in 1895, narrative films became one of the most dominant and important media that are utilised to tell stories. The record of the worldwide highest-grossing films by year have soared up from 4 million US dollar[8] in 1929, the first year of the Academy Awards, to 2.8 billion US dollars[9] in 2019 (Wikipedia, 2021) [10].

Characters in Narrative Films

> *I think the best stories always end up being about the people rather than the event, which is to say character-driven.*
>
> — Stephen King, in *On Writing: A Memoir of the Craft* (King, 2000)

In contrast to the non-narrative film, as a form of art or experimental oriented rather than entertaining oriented, which usually contains non-representational elements, in narrative films, the relations between the shots and/or between the elements of

[7] *La Sortie de l'usine Lumière à Lyon*, literally, "the exit from the Lumière factory in Lyon", or, under its more common English title, *Workers Leaving the Lumiere Factory*.

[8] *The Broadway Melody, 1929*, directed by Harry Beaumont.

[9] *Avengers: Endgame, 2019*, directed by Russo Brothers.

[10] The reason the stats only traced back to 2019 is that, due to Covid-19 Pandemic, the cinematic industry has been impacted catastrophically since early 2020 across the world. The highest-grossing film in 2020 is *The Eight Hundred* with 400 million US dollar box office.

the image are supposed to be perceived, such as temporal, sequential, or cause-and-effect (Aumont, 1992) . This is to say that in narrative films, the narrative information, or the story itself, is stored in and represented by the relations of the elements. Among the multiple elements (Young, Ware, Cassell and Robertson, 2013) in a story, McKee (1997) points out particularly that character is a central and essential **substance** for a story, no matter it is singular-protagonist or plural-protagonist. He describes it with an analogy "*far more profound than mere words beats at the heart of a story*". In other words, a story is driven by the characters, as well as the relations among characters. By creating characters on the screen, the intentions of the authors are able to convey; By watching characters on the screen, the audience is able to get resonated in their hearts. It is characters that start a story and progress the storyline, cause events happening and resolute events ending, reflect authors' intentions and bond the story with the audience (Bordwell et al., 2020).

As a narrative text format for representing a story, screenplay, or film script is special because it contains more structured discourse information (Herman et al., 2010; Jhala and Young, 2010; Papalampidi et al., 2020; Winer and Young, 2017) than other narrative texts such as news stories (Chambers and Jurafsky, 2008) or fables. In a screenplay, scriptwriters follow a standardised and rigid format (Riley, 2009) to compose its elements (e.g. dialogues, staging directions). Among these elements, other than a few short headings and transitions that are less related to characters, most of the elements are strongly relating to characters, including what and how a character speaks out (dialogues), as well as what and how a character acts and thinks about (actions, parenthetical directions).

2.2.3 Narrative Video Games

Unlike some formalised video games (such as sports games) which challenge players' reactions, operations, and strategies in a playfield, narrative-based video games provide

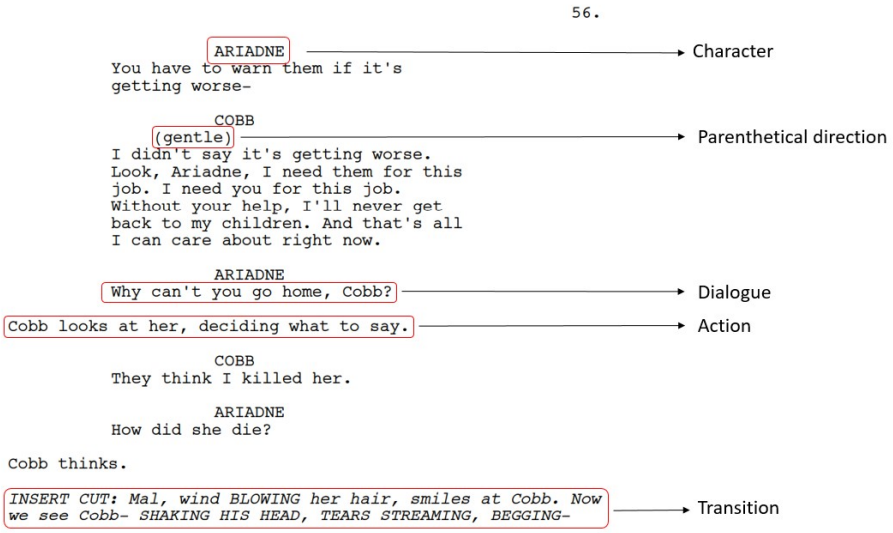


Fig. 2.2 An example of screenplay excerpt in *Inception, 2010*, directed by Christopher Nolan.

players with engaging scenes to participate, interact, and experience in a fictional world inspired by the real world. This kind of theoretical distinction is mostly a consensus in the community of game producers (Ryan, 2006). To some extent, they are more like films rather than video games.

Ludologists usually do not tend to regard computer video games only as games but not as narratives. For example, Juul and Ping-ping (2010) and Eskelinen (2001) argued that the rearrangement of events in games is not allowed: "*Games almost never perform basic narrative operations like flashback and flash forward*". However, as there are more cinematic cut scenes with non-chronological order in narrative films (e.g. *Memento*[11] and *Tenet*[12]), it also exists in video games, such as *Life is Strange*[13]. In *Life is Strange*, players are assigned the ability to let the protagonist Maxine Caulfield rewind time, undo any of her actions, and provide replay value to the game. For example, the player can rewind after examining an object or having a particular conversation, in order to use that new information on Max's benefit in the game. The time rewinding

[11] *Memento*, 2000, directed by Christopher Nolan.
[12] *Tenet*, 2020, directed by Christopher Nolan.
[13] *Life is Strange*, developed by DONTNOD Entertainment and published by Square Enix.

can either be forced for progressing the storyline (such as Maxine can only find the way to prevent her friend Chloe to be killed in the bathroom after rewind time) or be optional for additional achievements (such as collecting photos).

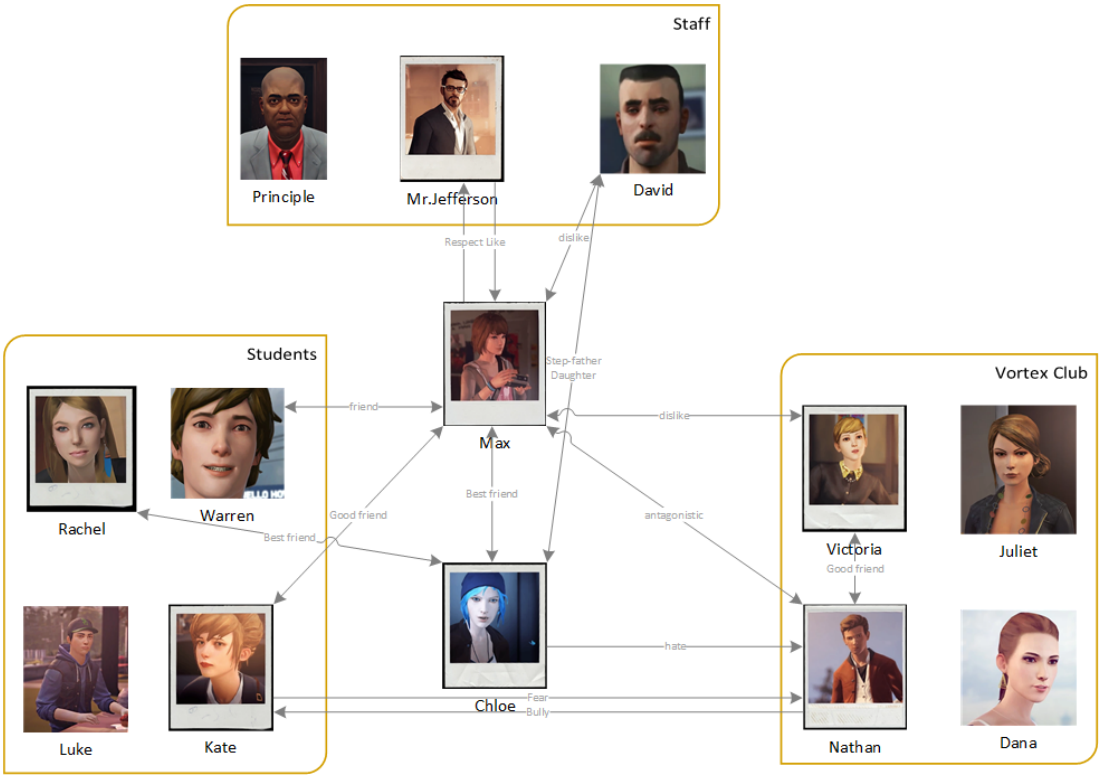


Fig. 2.3 The relationships between main characters in *Life is Strange*.

Characters in Narrative Video Games

As in narrative films, the character(s), whether the one we play as or the ones we interact with, is also an important element of telling stories and evoking an emotional response in the video game world. Adams (2014) points out a good character in a video game should be designed to be competently constructed, credible, and easily identified. Apart from the last goal that is towards business consideration, the first two goals are regarding game quality consideration, and Adams (2014) also mentioned that compared with past video games, which limited characters' attributes to physical details, recent games "*have made an effort to model social relationship and emotional states*". These efforts, including the design of characters' personalities and language, are important

to help make a character more believable, as well as make a game more appealing. In *Life is Strange*, the relationships (Figure 2.3) among characters are designed around protagonist Maxine, and the degree of advancement in which is mostly up to Maxine's (players') choices. Also, in different stages of these relationships, the usage of the language varies in order to reflect characters' personalities and emotions accordingly.

2.2.4 Dialogues in Visual Narrative Media

Although in such visual visual narrative media, it is capable or recommended to convey information visually as much as possible using actions, shots, focuses, customs, and settings, there still exists much essential and subtle information that can subtle be conveyed by dialogues as German-born American political theorist Arendt (1968) writes, *"We humanize what is going on in the world and ourselves only by speaking of it, and in the course of speaking of it we learn to be human"*. And to the most extent, this kind of information is about the realisation of designed stories and characters.

However, *dialogue* in narrative products should never be treated as equal as *conversation* in real life, as *character* is different from human being. The dialogues in narrative products have their own characteristics that are distinct from real conversations. Both McKee (1997) and Kozloff (2000) agree that the core difference between these two is that screen dialogue at the bottom is always an imitation of natural conversation, regardless they could be both spoken in the same language, or both followed by the same grammar and syntax. Because the screen dialogue is "*scripted, written and rewritten, censored, polished, rehearsed, and performed*" (Kozloff, 2000), while the real natural conversation is "*full of awkward pauses, poor word choices and phrasing, non-sequiturs, pointless repetitions*" (McKee, 1997). One step further, the natures of them bring out another functional difference between them, which is screen dialogue serves as a bridge for linking creators and spectators, even it is performed by characters. On the

contrary, the natural conversation is for establishing and developing relationships between speakers.

Functions of Dialogue in Narrative Media

During the development of the audible and visual narrative-based entertainment content, the functions of dialogue have been kept tweaking for complying with cultural fashion and popular genres of content in some certain eras. However, there are still some principal functions centred in these creative and imitative dialogues for narrative. According to Bednarek (2017) and Kozloff (2000), the functions can be categorised into two classes, the ones fundamentally involved in the communication of the narrative:

1. Anchorage of the diegesis and characters
2. Communication of narrative causality
3. Enactment of narrative events
4. Control of viewer evaluation and emotions
5. **Character revelation (including character traits and character relationships)**
6. Adherence to the code of realism

and the ones relate to aesthetic effect, ideological persuasion, commercial appeal:

1. **Exploitation of the resources of language**
2. **Thematic messages/authorial commentary/allegory**
3. Opportunities for star turns

In the following chapters of this document, the functions in bold are the targets to be discussed, analysed, and experimented.

Chapter 3

Literature Review of Dialogue Generation

In this section, we introduce the related works from both conceptual and technical perspective of dialogue generation to discover and discuss the junction point between narratives and text realisation.

From the conceptual perspective, we will first introduce traditional dialogue generation approaches and systems, which are mostly used for basic task-oriented purposes. Furthermore, we investigate various routes, i.e. conditional dialogue generation approaches, through which dialogues can be enriched with reach higher-level goals, resulting in more variety and styles of the content of dialogues.

Secondly, we investigate dialogue generation from the technical perspective. In this sub-section, we will introduce the applications of deep learning techniques in dialogue generation, along with the pre-trained language models and datasets.

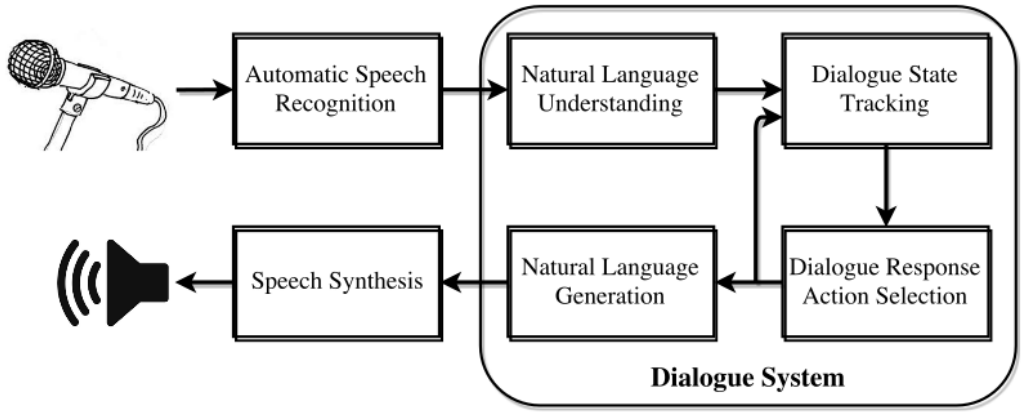


Fig. 3.1 A normal spoken dialogue system workflow (Figure from Serban et al. (2018)).

3.1 Dialogue Generation - Conceptual

3.1.1 Traditional Dialogue Generation

Traditionally, dialogue systems are designed and deployed in call centre application (Young, Gašić, Thomson and Williams, 2013) to reduce the workload of human operators and overall costs. Most of these systems are also called task-oriented (or goal-driven. For example, restaurant reservation (Huang et al., 2020; Williams et al., 2014), flight ticket booking (Bobrow et al., 1977) dialogue systems because of their clear and specific purposes, as well as have a similar workflow, which contains several components as Figure 3.1 shown.

It can be noticed that in Figure 3.1, *Automatic Speech Recognition (ASR)* and Speech Synthesis (e.g. *Text to Speech (TTS)*) are two components that are designed to convert verbal content into textual content, and vice versa. While the other four are the major components that process text and generate responses. Therefore, the so-called dialogue system refers to a system with these four components (Chen et al., 2017; Serban et al., 2018) as we are only focusing on text-based dialogue systems.

Jurafsky (2000) also named dialogue system using this workflow as a “frame-based dialogue system” as most of task-based dialogue systems use frame-based structure.

The dialogue system with this structure pre-defines one or more frames with slots and default values for each slot. Using this structure, the processes can be clearly abstracted and modularised for task-oriented dialogue generation. For example, in order to process a flight booking request from a user, the dialogue system is supposed to recognise some necessary information, such as the name of the city (departure and destination), the take-off date, and take-off time, which are set as *slots*. Once recognise the user's request, the system can then select the most likely dialogue action, a representation combined by speech act (e.g. inform, ask, confirm), as well as the type and value of the slot (e.g. time=19.00). To generate sentences for acquiring such information, templates including variables that can be used. For example, "*What time do you want to depart from [CITY-DEPT]?*", such sentence is targeting to acquire the desired time, as well as to implicitly confirm whether the filled slot [CITY-DEPT] is correct. The difficulties are mostly about detecting the current states correctly and conducting the next action correctly according to current states, while not about generating the actual text of dialogues. Although frame-based dialogue system is able to complete task-oriented dialogue, there are two limitations to be concerned. The first is there have to be different definitions of slots and values for different tasks. For example, the dialogue actions defined for restaurant reservation are not likely to be reused for flight booking as they need different information (e.g. table size and destination). Secondly, these systems tend to use template-based generation. Therefore, most of the words in the sentence need to be created heavily by designers' hands in advance.

The most important aim of these systems is to complete a certain task, which means that to correctly understand users' requests and correctly return the information with template-based generated sentences, of which the utterances are grammatically correct and logical. Since the goal of these dialogue systems is only to complete tasks and templates are leveraged, there is less necessity and ability to generate various dialogues in different formats. These systems have made the effort to address the

issue for the aspect of "what to say", i.e. to express and convey the correct content information in terms of grammars, syntaxes, topics, etc. However, the aspect of "how to say" is still in a rather fledgling state. In dialogue generation, this aspect is mostly relevant to the attributes of characters (e.g. profiles, personalities) who are the entities and assets involved in the turns of dialogue, affects (e.g. emotions, sentiments), and sometimes to the literal types (e.g. themes, genres). We will discuss the stylistic text generation in the next section.

3.1.2 The "Style" in Dialogue Generation

The word "style" is ambiguous that can be interpreted variously in different domains and definitions, as well as can be evaluated in different ways. There are many different aspects where the works of NLG have been focusing on over the past decades or so.

The classical understanding of the style of text is from the ancient world back to the 5th century B.C. which mostly concentrated on the rhetorical form. For example, Aristotle countered Plato's argument that rhetoric was mere sophistry, deceitful reasoning, by providing a system of instruction that treated the theory and practice of style as a valid discipline, designed to appeal simultaneously to reason and emotion (Cope and Sandys, 2010).

In the digital image processing community, image style transfer is a long-last hot topic. Given a normal picture that is shot by the camera, then the trained neural network system can produce new pictures with various typical styles of a specific famous artist, such as Van Gogh or Picasso, or of a specific drawing, such as sketch or comic (Johnson et al., 2016). Like altering pixel information for stylistic processed images, recent NLG works also generate stylistic text by representing textual information on various aspects.

It is natural to relate "style" in NLG to linguistic choices as the style in image processing is reflected with the stokes and colours. In NLG, various styles are re-

flected by linguistic features, which usually are referring to lexical, grammatical, or syntactical variation. DiMarco and Hirst (1993) develop a customised vocabulary of style by defining stylistic abstract and primitive elements in terms of grammatical and syntactical functions (e.g. balance, dominance, and position for abstract elements, as well as connective ordering and hierarchic ordering for primitive elements). They also develop a methodology to build up a stylistic grammar for correlating these stylistic elements with specific stylistic goals and defined rules. And then this stylistic grammar is interpreted into text using a specific parser. This is an example of a generation mode that achieves style by transiting from conceptual representation level to linguistic level directly (McDonald and Pustejovsky, 1985).

However, what causes these various linguistic features to appear in different text, or more specifically, in the context of NLG for dialogue generation, i.e. what cause stylistic dialogues, are more interesting and closer to the "style" discussed in this document, and are also more concentrated in recent NLG world. Many researchers consider the influence of speakers to be the source of dialogue style, as speakers are the direct agents who act or conduct the dialogues. Gatt and Krahmer (2018) argue that the main factors of speakers (individuals) behind the style variety are the personality and feelings, which is based on whether the factors are individually stable across time and transient every now and then respectively.

There are a lot work (Mairesse and Walker, 2007, 2010, 2011; Mairesse et al., 2007) focusing on achieving language generation variation with "Big Five" model (John and Srivastava, 1999) of personality traits, a standard in psychology. The authors of these works systematically explore and analyse the correlations between nearly exhaustive linguistic features and the 5 trait dimensions of the Big Five model (Extraversion, Emotional stability, Agreeableness, Conscientiousness, and Openness to experience). For example,

1. *Basically, actually, I am sure you would like Le Marais. It features friendly service and acceptable atmosphere and it's a french, kosher and steak house place. Even if its price is 44 dollars, it just has really good food, nice food.*
2. *Err... it seems to me that Le Marais isn't as bad as the others.*

The utterances above are collected from Mairesse and Walker (2007) which are generated on extravert set and introvert set respectively. As it shows, an utterance reflecting high extraversion might be more verbose and involve more use of expletives (1), compared to a more introverted style, which might demonstrate more uncertainty, for example through the use of stammering and hedging (2). So, the correlations between each personality trait and linguistic features are built up either derived from psycholinguistic findings, human judges or data-driven. And with these mappings, the system is supposed to generate various utterances with specified personality traits by tweaking the values of parameters of linguistic features. However, the mappings between psychological personality and linguistic features are possibly not clear or direct. And the fact exists that these personality-based styles are difficult to be perceived (Oberlander and Nowson, 2006; Youyou et al., 2015).

Another mainstream is focusing on the influence of feeling, or emotion, or affects of either speaker or listener involved in a dialogue. As the strategy of conducting personality-based styles, various emotional states can also be used to impact linguistic choices. The difference between these two categories of style is the personality of a person is relatively stable compared with emotions and affects that are more transient. To map human emotions to language features, researchers (Osgood et al., 1957; Russell, 1980, 2003) apply several influential factor analyses with three main dimensions,

1. dimensions of word meaning are **Valence** (e.g. positiveness/negativeness, pleasure/displeasure)
2. **Arousal** (e.g. active/passive, excitement/calmness)
3. **Dominance** (e.g. dominant/submissive, powerfulness/weakness)

which are usually called VAD lexicons. Colombo et al. (2019) use both categorical representation and continuous representation in a VAD space to model six basic emotions [1] proposed by Ekman et al. (1983), as well as apply this affect model to a GRU-based RNN[2] neural dialogue system to generate affect-stylistic responses according to the desired four emotions out of six. Buechel et al. (2020) introduce a methodology for creating almost arbitrarily large emotional lexicons for any target language, which aims to break the bottleneck that manually built-up lexicons contain limited lexical units and feature limit emotional variables. Huang et al. (2018) trained an LSTM-based emotional classifier for 9 emotions, which were used on a seq2seq dialogue system to generate dialogues expressing corresponding emotions.

Also, there are many other styles of interest with which it can do benefit to imitate the speaking types of specific categories of people. For example, to distinguish old people and young people (Hovy et al., 2020), to transfer expertise language into plate language for layman (Cao et al., 2020).

3.1.3 Conditional Dialogue Generation

Before the propaganda of deep neural approaches, " *'how to say' is often defined by simple templates or hand-coded rules which define appropriate word strings to be sent to a speech synthesizer or screen.*" (Lemon, 2008). Although the success of some generation systems (Cavazza and Charles, 2005) has proved the feasibility of this approach, there still exist limitations with this approach, of which the requirement of expertise for designing the templates and rules is the most significant one.

While since there are a few recent works about altering the style of utterances and sentences in various ways, it is evidenced that deep learning approaches with big datasets are promising for generating stylistic narrative dialogue.

[1] Anger, disgust, fear, joy, sadness, and surprise.
[2] **R**ecurrent **N**eural **N**etwork

Wen et al. (2015) propose a semantic controlled LSTM cell by adding a dialogue action cell to a traditional standard LSTM cell (Hochreiter and Schmidhuber, 1997). By this model, utterances in natural language can be generated based on the dialogue action vector representation and 12 attributes in the restaurant domain and hotel domain.

Recently, various additional features or attributes are incorporated into deep neural networks in order to alter or control the generated sentences, such as speaker profile (Dong et al., 2017; Li et al., 2016; Zheng et al., 2020), big-five personality (Herzig et al., 2017; Oraby et al., 2018; Xu et al., 2020), sentiment or emotion (Ficler and Goldberg, 2017; Ghosh et al., 2017) and tense (Hu et al., 2017). Although these features are not applied in narratives and they are used for altering the styles within a single sentence, we believe these works could be good references for our research project since we have good potential to reuse some of these features in narrative context and extend the scope of effect from a single sentence to a conversation.

3.2 Dialogue Generation - Technical

As mentioned in previous parts, most computational narrative systems utilise plan-based approaches with hand-coded rules to generate stories and discourses, which remain beset by issues of low efficiency, and with difficulty to scale up and limited variety.

3.2.1 Neural Network and Dialogue System

Deep neural networks, consisting of numerous small computing units which take a vector of input values and produce a single output value, has been a powerful technique in the AI community in the last years. Particularly, in natural language fields, it is flourishing and helping researchers reach huge achievements in many different tasks,

such as machine translation (Cho et al., 2014; Luong et al., 2015), Summarisation (Yu et al., 2017), and Text (Dialogue) Generation (Li et al., 2016; Wen et al., 2017).

"sequence to sequence (seq2seq)" model, also called "encoder and decoder" (Figure 3.2), has proven its success in machine translation task (Bahdanau et al., 2015; Cho et al., 2014; Sutskever et al., 2014; Vinyals and Le, 2015). In this model, an original variable-length sentence is fed into the encoder, and is encoded into a shared vector representation. This shared vector then is decoded into an output vector representation, which will be compared with the target vector of the target sentence, in order to find the mapping between the input and output by updating the parameters in encoder and decoder over iterations. There are two major advantages of this model. One is that it is an end-to-end model which requires much fewer hand-crafted rules. Also, it has the capacity to generate sentences with unfixed length.

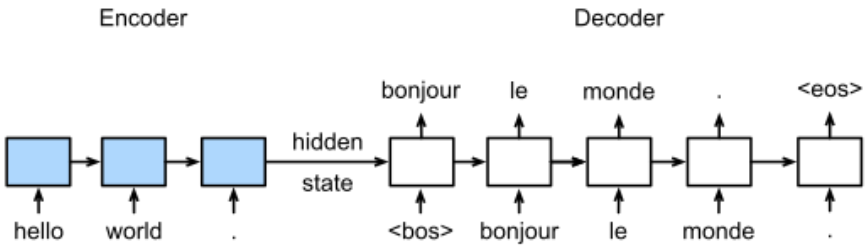


Fig. 3.2 A seq2seq model for machine translation (Figure from Zhang et al. (2021)).

On the basis of the seq2seq model, some additional mechanisms have been incorporated into this paradigm. For example, the attention-based mechanism has been a standard mechanism for seq2seq, which assign more weights to parts of encoded input vectors when predicting certain portions of the output vectors during decoding (Bahdanau et al., 2015; Luong et al., 2015). Attention-based is widely used in natural language processing/generation and has been proven effective (Yin et al., 2016), because it improves the performance of processing long input sentences by leveraging dynamic context vectors, which contain dynamic weights for input tokens, for each decoding step with the aligned position. Compared with the plain seq2seq which uses a stationary context vector for all decoding steps, this mechanism can help the system

"memorise" longer sentences without "forgetting" the beginning of the sentences, as the word "attention" denotes.

Unlike classical seq2seq, which mostly uses RNN computing unit (LSTM or GRU[3]) along with attention-based mechanism, Vaswani et al. (2017) proposed a "transformer" model which entirely built on the self-attention mechanisms without using sequence-aligned recurrent architecture. This model solved the constraint of sequential computation that RNN owns inherently.

Although seq2seq is designed for machine translation initially, it is commonly used for dialogue generation as well, as the processes of these two tasks are similar to each other. Serban et al. (2016) extend the hierarchical recurrent encoder-decoder neural network to improve the end-to-end dialogue system, which decreases the dependency of massive hand-crafted features and unlock the restriction of the applicable domain like the earlier works using partially observable Markov decision process (POMDP) models (Gašić et al., 2013; Young, Gašić, Thomson and Williams, 2013)

3.2.2 Pre-trained Language Model

As this thesis is about dialogue generation, the process of which can be approximately described as predicting upcoming words from prior context. This description is also commonly viewed as the definition of the conditional language model, where output is generated by sampling tokens (normally words) from a distribution conditioned on input features, which could be semantic, contextual, and stylistic (Gatt and Krahmer, 2018).

Word Embedding

Feedforward neural network language model (NNLM) (Bengio et al., 2003) is an early introduced language model and is treated as a prototype of neural language models.

[3] **G**ated **R**ecurrent **U**nit (Chung et al., 2014)

However, it has some limitations, such as the need to specify a fixed sliding window as input context, and the lack of ability to represent complex patterns. To overcome such limitations, improved language models based on recurrent neural network (RNNLM) were proposed (Bengio and Lecun, 2007; Mikolov et al., 2011). The key difference between NNLM and RNNLM is that NNLM only takes the token vectors in the current sliding window as the input context, while RNNLM takes the input context containing not only the current token vector but the hidden states that includes the information of previous time steps (i.e. from current token back to the beginning of sequence). This modification does improve the performance of RNNLM in spoken and written language although it is difficult to reason about (Jurafsky, 2000).

Many works focusing on language model build-up benefit from NNLM and RNNLM, and word embedding can be treated as a byproduct along with the whole language model [4]. The trained word embedding, or the trained parameters of the encoding layer, which like a small simple linear language model, can be reused for initialising the encoding layer in other NLP tasks and improve their performances. Moreover, some works modify the optimisation objective with a technique they called "Negative Sampling" (Mikolov, Chen, Corrado and Dean, 2013), avoiding training all parameters of layers but only the encoding layer. Mikolov, Chen, Corrado and Dean (2013) proposed Word2Vec with two different training strategy variants, Skip-gram, and CBOW[5]. With the research of creating effective word embedding, the relations between words are explored and are able to be vectorised. For example, it is found that the words of country-capital pairs have similar distances and angles illustrated in vector space (Mikolov, Sutskever, Chen, Corrado and Dean, 2013), which means word embeddings are able to learn and represent the relationship between words. Another

[4]Because language models are trained to predict the next word based on previous words, in which each input word is encoded (e.g. one-hot encoding) and then multiply an initially random trainable matrix Q as the context for further computation in the hidden layer(s) and softmax layer. Therefore, when the training process is completed, we not only have a whole language model with trained parameters of each layer, but also obtain a trained matrix Q, which contains the vectorised representation for each word, a.k.a. *word embedding*.

[5]**C**ontinuous **B**ag **o**f **W**ords Model

example GloVe[6] (Pennington et al., 2014) improves Word2Vec by efficiently leveraging statistical information of global co-occurrence counts rather than local context window methods. However, the results of these works are static word embeddings, which means in some situation where polysemy appear, they cannot switch the vector representations for the correct meanings. To solve this issue, Peters et al. (2018) propose ELMo[7] that uses both semantic information of words and the contextual information before and after the words on the architecture of bidirectional LSTM.

Advanced Pre-trained Language Model

Most of the above early generation of pre-train methods is based on the unidirectional language model theoretically and/or on RNN (e.g. LSTM or GRU) technically. After transformer architecture (Vaswani et al., 2017) has been proposed, it has been proved that the language models trained with transformer have significantly achieved better performance than ones on RNN (Devlin et al., 2019; Radford et al., 2018, 2019). Because the transformer is able to process much longer sequence with self-attention mechanism, and performs better in parallel computing than RNN due to its own recurrent nature. Also, applying the masked bidirectional language model in the training process has also achieved better performance than the unidirectional one (Devlin et al., 2019; Radford et al., 2019). Because with the unidirectional language model, only the context before the target word is encoded for training, i.e. predicting the next word according to prior words or the traditional language model. However, encoding both the context before and after the masked word allows transformers to predict the target word more correctly as the semantic meanings of some words do not only depend on the prior context but also subsequent context. Moreover, with much more massive corpora used and rapid development of hardware, the new generation of language models can be

[6] **Glo**bal **Ve**ctors for Word Representation
[7] **E**mbeddings from **L**anguage **Mo**dels representations

designed more complexly with more layers and blocks, and be capable of solving more generic NLG and NLP tasks.

All of the above upgradations and improvements make these advanced language models have a huge impact on the field of NLP, and are now central to most NLP systems and research (Bommasani et al., 2021). Clark et al. (2021) demonstrate that non-experts have difficulty distinguishing short-form English text that was written by GPT-3 (Brown et al., 2020) from that written by humans.

Such advanced pre-trained language models are not just vectorised word representations, like word embeddings, but backbone-like language models which can be adapted and transferred to solve multiple different downstream NLP tasks in various domains, with a minimal number of parameters need to be learned from scratch. Because these language models are pre-trained on corpora with an enormous amount of text in the real world, which makes them "*store the fundamental knowledge that closely represents the state of the world* (e.g. the lexical, grammatical, syntactical knowledge), *independent of modality*" (Bommasani et al., 2021). Devlin et al. (2019) introduce that BERT[8] can be easily adapted for single sentence classification, sentence pair classification, and single sentence tagging tasks.

OpenAI's GPT-2[9] (Radford et al., 2019) was trained simply to predict the next word in 40GB of Internet text, which has demonstrated that transformer models trained on very large datasets can capture long-term dependencies in textual data and generate text that is fluent, lexically diverse, and rich in content (Samples generated by GPT-2 collected from OpenAI blog[10] presented as Table 3.1).

Many achievements have been noticed on the basis of the pre-trained GPT-2 language model in recent years (Mao et al., 2019; Yang et al., 2020; Zhang et al., 2020; Zheng et al., 2020). Zhang et al. (2020) present DialoGPT, which models a multi-turn dialogue session as a long text and frame the generation task as language modelling on

[8] **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
[9] **G**enerative **P**re-trained **T**ransformer 2
[10] https://openai.com/blog/better-language-models

Table 3.1 A partial samples of text generation by GPT-2 from OpenAI

SYSTEM PROMPT (HUMAN-WRITTEN)	**A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.**
MODEL COMPLETION	The incident occurred on the downtown train line, which runs from Covington and Ashland stations. In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief. "The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation," said Tom Hicks, the U.S. Energy Secretary, in a statement. "Our top priority is to secure the theft and ensure it doesn't happen again." ...

an enormous dataset extracted from Reddit. According to Zhang et al. (2020), DialoGPT is able to generate more relevant, contentful and context-consistent responses. Yang et al. (2020) inherited DialoGPT with KL(Kullback-Leibler) loss and style classifier to generate responses towards the target styles in both word-level and sentence-level. Particularly, they construct two style corpora with arXiv-style and Holmes-style, which are trained on for providing language styles implicitly along with the plain conversation dataset from Reddit. Zheng et al. (2020) also present a personalised dialogue generation model on the basis of a pre-trained language model and fine-tune it on a persona-sparse dataset, as they argue that a large scale part of dialogues is not persona-related (Their persona consists of gender, location, and personal interest, or profile). Further, the GPT-2 language model is also leveraged for story generation by Mao et al. (2019). They use auxiliary training signals from datasets designed to provide common sense grounding, and achieve quantitatively better common sense reasoning in language models.

3.2.3 Corpora and Dataset

In recent years, the trend of dialogue system research has been focusing on data-driven approach greatly, compared to the previous systems built up with rule-based

approaches and expert knowledge. If we view an advanced dialogue system like a car, and view the deep neural networks as car body and engine, then it is making sense that datasets or corpora would be the fuel for the car. The fast developing deep neural networks are upgrading their ability of feature extraction and computation. However, like a high-performance car needs not only a solid car body and powerful engine but also high-purity fuel that matches the car, the performance of a data-driven dialogue system also relates to the quality of its fuel – dataset, or corpora, and even more so than the fuel in car analogy.

Although categorising such a huge amount of existing corpora is difficult, it is still necessary to do so in order to select the most appropriate one for the given NLP/NLG task. Regarding the target of this document, I will focus on dialogue corpora particularly and discuss them from dialogue source aspects of interest referring to the work of Serban et al. (2018).

Screen Dialogue

The corpora with screen dialogues, or scripted dialogues, are effective alternatives to the ones with natural conversations in data-driven deep dialogue system learning. Because screen dialogues are sourced from either film screenplays, TV series manuscripts, or novels, these dialogues are created with rich dramatic information. And they are created, adapted and revised by the hand of expertscript writers and novel authors, which guarantees the quality of the correctness of the usage of language (e.g. the grammar correctness), as well as the consistency of the content.

There are a lot of existing corpora based on film and TV series that are widely used in NLP/NLG tasks. Among them, the **OpenSubtitles** (Lison and Tiedemann, 2016) is a primary corpus, which are based on the OpenSubtitles website[11]. The OpenSubtitles collects an enormous amount of film dialogues with around 140 million utterances and 1 billion words, as well as multiple language support. These two features make it

[11]https://www.opensubtitles.org/

an appropriate dataset for neural machine translation (NMT) tasks. However, because the utterances in the OpenSubtitles are not speaker-aligned, it is lack of potential to serve for character-based stylistic related dialogue generation.

Apart from the textual information, the utterance text, there are many additional features as they deliver information as well. The **Cornell Movie-Dialogue Corpus** (Danescu-Niculescu-Mizil and Lee, 2011) provides not only the dialogues but also various metadata on different levels. In this corpus, speaker, film name, genre, release year and more metadata are provided, which makes it possible to learn specific types of dialogues based on these features. Some work has been done for certain additional features, such as character-related ones. Walker et al. (2012) proposed the **Character Style from Film Corpus** collected from IMSDb[12] archive. In this corpus, Features, such as the sentiment behind the utterances, are automatically extracted and used to build up models on different characters, which are capable to generate new utterances with similar styles to those spoken by the character.

Some corpora are focusing on the dialogue structure, such as **Filtered Movie Script Corpus** (Nio et al., 2014) and **Movie-Triples** (Serban et al., 2016) Both of which filtered the collected subtitles and structured them as X-Y-X triples, where X is spoken by one actor and Y by another, and each of the utterances shares some semantic similarity.

Because of its nature, the corpus with screen dialogues only provides the textual information in films or TV series, in which much more visual information cannot be observed, such as the performance of characters and scenic settings. Also, the differences between scripted dialogues and real conversation cannot be ignored. Nevertheless, it can still be helpful for data-driven dialogue learning since movie dialogues are more compact, follow a steady rhythm, contain less garbling and repetition, and convey information clearly to the viewers or readers (Dose, 2013; Forchini, 2012).

[12]https://imsdb.com/

Natural Conversation

Broadly, all the conversations that happen in the real life naturally, other than the ones created intentionally, are supposed to be classified into this sub-section. However, there are still finer branches according to the ways that are used for collecting natural conversation data. Serban et al. (2018) notes that the distinction between the spoken and the written conversation is important. Because for the spoken conversations, which are mostly recorded by real people and then transcribed into text, they are "*interpersonal, situation-dependent has no narrative concern, belonging to a highly interactive, situated and immediate text type*" (Forchini, 2012). Also from the linguistic aspect, spoken conversations contain more pronouns and modal particles, as well as tend to use shorter words and phrases.

On the other hand, the written conversations are mostly collected from online social network websites, such as forums, blogs or micro-blogs, and chat-room. Because in such ways, conversations between interlocutors do not happen simultaneously, then users are able to reflect on what they are writing before they post a message and thus are more precise. However, there are still differences among these social networks, for the ones that are more replied on instant message exchange, like Twitter[13], or some chat-rooms, conversations have more similarities to the spoken conversations as these media intend to create engaging online scenes for imitating the occasions where real spoken conversation happens.

3.3 Background Conclusion and Summary

Contemporary entertainment content has been becoming a vigorous and developing medium for the human to tell stories, which offers abundant and workable elements and plots of narrative in multi-modal ways (such as visual and textual). Narrative systems have explored a wide range of approaches to both generating and presenting

[13]https://twitter.com/

narratives (Riedl and Young, 2003) and have experienced mixed results (Hargood, 2011). However, most narrative systems are focusing on the narrative structures and narrative events generation, or the so-called plot, but give little attention to text realisation, including by the means of dialogue, as pointed out by some (Cavazza and Charles, 2005; Gatt and Krahmer, 2018).

Meanwhile, dialogue generation is a speedy up-trending research branch in natural language processing in recent decades with the support of deep neural network models and big data technology. However, the influence on the dialogue generation by incorporating additional information from the authorial aspect into neural networks has not been sufficiently explored. Considering the successes of the dialogue generation research have achieved so far, such deep learning algorithm and rich corpora can be expected to not only generate natural conversations in real life, but also for generate dialogues in narratives, as well as contributing to enrich the ways of representation and realisation of narrative structures and story plots.

In the next chapter, we explore an approach for data-driven narrative dialogue generation on a customised dialogue corpus built on film screenplays.

Chapter 4

Personalised Dialogue Generation

As we mentioned in Chapter 1, we are intending to investigate the impacts that characters with different attributes make on the dialogue generation in the fictional world. And particularly, we select the personality of a fictional character as the main factor among the attributes of a character according to literature review (Chapter 3).

Therefore, we are introducing the workflow of our approach , including the model of personality we choose to apply, the process of parameterising personality during the building of our dataset, the structure of neural network adapted, and the ways to incorporate personality into neural network.

4.1 Character Personality

4.1.1 Personality Model

There are many existing psychological models that describe human personality from different perspectives, for example, Big-Five Personality Traits, Myers-Briggs Type Indicator (MBTI), Enneagram of Personality, 16PF Questionnaire, etc.. Among these models, Big-Five model and MBTI are two models that are deeply investigated and have been applied in language generation using deep learning techniques.

Myers-Briggs Type Indicator uses four categories to describe personality type, with dichotomies for each category:

- **E**xtraversion/**I**ntroversion
- **S**ensing/i**N**tuition
- **T**hinking/**F**eeling
- **J**udging/**P**erceiving

This model is originated from Swiss psychiatrist Carl Jung's observation and speculation that human beings experience the world using principal psychological functions (Jung, 2016). The model provides 16 preferred personality types as each category has two polar orientations (the bold letters denote the polar orientations, for example, ISTJ means Introvert, Sensing, Thinking, and Judging). Keh et al. (2019) tested the performance personalised generation by embedding 16 MBTI personality types using BERT pre-trained language model, and achieved training losses around 0.02 overall. Although this model is used in career counselling programmes , it is under criticism for various reasons. Firstly, Jung's original theory is based on his clinical observation rather than controlled scientific studies, which might be an explanation of some dimensions of MBTI is correlated (**J**udging/**P**erceiving and **S**ensing/**I**ntuition). Also, the results of test-retest indicate the reliability of MBTI is tending to be low (Pittenger, 1993).

Big-Five personality traits are derived from trait theories in psychological research (Allport, 1961). Although there are many traits that can be used to describe personality with habitual patterns of behaviour, thought, or emotion, it is widely acknowledged that these traits can be reduced into 5 major traits using factor analysis:

- Extraversion
- Neuroticism
- Agreeableness
- Conscientiousness

- Openness to experience

Also, the results of factor analysis on personality survey data reveal some semantic associations, i.e. some linguistic expressions are often used to describe aspects of personalities of the same person. John et al. (1988) introduced the definition of lexical hypothesis by elaborating prior works, which is:

> *Those individual differences that are most salient and socially relevant in people's lives will eventually become encoded into their language; the more important such a difference, the more likely is it to become expressed as a single word.*

Although there are some disagreements about division of the 5 traits in the Big-Five model, such as criticism for orthogonal structure exists between traits, the relation between lexical options and personality descriptions has been commonly agreed (Goldberg, 1993). A compendium of Allport and Odbert (1936) includes approximately 4,500 trait-descriptive terms in natural English. All these researches build up the methodological foundation or rationale of the research of Big-Five personalities by the means of lexical descriptors in natural language. However, using this kind of verbal descriptors to present personality is likely to introduce inevitable limitations. To sum them up, they are all about the interpretations of lexical descriptors for personality may vary across language communities or cultures, or between scientific researchers and laymen, as well as may change over time.

4.1.2 Personality Definition

According to narrative theories (Dyer, 1989; McKee, 1997), the personality of characters in a single narrative is set by the author and should remain consistent throughout. For example, "*A character's personality in a film is seldom something given in a single shot. It has to be built up ... across the whole film. A character is a construct from*

the very many different signs deployed by a film", writes Dyer (1989). Therefore, we hypothesise that the personality of a certain character in a film remains consistent. Given the research progress of personality descriptor and natural language (Allport and Odbert, 1936; John et al., 1988; John and Srivastava, 1999; Mairesse and Walker, 2011; Mairesse et al., 2007), we use the properties of the Big-Five model to define characters' personalities by their dialogues presented in textual form. As mentioned before, the big-five model contains 5 major personality traits. It is easier to memorise these 5 traits by combining the abbreviation of trait names into a word. Therefore, there are some aliases of the big-five model, such as the "OCEAN" or "CANOE", which are widely acknowledged and used. Also, the name and the abbreviation of the same trait can be selected differently. Particularly, the trait "Emotional Stableness" in the Big-Five model refers to trait "Neuroticism" in "OCEAN" model, which are two polar orientations of the same dimension of personality.

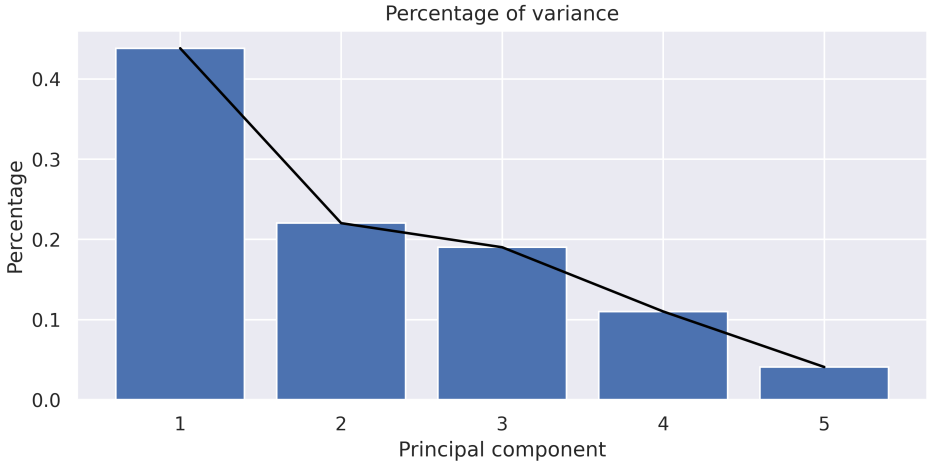


Fig. 4.1 The percentage of the explained variance of each original trait dimension.

4.1.3 Personality Recognition

We use a *personality recognizer* (Mairesse et al., 2007) to calculate the personality of characters in films. Its core foundation is based on the correlations between linguistic variables and personality traits in different linguistic levels, such as lexicon and syntax.

For example, compared with an introvert person, an extravert person tends to use language with more diversity on lexical level, and with more verbs on syntax level. Mairesse et al. (2007) trained their personality recognizer models on two corpora, which are essay corpus (Mehl et al., 2006) in written language and EAR (electronically activated recorder) corpus (Pennebaker and King, 1999) in spoken language. The essay one contains self-assessed personality, while the EAR one contains the personality observed by external judges. In this thesis, we use the model trained on EAR corpus because the models of observed personality outperform those of self-assessed personality, according to their experimental results, as well as the dialogues in films is a type of spoken language.

Fig. 4.2 The matrix of eigenvectors. The numbers in figure denotes the correlation coefficients between the new primary components (row) the original 5 personality traits.

In order to analyse the data of personality, we use Principal Component Analysis (PCA). We first calculate the personality scores of each character on all five traits using personality recognizer, and standardise these scores by removing the mean and scaling to unit variance on each trait. Then we conduct PCA and show the results in Figure 4.1 and Figure 4.2. In Figure 4.1, we notice that the first three dimensions take around 86% of the variance ratio, which indicates that the trait

dimension of personality can be reduced from 5 to 3[1]. And Figure 4.2 shows the matrix of correlation coefficients between the original 5 traits and the new 3 primary components. Normally we need to construct new 3-dimension data with this matrix. However, this will include mixed new traits that are difficult to be interpreted, which will further affect our evaluations negatively. Therefore, we select 3 traits from the original 5 trait with the largest coefficients. It is easy to be aware that **Extra**version relates to the new primary component *2* and **Emoti**onal stability relates to the new primary component *1* dominantly, with the numbers of 0.93 and 0.8 respectively. Although the new primary component *0* relates to **Consc**ientiousness with the highest number of 0.56, we do not select this trait as the third trait in our study although the correlation coefficients of **Agree**ableness and **Openn**ess are slightly lower than it. There are two main reasons for this selection: Firstly and essentially, Mairesse et al. (2007) reported that MRC[2] (Coltheart, 1981) features are the most important for classifying trait **Consc**ientiousness for spoken language. However, during calculating the personality scores for utterances, we noticed that there are 5 MRC features that cannot be calculated, which take up 36% of total MRC features (14 in total). Thus, the personality scores for trait **Consc**ientiousness are likely to be impacted negatively. Secondly, in Mairesse et al. (2007), we noticed that the linguistic LIWC (Pennebaker, 2001) markers that correlate to trait **Agree**ableness are similar to those correlate to trait **Consc**ientiousness (also reported in Mairesse et al. (2007)). Although it is possible to view the new primary component *0* as a mixture of multiple traits as the correlation coefficients are not dominant (3 out 5 are from 0.46 to 0.56) as the other two new components, we believe it would bring us more difficulties to interpret a new mixed trait. Therefore, regarding the Figure 4.1 and Figure 4.2, we select **Agree**ableness as the third primary trait according to these reasons and we believe to select individual

[1] Another reasonable option is to reduce the trait dimension from 5 to 4. Here we select the simpler one.

[2] A computerised database of psycho-linguistic information. Semantic, syntactic, phonological and orthographic information

Table 4.1 Character personality matching rate between the overall personality score in film-level and the average score of scene-level.

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

traits as original is able to maintain the initial definition meanings of the Big-Five personality model.

4.2 Dataset

4.2.1 Why Not Use Existing Corpora?

In Chapter 3.2.3, we have discussed existing corpora from the perspective of the differences between screen dialogue and natural conversation. Considering the research goals and the features of existing corpora, we recognise the necessity to create our own corpus to overcome the limitations and inaccuracies of existing ones.

As we are focusing on dialogue generation in the context of narrative film, it is reasonable and necessary to narrow down our interests to corpora with screen dialogue. Some existing corpora with screen dialogue are collected from films and TV series. These corpora only contain textual information of dialogue, i.e. the utterances or sentences themselves (Lison and Tiedemann, 2016), or contain very few other

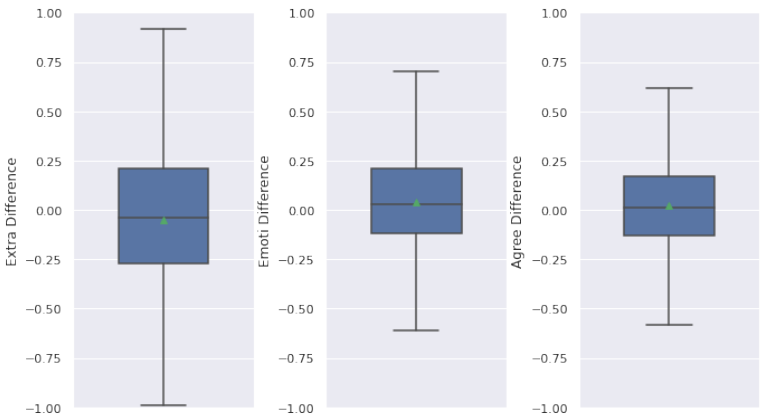


Fig. 4.3 Character Personality difference between overall personality score in film-level and personality scores in scene-level.

basic metadata as well, e.g. speaker and film name (Danescu-Niculescu-Mizil and Lee, 2011). However, for the dialogues within the context of the narrative, these kinds of information are not able enough to represent narrative characteristics. Therefore, we aim to build up our own corpus from screenplays and include features that represent narrative characteristics. For example, we aim to retain the structure of a dialogue rather than just treat dialogue as a start utterance and its responses simply.

Besides, on the aspect of conditional dialogue generation, the additional "conditions" appear more interestingly and importantly than the textual information of utterances. We observe that some work focuses on conditional dialogue generation use character persona as an additional feature. Such persona commonly contains the information of speaker's gender, age, job, and the location (Li et al., 2016; Qian et al., 2018; Zheng et al., 2020), which we call "*profile*". Such features are able to help dialogue systems to generate sentences with corresponding "profile" in an explicit way but not affect the style of linguistic expression in essence. Also, the other additional features, such as sentiment and emotion, are labelled on individual and independent pieces of text (e.g. a tweet feed) (Colombo et al., 2019; Klinger et al., 2018), which lack of overall characteristic of a speaker. We acknowledge the improvements of dialogue generation with these features. However, these features are not fit our requirements properly, as we aim to generate dialogues with overall character personalities which are related to linguistic expression and reflect authorial intentions.

4.2.2 Corpus Build-up

The dialogues are parsed from raw screenplays on IMSDb by following Winer and Young (2017), in which they pointed out that rich narrative knowledge can be extracted from screenplays, as "*screenplays contain more structured discourse information* (Jhala, 2008) *than other narrative texts such as news stories* (Chambers and Jurafsky, 2008) *or fables*".

A screenplay is supposed to be written in formalised rules, which means it is feasible to identify elements in screenplays with certain typesetting patterns (Riley, 2009). For example, in Figure 2.2, the name of a character is supposed to be typed with all capital letters, and is placed in the centre of the page with a fixed indentation. Therefore, it is reasonable to extract dialogues, dialogue structures, and characters from original screenplays by utilising these formalised patterns.

Winer and Young (2017) created a grammar based on the rigid syntax of shot headings to extract short descriptions (e.g. stage directions) and other discourse elements for automatic screenplay annotation. The main strategy that their annotator uses for identifying these elements is to segment elements by different indentation and unique formats for certain elements. For example, each line in screenplay (supposed to be 58-character length) is divided into three blocks, which are *RANGE_LEFT*, *RANGE_MID*, and *RANGE_RIGHT*, with 18-character, 24-character, and 16-character length respectively. And then the annotator is supposed to identify a piece of text in a line is a name of a *Character* if the first letter of this text locates within *RANGE_MID* with all capitalised letters. This rule-based on indentation is able to identify characters' names correctly if a character's name is placed in a standardised position. However, we observed that the raw screenplays from IMSDb are not with a standard format completely, such as varied positions for character's name (which might be in all three blocks they defined), or wrong indentations of *dialogue* text (supposed to be with different indentations from *character*). These wrong formats can cause mistaken identification for elements in screenplays, especially for dialogues.

Therefore, we create a heuristic-based process of screenplays to recognise and segment a series of elements with the strategy of *Finite State Machine (FSM)* referring to the work of Zhang et al. (2019). Different from the strategy of Winer and Young (2017) that relies heavily on indentation, we only use conditions based on the format to identify characters' names, as well as dynamic rather than fixed indentation to identify the start and end of each utterance. With these two main strategies, we are

able to process some frequent situations more correctly where character-utterance pairs with wrong formats appear as follows:

- Varied and wrong indented character and/or utterance.
- An unexpected empty line between character and utterance.
- No necessary empty line for separating the utterance spoken by previous character and the next character's name.

Table 4.2 Overall characteristics of the parsed IMSDb dataset.

Genre	Film	Dialogue	Turn
Action	315	96k	213k
Drama	631	197k	502k
Romance	203	62k	164k
Thriller	389	113k	274k

Corpus Statistics

Initially, We collected 1,160 screenplays from IMSDb, and use our heuristic-based script to parse these screenplays. We discarded some screenplays from which no dialogue can be parsed, because they contain little or no formatting information such as no capitalised words, which makes it impossible to distinguish characters from other elements. Consequently, we parsed 1,064 screenplays, which is around 92% of the screenplays we collected. Each screenplay is parsed into dialogue sessions based on transitions[3] within each film, along with corresponding character (speaker) for each utterance, and categorised by genres. Specifically, in our following experiment, each big dialogue session is segmented into several dialogues for training, ensuring each dialogue is always taking place between two characters. We make this simplification because "A-B-A-B" exchange is the simplest and most common dialogue structure in film screenplays. In a dialogue session, descriptor *CONT'D* from screenplay before a turn might appear, which denotes this turn is continued to the last one spoken by the

[3]We recognise a transition by some special staging directions, e.g. when the locations are switched.

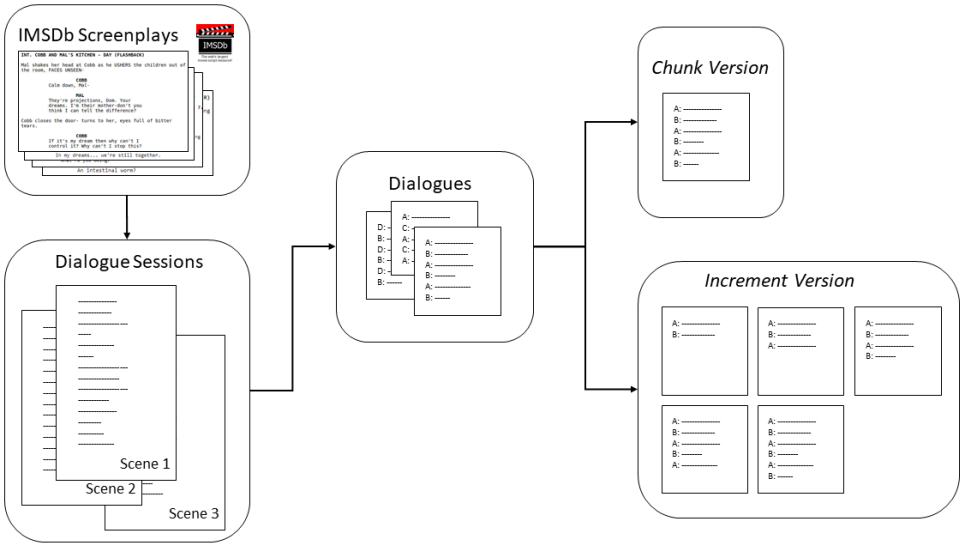


Fig. 4.4 The steps in building up our corpus.

same character. Here two strategies are used for processing this: *discarding* one turn or *concatenating* two turns. both ways result in the same number of turns.

Also, all parsed dialogues are extracted from the screenplays on 4 major narrative-rich genres: Drama, Romance, Action, and Thriller. (Table 4.2 shows the original dataset's characteristics).

In this document, we select dialogues composed of between 2 to 6 turns, and extend them to 2 versions of dialogue set: *chunk* and *increment* (as Figure 4.4 shows). For both sets, we keep the original structure of selected dialogues. Particularly for *increment* one, for each selected dialogue, we expand it into $n-1$ sub-dialogues, where n denotes how many turns in such selected dialogue. We use *increment* set with the statistics as Table 4.3 shows. Yang et al. (2020) use ~1M sentences of arXiv-style and ~38k sentences of Holmes-style for training stylistic dialogue generation. Here our dataset is with a not less magnitude of size[4]. We split it into the training set and the evaluating set with ratios 90% and 10% respectively, to conduct all the experiments and studies in this document.

[4]One turn consists of at least one sentence.

Table 4.3 Characteristics of the selected and parsed IMSDb dataset (*increment*) used in this study.

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

Table 4.4 Character personality matching rate between the overall personality score in film-level and the average score of scene-level.

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

4.2.3 Labelling Personality

We label the parsed dialogues with the selected three traits (**Extra**version, **Emoti**onal stability, and **Agree**ableness, or with abbr. *Extra*, *Emoti*, and *Agree*) regarding the results of PCA. To start labelling, we first calculate the personality scores on three traits for each character using personality recognizer (Mairesse et al., 2007). scores here are calculated from all the utterances spoken by a single character from a complete film, representing a *film-level overall* personality score for this character. They are within the range from 1 to 7, with 4 as the neutral following Mairesse and Walker (2011); Mairesse et al. (2007). And we then divide them 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[5]. For instance, for the extraversion trait, the label *Low* denotes more introvert and the label *High* denotes more extravert. We set the thresholds identically for each trait, for maintaining the original distribution from the raw data as they are already scaled by Personality Recognizer. We then label each utterance in the dialogue session with a 3-trait personality pertaining to the character who utters this sentence. For example, a sentence is assigned with the label (*High*, *High*, *High*) if the character utters this very sentence is extravert, emotionally stable, and agreeable.

[5]We also use another threshold scheme in which *Medium* between [3.5, 4.5], with *Low* smaller than 3.5 and *High* greater than 4.5.

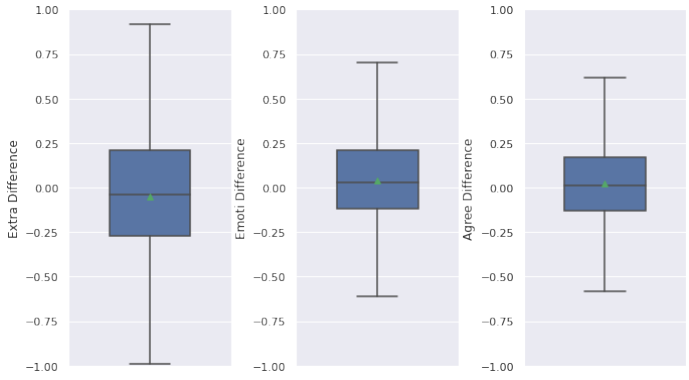


Fig. 4.5 Character Personality difference between overall personality score in film-level and personality scores in scene-level.

During the investigation of the overall personality, we noticed the existence of volatility in term of recognised personality scores over the progress of a film based on the aforementioned calculation. We did the same calculation and label work on the 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. In order to compare the differences between film-level personality and scene-level personality of each character, we firstly calculated a weighted normalised 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 matching degrees on the trait labels between film-level personalities and scene-level personality averages. Table 4.4 shows that there are 79% of characters who have at least 2 out 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 4.5).

We set two versions of corpus: *Corpus 1* as a pilot corpus using *discarding*, [1, 3.5, 4.5, 7] scheme, and first *chunk* in a dialogue session; while *Corpus 2* as a final corpus using *concatenating*, [1, 3.8, 4.2, 7] scheme, and *increment* set [6]. For each corpus and each genre, characters' personality scores are calculated and their medians are compared. For example, for label *High* of Extraversion, if the median of scores in *Corpus*

[6]In the following evaluations, we use *Corpus 1* for User Study i, and *Corpus 2* for the others. In this document, "corpus" refers to *Corpus 2* unless specified.

2 minus that in *Corpus 1* is positive, which denotes *Corpus 2* has more percentage of extreme scores for this trait, otherwise the *Corpus 1*. For 8 ($2 \times 2 \times 2$) personalities with *High* and *Low* labels, we have 96 ($8 \times 3 \times 4$) median differences. Among them, *Corpus 1* and *Corpus 2* have 51 (avg. 0.698, std. 0.48) and 45 (avg. 0.736, std. 0.58) traits respectively of which the medians greater than the other, i.e. more extreme. We consider these two corpora have similar percentage in extreme personality with 0.72 of p-value, while *Corpus 2* has better parsed dialogue structures, more utterances with extreme personalities, and more amount of utterances.

4.3 Approach

4.3.1 Problem Formalisation

Our aim is to generate utterances that corresponds to a given dialogue context and a representation of composite target personality traits. This aim can be decomposed into two sub-aims, which are:

1. to generate contextually consistent utterances.
2. to incorporate target personalities that is able to be presented through generated utterances correctly.

The overall workflow of our approach is illustrated as Figure 4.6.

Basic Language Model

We have introduced that the state-of-the-art neural language model has proven its success in generating contextually consistent responses with given context for various NLP tasks in Chapter 3.2.2. The traditional language model using N-gram estimates conditional probabilities by counting the ratio between a certain N-gram and other N-gram with the same context history (input word sequence) from a corpus. Compared with this N-gram based language model, the neural language model is able to extract

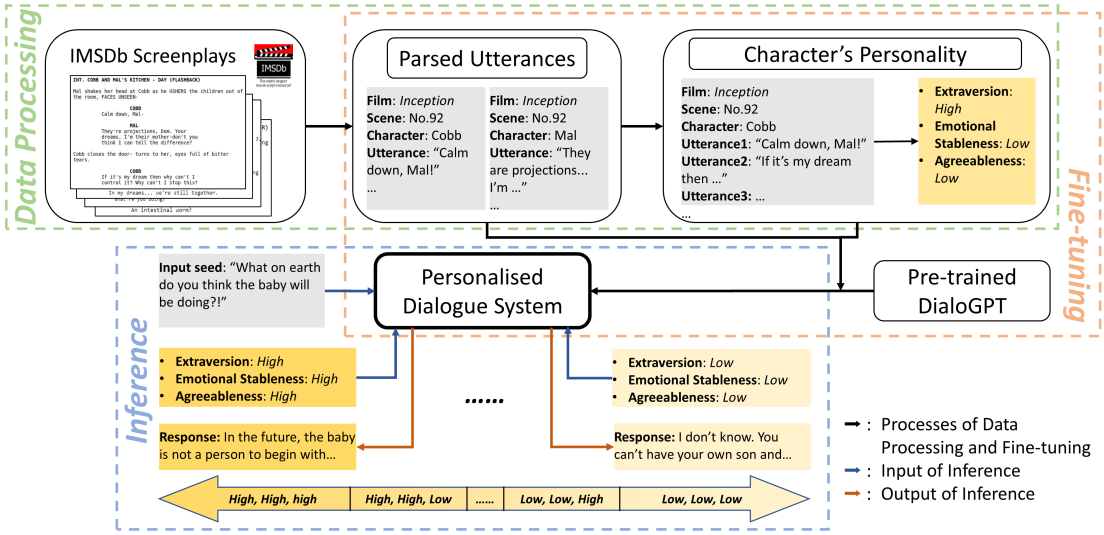


Fig. 4.6 Our approach workflow, including stages of Data Processing, Fine-tuning, and Inference.

features and mappings from much longer context history, as well as to be more generalised over the same context input.

Initially, the neural language model was designed to predict upcoming word(s) depending on the given word sequence (firstly proposed by Bengio et al. (2003)). Suppose provided a piece of text with n words $w_1, ...w_n$, then we denote the first m words as context C, and $w_{m+1}, ..., w_n$ as target response. Therefore, the language model can be described as the product of a series of conditional probabilities as follows:

$$P(R|C) = \prod_{i=m+1}^{n} P(w_i|w_1, ..., w_{i-1}) \tag{4.1}$$

Since the transformer architecture proposed, which is capable to process longer sequence than RNN network, it is possible to feed the longer sequence like a series of sentences instead of words. Following Zhang et al. (2020), the Equation 4.1 can be extended to

$$P(R|C) = \prod_{i=m+1}^{n} P(T_i|T_1, ..., T_{i-1}) \tag{4.2}$$

where $T_1, ..., T_m$ is assigned as context, (i.e. the first m^{th} turn of a dialogue session with T_n turns), and the turns of $T_{m+1}, ..., T_n$ are set as response. Consequently, optimising

a single objective $P(T_n, ..., T_2|T_1)$ can be perceived as optimising all $P(T_i|T_1, ..., T_{i-1})$ source-target pairs.

Adapted Language Model

As prior work about conditional dialogue generation (Colombo et al., 2019; Huang et al., 2018; Yang et al., 2020; Zheng et al., 2020) 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:

$$L_{NLL} = -logP(R|C') \tag{4.3}$$

Text Representation

Most previous language models for dialogue generation or text generation treat **word** as the smallest unit (token) (Li et al., 2016), i.e. a piece of text consists of a set of words. Such models can only compute a restricted space of model-able textual strings, which need to be heavily pre-processed to achieve a certain format, such as lower-case and punctuation tokenisation, as well as lead to some limitations, especially the incapability of addressing out-of-vocabulary words.

An alternative approach, which is used for processing strings as a sequence of UTF-8 **bytes**, is able to address such limitations (Gillick et al., 2016), but the performance of byte-level language models is not competitive with word-level language models on large scale corpora. Byte Pair Encoding (BPE) (Sennrich et al., 2016) is a practical middle ground between character and word level language modelling which effectively interpolates between word-level inputs for frequent symbol sequences and character level inputs for infrequent symbol sequences, which results in the feasibility of language models for handling out-of-vocabulary words. And Radford et al. (2019) adapt this original BPE method by reducing unnecessary merge character categories

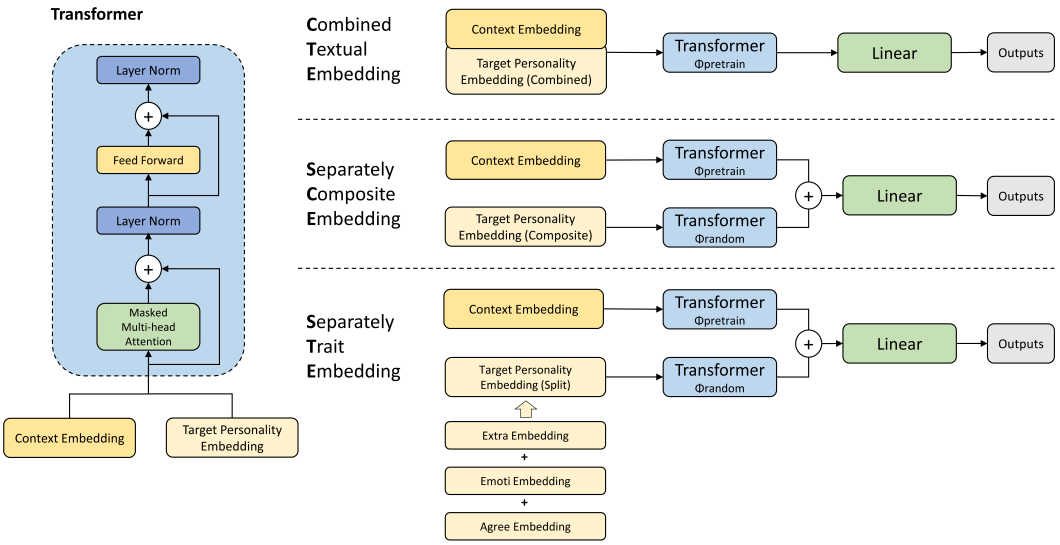


Fig. 4.7 Our approach's framework. The basic Transformer inherited from prior works on the *left*. The 3 methods for embedding characters' personalities are presented on the *right*. Each transformer has the same configuration and uses initialised parameters from pre-trained DialoGPT or randomisation as noted.

for any byte sequence. We here inherit this method for text representation, which combines the benefits of word-level language models with the generality of byte-level encoding methods. Also, it allows us to avoid heavy pre-processing of text.

4.3.2 Personality Incorporation

As we introduced above, to achieve our personalised dialogue generation, it is necessary to investigate how to represent the calculated personality trait labels, as well as how to incorporate target personality into our dialogue system. Here we implement three different representations and incorporating methods for target personality, and evaluate them respectively.

Combined Textual Embedding (CTE)

First, we use a simple and naive method to incorporate personality, which is to treat the target personality as another turn of dialogue in text, referred to ***Combined Textual***

Embedding (CTE). For example, in Equation 4.2, given a dialogue session with the first $i-1$ turns as context, and the i^{th} turn as the response in fine-tuning stage. Also, the i^{th} turn of dialogue is uttered by a character with a certain personality trait combination. Suppose this character is extravert, emotionally stable, and agreeable, which denotes the combination with trait labels is *Extra: high, Emoti: high, Agree: high.* And then the labels of personality traits are collected in text and concatenated with the trait order (Extra, Emoti, Agree). Therefore, the final personality of this character is represented as $Psn = \#high\#high\#high$, where # is the splitter, and concatenated after the context of dialogue as a special turn, leading to the final input for fine-tuning

$$T_1, ..., T_{i-1}, \#high\#high\#high \rightarrow T_i$$

For this embedding method, we explicitly represent the personality with the labels of the three specific traits, concatenate it along with context as the input sequence, and feed it in a single transformer initialised with pre-trained GPT-2 parameters (Figure 4.7).

Separate Composite Embedding (SCE)

The second method is to embed context and target personality separately, referring to ***Separate Composite Embedding (SCE)***. For this representation, we treat each personality with 3 traits as a composite, and label all personalities as $Psn_i(i = 0, 1, ..., 27(3^3))$. Also for the same example, suppose the response turn of dialogue is uttered by a character with a certain personality extravert, emotionally stable, and agreeable, which denotes the combination with trait labels is *Extra: high, Emoti: high, Agree: high* and will be assigned the label 1. All the combinations are assigned a unique label from 1 (*Extra: high, Emoti: high, Agree: high*) to 27 (*Extra: low, Emoti: low, Agree: low*) respectively. Here we particularly denote 0 as none personality. Instead of feeding context embedding and target personality embedding independently into on single

transformer as Zheng et al. (2020), we use these two embeddings as input into two same transformers respectively like Mazare et al. (2018), but with different initial parameters. In other words, the transformer receives context embeddings is initialised with pre-trained DialoGPT weights and the transformer receives personality embeddings is initialised with random weights, as (Figure 4.7) shows.

Separate Trait Embedding (STE)

The 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. (2020), named as ***Separate Trait Embedding (STE)***. For example, the character with a certain personality of extravert, emotionally stable, and agreeable is assigned (0, 0, 0), as per (*Extra: high, Emoti: high, Agree: high*) of semantic label. By conducting this embedding method, the labels of each trait can be embedded and controlled by three different embedding layers independently. Also, the context embedding and aggregated target personality embedding are fed into two same transformers respectively with different initial parameter weights as such in the second method (Figure 4.7).

Summary

We describe three methods for incorporating personality into neural networks. For CTE, we regard the personality as a textual representation and concatenate it with the textual representation of dialogue context. Then we transform this concatenated textual representation into a vectorised embedding in one embedding layer, and followed by the fine-tuning process in transformer. For SCE, we separately transform the dialogue context and personality into embeddings in two embedding layers respectively. Then these two embeddings are feed into two transformers and their outputs are added up to complete incorporation. Last for STE, three personality traits are separately fed into three embedding layers, resulting in three trait embeddings, which

are added up afterwards as a complete personality. Finally the following process is conducted, in a similar manner as SCE.

Chapter 5

Automatic Evaluation

5.1 Methodology

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 with 12 hidden layers, 12 attention heads for each layer, 768 dimensions hidden states, to fine-tune and evaluate our approach. We conduct fine-tuning process from original DialoGPT on our datasets for 2 epochs following their recommendation[1] with manually optimised hyperparameter values as follows: learning rate of 5×10^{-6}, learning rate decay of *noam* (Vaswani et al., 2017), maximum sequence length of 128, and batch size of 8. Then we select the models with the lowest values of evaluating loss and perplexity during fine-tuning.

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 that 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 personality (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

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

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 previous section plus *SDG* for **S**tyle**DG**PT (Yang et al., 2020)). Therefore, 6,000 ($15 \times 50 \times 8$) sets of dialogues are generated for each personality embedding, per level of personality and per genre.

All the acronyms of the embedding methods used in this chapter are list as below:

- **C**ombined **T**extual **E**mbedding: **CTE**
- **S**eparately **C**omposite **E**mbedding: **SCE**
- **S**eparately **T**rait **E**mbedding: **STE**
- **S**tyle**D**(ialog)**G**PT: **SDG**

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 accumulated and evaluated by the same tool as per for labelling.

5.2.1 Overall Aspect

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 5.1, "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 matches.

In Figure 5.1, we show an overall personality identification accuracy with two levels of personality and three embedding methods. From the perspective of scene-level personality (Figure 5.1), we notice that with both *SCE* and *STE*, over 80% of target personalities across 4 genres can be correctly identified, with all three correct

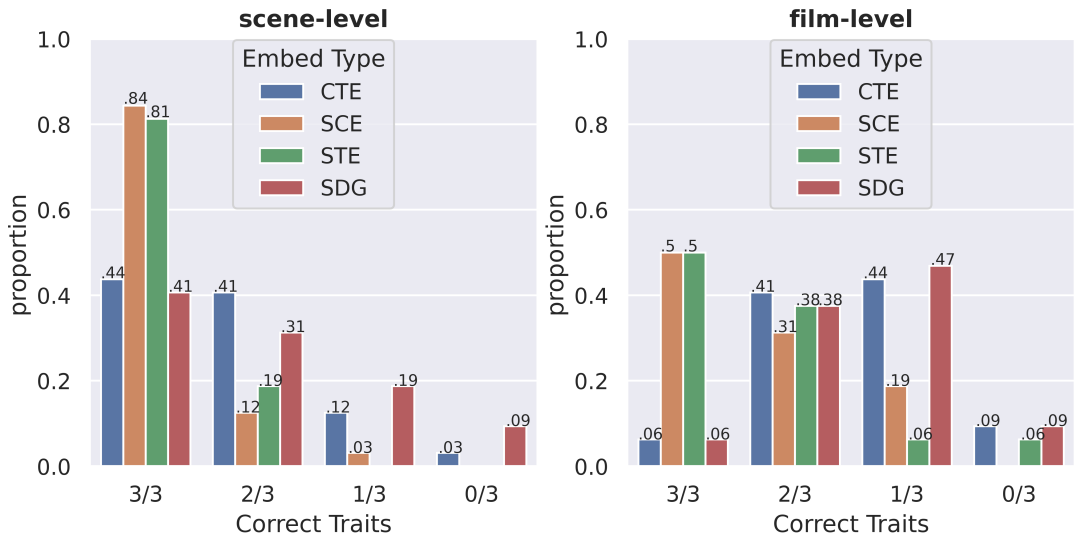


Fig. 5.1 The accuracy of personality identification for generated dialogues with scene-level (left) and film-level (right) personality and 4 embedding methods.

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*, which has more than 40% personalities with all three traits correct and at least two traits correct. From the perspective of film-level personality (Figure 5.1), we can also obtain the similar observation that *SCE* and *STE* have better performance than *CTE*.

Comparing the performance between scene-level and film-level personality (Figure 5.1), 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.

Figure 5.2 shows the scores of identified personalities on a finer-grained trait aspect. With the target labels for each trait, the differences of the scores between the *high* and *low* are the focuses of interest, where the greater 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 the trait perspective, we observe that the trait extroversion is the most correctly identified trait among all three traits with the great difference of score

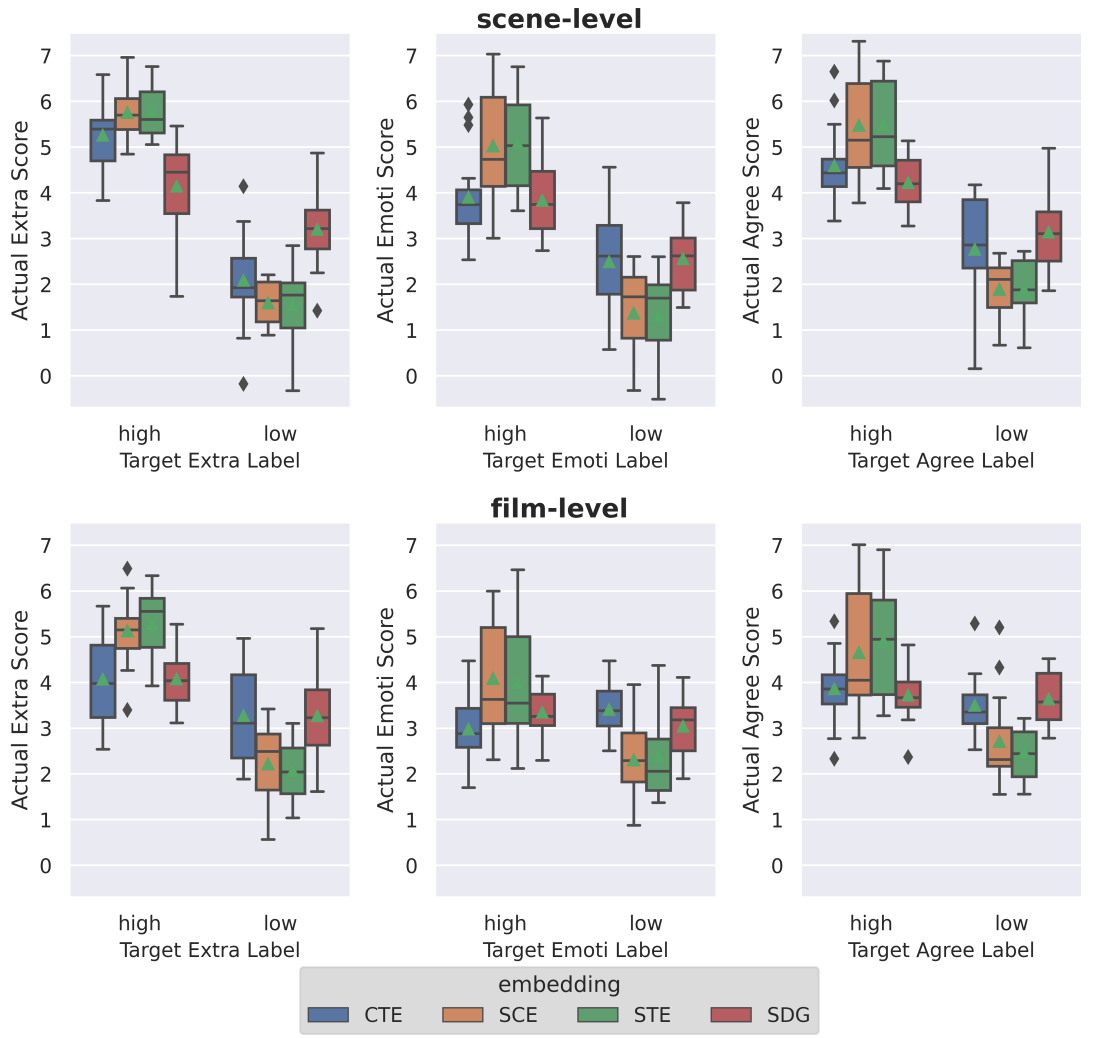


Fig. 5.2 Results of personality identification for generated dialogues on trait aspects with identified scores. For each aspect, the first row shows the results with scene-level personality, and the second row shows the results with film-level personality. All results are grouped by embedding methods.

distribution between two labels and 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.

With or without personality

We also compare the distributions of personality identification scores between our method and the plain dialogue generation approach, i.e. without personality. We particularly select *STE* group (including the results of personality identification with

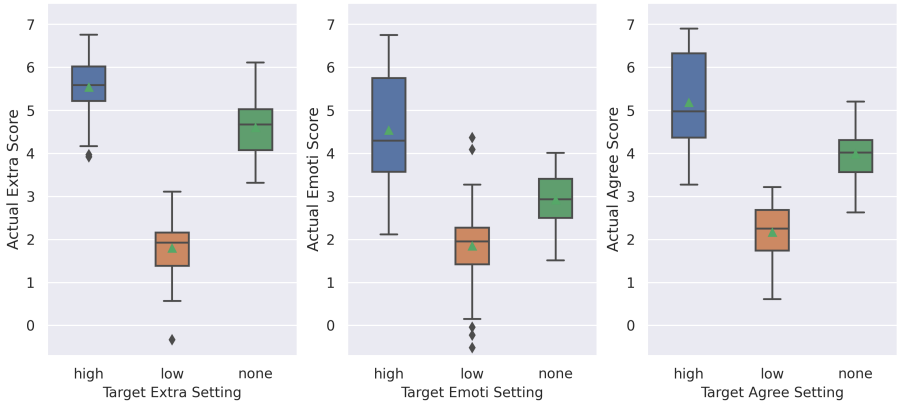


Fig. 5.3 The comparison between the generations with personality (STE film-level & scene-level) and w/o personality (DialoGPT).

film-level and scene-level personality on all 4 genres) to represent the generation with personality because *STE* has the best performance in personality identification evaluation across all personality embedding methods, and *SCE* has the similar distributions as *STE*. We calculate the personality identification scores for three traits, and group them by personality settings. For each trait, we calculate scores for three settings, of which the dialogues with personality settings (*high* and *low* labels) are generated using method *STE*, and the dialogues without personality settings are generated using DialoGPT directly. All dialogues are generated with the same input seeds.

In Figure 5.3, we observe that the boxes with personality settings reflect more polar distributions, while the ones without personality settings tend to reflect relatively less polar distributions (i.e. the boxes of none are positioned between generations with personalities regarding the scores). Therefore, compared with the plain dialogue generation system, which tends to generate dialogues with mostly plain personality, our method is able to control dialogue generation with given extreme personality accordingly.

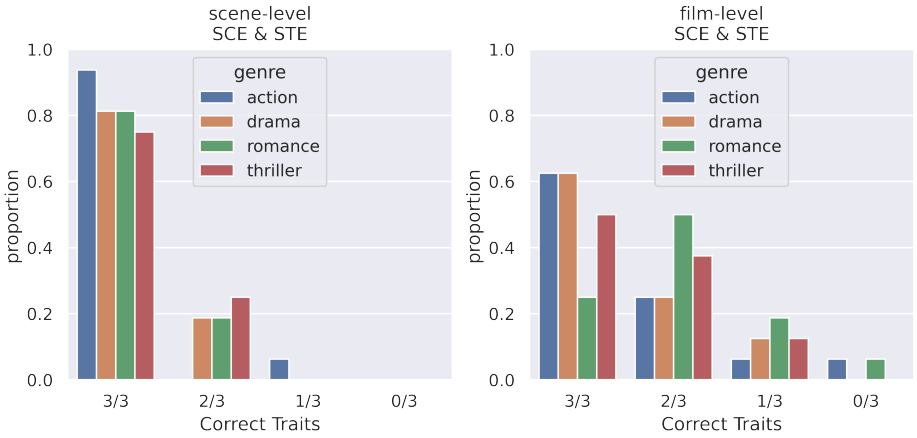


Fig. 5.4 Results of personality identification for generated dialogues on overall aspect with trait matching accuracy for all 4 genres, with *SCE* and *STE* embedding types, for both film-level and scene-level personalities.

5.2.2 Genre Aspect

In Figure 5.4, we specifically select the results of personality identification with embedding methods of *SCE* and *STE* since they perform better than *CTE* and *SDG*, and present the results on both scene-level and film-level across 4 genres. From Figure 5.4, we obtain the similar observations that scene-level personality contributes more positive impact on personality identification rather than film-level personality does as Chapter 5.2.1 presented, regardless genre. For example, around 80% or more personalities with all three traits are identified correctly (3/3) on all 4 genres with scene-level personality (Figure 5.4 *left*), while the number only reach up to 60% with film-level personality (Figure 5.4 *right*).

Also, we observe that with scene-level personality (Figure 5.4 *left*), the results on the genres of **drama**, **romance** and **thriller** tend to have relatively similar personality identification accuracy, while the one on **action** has higher accuracy. Comparably, with film-level personality (Figure 5.4 *right*), the results on the genres of **action**, **drama**, and **thriller** have the similar distributions as the ones with scene-level personality (i.e. **action** and **drama** are the highest, while **thriller** has slightly less accuracy). However, the accuracy (here we only focus "3/3") of the genre **romance** drops down

significantly from levelling with **drama** and **thriller** in Figure 5.4 (*left*) to less than half accuracy in Figure 5.4 (*right*). A probable reason for this accuracy dropping down could be, comparing with the other three genres, **romance** has the smallest corpus size, plus the fact of relative weakness of film-level personality.

Table 5.1 Statistics of generated dialogues grouped by personality level and embedding method. The up arrow denotes the expectation of greater numbers, and down arrow on the contrary.

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

5.3 Basic Analysis

Apart from personality identification accuracy, we also evaluate generated dialogues on some metrics widely used in computational linguistics.

5.3.1 Metrics

The major purpose for us here to is to evaluate whether incorporating additional characters' personalities affects the variety of generations. Hence, we select two metrics, edit distance and semantic cosine similarity to evaluate generations from surface text perspective and semantic perspective respectively. For edit distance, we particularly use Levenshitein distance following Porteous et al. (2013) as they use this metric to measure the difference between pairs of narratives generated with different relationships between characters.

We also use BLEU (Papineni et al., 2002), which is one of the commonly used word-overlap metrics in NLG research. By comparing the generated dialogues and golden references, the word overlapping rate denotes the similarity of generations to the references. Normally, a higher word overlapping rate represents higher quality. In our experiment, we evaluate the generated dialogues on BLEU on both utterance-level and dialogue-level.

And we also calculate the simple sentence count and word count of each generated dialogue for evaluating the capability of each embedding method to our approach.

5.3.2 Results

We present our findings for a sample of generated dialogues in Table 5.1.

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 numbers among 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 are all over 1.33 sentences per turn, which are more than dialogues from original DialoGPT, but much less than the written screenplay. And the word counts per turn are around 10 words, which are less than DialoGPT and written screenplay. This observation probably indicates that in order to correctly control the target personality of dialogue, our approach loses some of the ability of 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 the same seed and same target personality (For DialoGPT, only the seed controlled), we calculate the edit distance (normalised by text length with range 0 to 1) using Levenshiten distance, and semantic cosine similarity (with range -1 to 1) using Universal Sentence Encoder (Cer et al., 2018) for these dialogues pair-wisely. From the table, we observe that the dialogues generated by our approach are able to provide higher edit distance, as well as lower semantic similarity. This observation could be regarded as our approach is able to generate dialogues with more variety given the same seeds from surface text perspective and semantic perspective.

Furthermore, we use Bleu (Papineni et al., 2002) (n=1,2) as a representative of word-overlap metrics, to evaluate generated dialogues against the sole reference from the original screenplays. We apply the Bleu metric on dialogue level and utterance level evaluations. As the results show in Table 5.1, from the personality level aspect, the Bleu scores for film-level personality are higher than the ones for scene-level personality. Also from the embedding method aspect, *CTE* reaches the highest Bleu scores in either personality level. Normally higher Bleu scores are expected, which denotes the generated samples and the reference(s) share more overlapped n-gram words. However, in our case, we noticed the settings that achieve higher Bleu scores (film-level personality, *CTE*) have lower personality identification accuracy (see Figure 5.1). This observation is understandable, because generations with different personality combinations are more similar to the sole reference also means they are more similar

Table 5.2 Generated dialogue examples from the two extreme personality trait combinations.

Target Personality (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.

to each other. Meanwhile, given the fact that we have only one ground-truth golden reference for each input seed, achieving higher BLEU scores indicates that these generated dialogues are similar to the reference and similar to each other, which might be against the goal that we expect them to reflect different target personalities correspondingly. 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 (Liu et al., 2016), especially in the case where there is a lack of golden references, e.g. the ones written by screenplay writers.

5.4 Generated Samples

We present several examples of dialogues generated with the setting reaches the best performance (scene-level, and *SCE*), given the same input seed (statement) and on 8 personalities (Table 5.3). And in Table 5.2 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 sub-dataset, and they are expected to take place between two characters as the dialogue structure in the dataset.

5.5 Conclusions

We evaluate the generated dialogues from the aspects of personality identification and basic linguistics.

From the aspect of personality identification, *SCE* and *STE* embedding methods have significantly higher accuracy than *CTE* and StyleDGPT. Also, scene-level personalities have higher accuracy than film-level ones. And the genre of drama tends to have higher accuracy with both film-level and scene-level personality, and the genre of romance tends to have the lowest accuracy, compared with the other three genres.

Also, from the results shown of comparisons between our approach and other dialogue systems, we notice that our approach performs better (generate more percentage of dialogues with given personalities correctly identified) than StyleDGPT, and has more capability of controlling personalities rather than plain language models.

And from the aspect of basic linguistics, we observe that the capability of generation of our approach is reduced on generated word quantity, but slightly increased on generated sentence quantity, which results in a more natural word-sentence ratio, i.e. similar to the original screenplays.

Table 5.3 Examples of generated dialogues from the full range of personality trait combinations (8 separate ones).

Target Personality (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.

Chapter 6

Human Evaluation - CTE

In the preceding chapter, we have demonstrated our dataset and approach are able to improve the dialogue quality regarding variety, as well as reflect the given target personality using automated metrics. While it is also necessary to evaluate the generated dialogues with human intelligence.

Therefore, in this chapter, we introduce two user studies we conducted, **User Study i** and **User Study ii**, including the survey design, the material to be evaluated generated with CTE method, the results, and the analysis. The user studies are aiming to evaluate the impact on dialogue generation by incorporating additional personality features. More specifically, to what extent do the additional narrative-based personality features affect the quality of generated dialogues using data-driven approach statistically, as well as to what extent do the participants perceive the target personality.

During user studies, the recruited participants will be grouped purposefully and be assigned to do surveys. In each survey, each participant is supposed to evaluate several pieces of generated dialogue with their own knowledge and with different feature settings according to linguistic and narrative measuring factors that are defined clearly and precisely.

6.1 Methodology

6.1.1 Survey Design

Following our aims, we design the survey for our user study to evaluate dialogues from 3 perspectives:

1. Dialogue Quality
2. Personality Identification
3. Genre Identification

We initially define our dialogue quality from 5 aspects in order to include semantic and pragmatic evaluation, as well as narrative-based evaluation:

- The Dialogue is Grammatically Correct
- The Content of Dialogue is Consistent
- The Content of Dialogue is Logically Believable
- The Dialogue is Easy to Comprehensive
- The Dialogue has Rich Dramatic Effect

After analysed 6 responses for a small pilot study, we noticed that some participants showed their impatience and frustration when they were doing this survey, which leading to more random responses, i.e. participants are answering questions merely for finishing the survey rather than making proper judgements. And such situation was more likely to be noticed progressively, especially at the second half of the survey. A major reason for this situation is the total amount of questions, which has reached 200 (5×40) only for dialogue quality aspect on such design. To answer such amount of questions, participants are probably losing their attention over time. Another reason is participants are required to answer same questions repeatedly to judge dialogues, which might cause the whole survey a tedious work. We also noticed that it is difficult to judge generated dialogues from narrative perspective as each dialogue only contain 4 utterances and is lack of context as a narrative, according to the feedback from participants in pilot study.

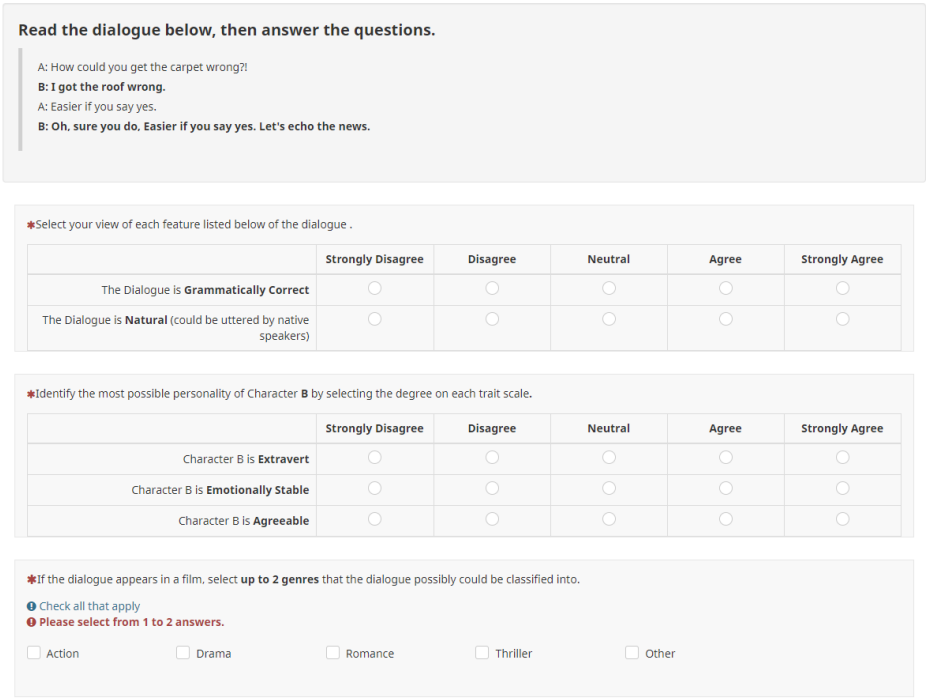
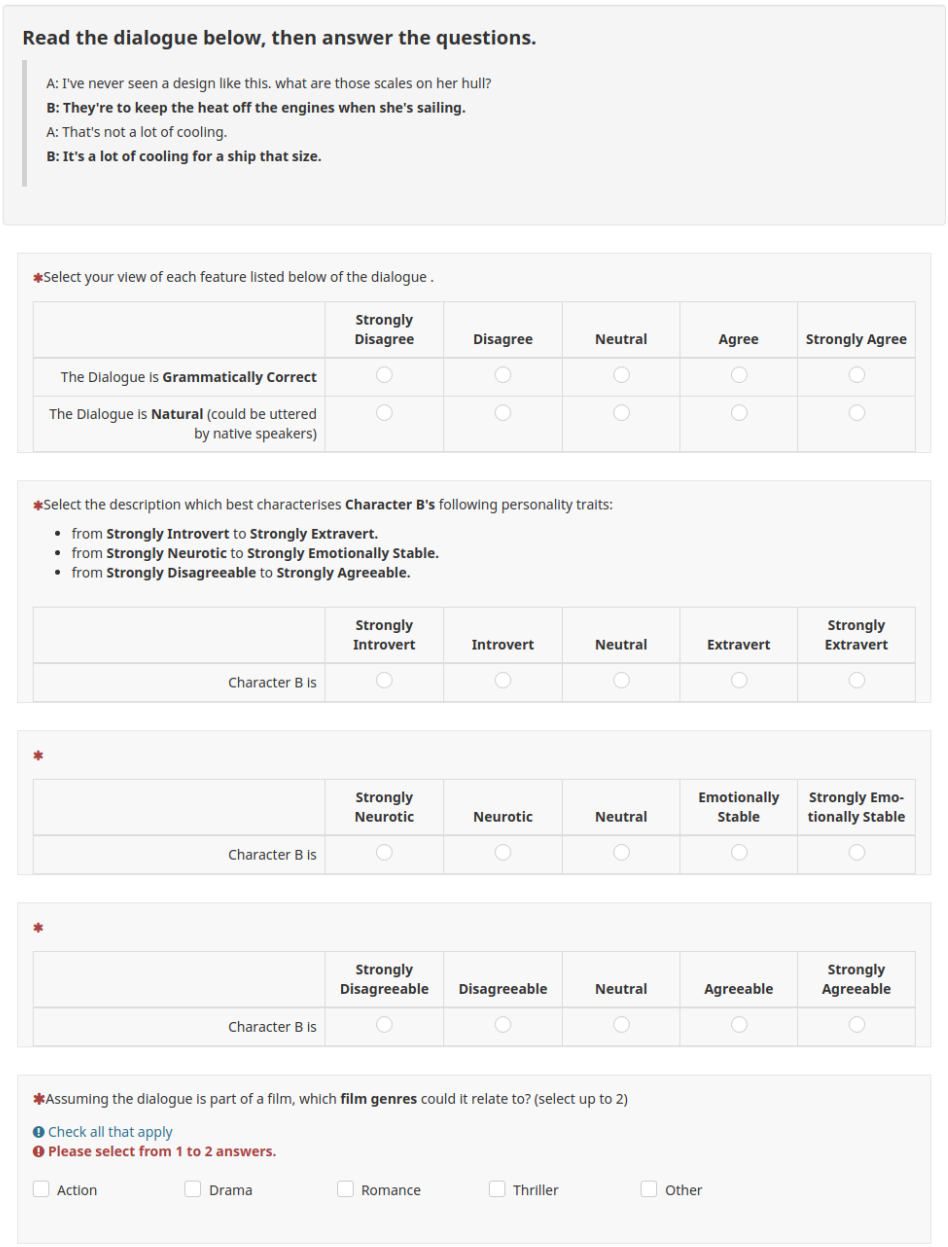


Fig. 6.1 A screenshot (one page) of the survey for **User Study i**. Each page contains a piece of either generated dialogue or original written dialogue, a question group for dialogue quality evaluation, another question group for personality identification, and a question for story genre identification.

Considering the results of our pilot study, we reduce the amount of questions and only consider the linguistic quality of dialogues. Novikova et al. (2017) use a 6-point Likert scale to evaluate dialogues for the following three aspects:

- **Informativeness** *(Does the utterance provide all the useful information from the meaning representation?)*
- **Naturalness** *(Could the utterance have been produced by a native speaker?)*
- **Quality** *(How do you judge the overall quality of the utterance in terms of its grammatical correctness and fluency?)*

Referring to Novikova et al. (2017), we adapt the factors for our dialogue quality evaluation, with Naturalness and Grammatical Correctness for our actual user studies (see Figure 6.1 and Figure 6.2). Here we require participants to judge dialogue quality on 5-point Likert scale by given positive statement (e.g. The dialogue is grammatically



Read the dialogue below, then answer the questions.

A: I've never seen a design like this. what are those scales on her hull?
B: They're to keep the heat off the engines when she's sailing.
A: That's not a lot of cooling.
B: It's a lot of cooling for a ship that size.

*Select your view of each feature listed below of the dialogue .

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The Dialogue is **Grammatically Correct**	○	○	○	○	○
The Dialogue is **Natural** (could be uttered by native speakers)	○	○	○	○	○

*Select the description which best characterises **Character B's** following personality traits:

- from **Strongly Introvert** to **Strongly Extravert.**
- from **Strongly Neurotic** to **Strongly Emotionally Stable.**
- from **Strongly Disagreeable** to **Strongly Agreeable.**

	Strongly Introvert	Introvert	Neutral	Extravert	Strongly Extravert
Character B is	○	○	○	○	○

*

	Strongly Neurotic	Neurotic	Neutral	Emotionally Stable	Strongly Emotionally Stable
Character B is	○	○	○	○	○

*

	Strongly Disagreeable	Disagreeable	Neutral	Agreeable	Strongly Agreeable
Character B is	○	○	○	○	○

*Assuming the dialogue is part of a film, which **film genres** could it relate to? (select up to 2)

Check all that apply
Please select from 1 to 2 answers.

☐ Action ☐ Drama ☐ Romance ☐ Thriller ☐ Other

Fig. 6.2 A screenshot (one page) of the modified survey for **User Study ii**. Each page contains a piece of either generated dialogue or original written dialogue, a question group for dialogue quality evaluation, another three separate questions for personality identification, and a question for story genre identification.

correct). Therefore, the selection of *(Strongly) agree* denotes good dialogue quality and the selection of *(Strongly) disagree* denotes bad dialogue quality.

For the part of personality identification, we also provide human participants with a statement of description of character's personality on the same 5-point Likert scale. In **User Study i**, The options 5-point Likert scale are set as same as ones for dialogue quality, And for identifying each personality trait, there is a positive statement presented (e.g. The character is extravert/emotionally stable/agreeable) as Figure 6.1 shown. Focusing on extraversion, the selection of *(Strongly) agree* denotes participants consider the given utterances are spoken by an extravert character. And *(Strongly) disagree* on the contrary. In **User Study ii**, we keep the same 5-point Likert scale, but with different scale descriptions. For example, here, we use *Strongly extravert* rather *Strongly agree* for *The character is extravert.* And all options are *Strongly extravert*, *Extravert*, *Neutral*, *Introvert*, *Strongly introvert.* And for the other two traits, the descriptions are modified similarly.

Last, for genre identification, the participants are required to select the 1 or 2 most likely genres from 5 options on given dialogues according to their judgements.

6.1.2 Material for Evaluation

In this chapter, all dialogues need to be evaluated are generated by the fine-tuned transformer-based neural network on our two film-level IMSDb corpora using *CTE* method. Given an input utterance as seed, three successive utterances are generated automatically and then packed as a 4-turn dialogue along with the input seed. In order to make the dialogue more readable for participants, we manually add annotations (**-A:** and **-B:**) by turns before each utterance, conducive to be recognised as a dialogue conducted between two characters. The format of a presented dialogue is as following:

-A: input seed.
-B: generated turn.
-A: generated turn.

-B: generated turn.

Aiming to evaluate the impact of whether incorporating additional personalities and incorporating different personalities, given a same input utterance, we conduct the material generating and grouping following procedures below:

- Dialogue set 1, or **No-Personality (NOP)**: Generating 1 pieces of dialogues with no additional personality trait combination.
- Dialogue set 2, or **Personality (PER)**: Generating 8 short pieces of dialogues with 8 personality settings, i.e. one dialogue for each certain personality trait combination.
- Dialogue set 3, or **Source-Script (SS)**: Collecting the dialogue starting with this input utterance from original screenplay. (Ideally, it is more reasonable to recruit an expertise writer to write 8 short pieces of dialogues with this input utterance and with 8 personality trait combinations respectively.)

For the personalities, or the trait combinations, we specifically select two extreme labels (i.e. *High* and *Low*) for each of three personality traits, but omitting label *Medium* as we did in Automatic Evaluation (Chapter 5), with the consideration of simplicity. Therefore, for the dialogue set **PER**, a total of 8 ($2 \times 2 \times 2$) dialogues with each personality trait combinations are used to be measured. An entire enumerate for all personality trait combinations are:

1. (extrovert, emotionally stable, agreeable)
2. (extrovert, emotionally stable, disagreeable)
3. (extrovert, neurotic, agreeable)
4. (extrovert, neurotic, disagreeable)
5. (introvert, emotionally stable, agreeable)
6. (introvert, emotionally stable, disagreeable)
7. (introvert, neurotic, agreeable)
8. (introvert, neurotic, disagreeable)

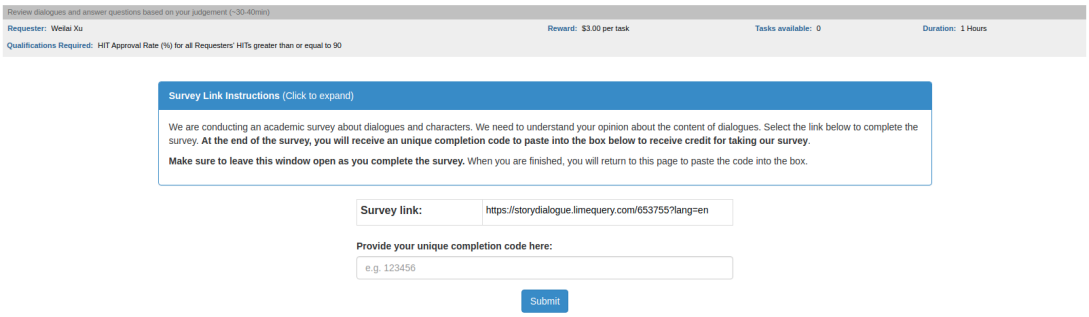


Fig. 6.3 A screenshot (one page) of the published request on AMT.

Consequently, 10 pieces of dialogue are generated for each input seed (1 for set **NOP**, 8 for set **PER**, and 1 for set **SS**). Also, we select 4 different input seeds from unused screenplays on 4 genres respectively. Therefore, each participant are asked to review $(1 + 8 + 1) \times 4 = 40$ pieces of dialogue and make judgements. All dialogues for evaluation are presented in Appendix B.1 and Appendix B.2[1].

6.1.3 Deployment

The survey is designed as workflow which requires participants to make their judgements on dialogues by answering same questions repeatedly. Considering this characteristic, some simple survey platforms such as Google Forms[2] do not meet our requirements, because they merely allow us to present contents on a single page. We finally select a more professional survey platform called LimeSurvey[3], which allow us to publish a survey with multiple pages and group questions up with same purpose (See Figure 6.1 and Figure 6.2). To make sure participants not to be informed that whether a dialogue is generated or collected from original screenplay, the order of all dialogues is randomised in the survey.

Because of the Covid-19 pandemic, it is not likely to recruit participants on the campus in person. Therefore, we publish user studies through Amazon Mechanical

[1]For **User Study ii**, we select 13 dialogues of which the personalities in *Corpus 2* consist of more than 2 traits that are more extreme than those in *Corpus 1* across 4 genres based on the comparison in Section 4.2.3, plus $(1 + 1) \times 4 = 8$ of set **NOP** and set **SS**, in total 21 dialogues.

[2]https://www.google.co.uk/forms/about/

[3]https://www.limesurvey.org/

Turk (AMT)[4] for recruiting crowdworkers as several NLG and storytelling works evaluate textual generations (Liu et al., 2016; Novikova et al., 2017; Yang et al., 2020). Considering the nature of the materials in these user studies, the only requirement for the participants of interest is that they have to be native English speakers. We assume the participants recruited who honestly qualified themselves proficient English speakers with the principle of *assuming good faith*. And then the recruited participants can be directed to LimeSurvey to start the survey (see Figure 6.3).

6.1.4 Measuring

For the evaluations of dialogue quality and personality identification, the participants are asked to review each piece of dialogue and make their own judgement on each question by grading according to an 5-point Likert scale, which then will be used as a ordinal scale in categorical statistical analysis, as well as be converted to a quantified scale (from -2 to 2) for numerical statistical analysis:

We use normal wording in 5-point Likert scale for evaluating dialogue quality, in which "Strongly Disagree" stands for worst quality and "Strongly Agree" for best quality as our statements are positive.

- -2 = Strongly Disagree (i.e. worst quality)
- -1 = Disagree
- 0 = Neutral
- 1 = Agree
- 2 = Strongly Agree (i.e. best quality)

And for the personality identification, we use a similar measuring scale as for the quality evaluation.

- -2 = Strongly Introvert/Neurotic/Disagreeable
- -1 = Introvert/Neurotic/Disagreeable
- 0 = Neutral

[4]https://www.mturk.com/

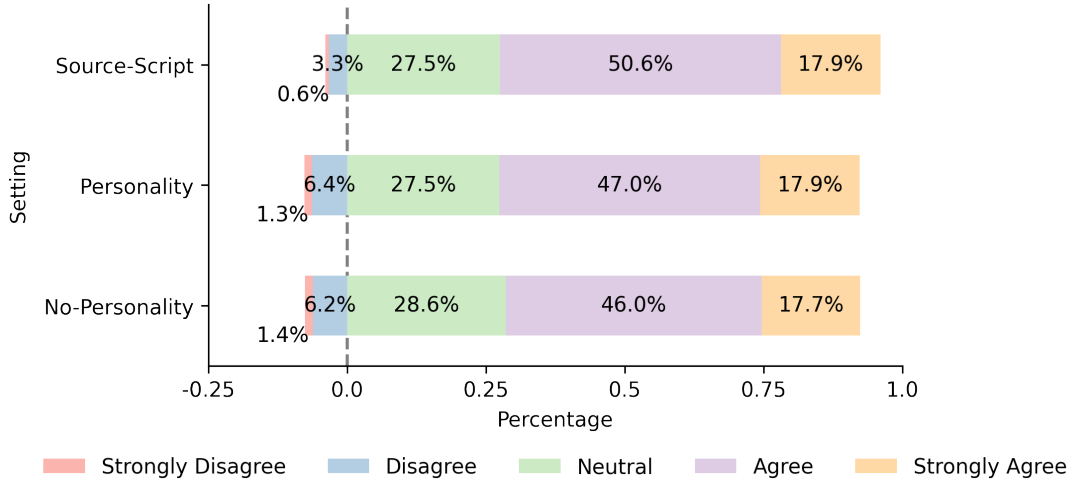


Fig. 6.4 The results of dialogue quality evaluation from perspective of setting in **User Study i**. Note the setting "Personality" contains all 8 personality combinations.

- 1 = Extravert/Emotionally Stable/Agreeable
- 2 = Strongly Extravert/Emotionally Stable/Agreeable

for the evaluation of genre identification, the measurement would be the percentage of precision and recall for selections.

6.2 Results

There are total **79** participants completed the **User Study i** and total **69** participants completed the **User Study ii**. In following sections, the results will be presented with the order of dialogue quality, personality identification, and genre identification perspective, and they will be compared between two user studies.

6.2.1 Dialogue Quality

Setting Perspective

Figure 6.4 and Figure 6.5 show the overall quality of generated dialogues in **User Study i** and **User Study ii** respectively, which tend to be positive on all three settings.

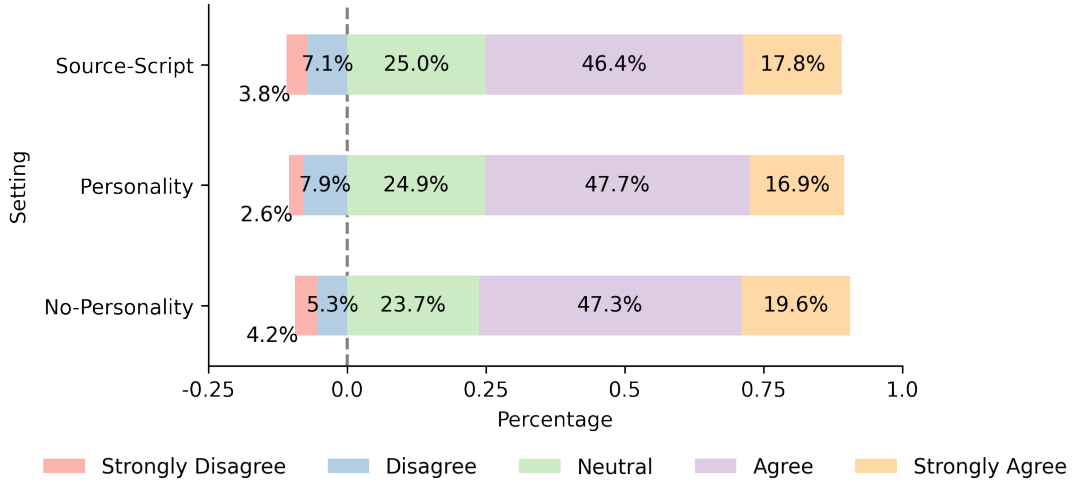


Fig. 6.5 The results of dialogue quality evaluation from perspective of setting in **User Study ii**. Note the setting "Personality" contains all 8 personality combinations.

We also conduct a 1-tail T-test for comparing one of three setting against other two pairwisely, in order to investigate the impact on dialogue quality of adding personality information. We show the T-test results in Table 6.1. In **User Study i**, we observe that dialogues from source script perform better than generated dialogues in the term of mean value, including with and without personality. And p-values show the differences are significant (0.993 and 0.977, equal to 0.007 and 0.023). Also, within the scope of generated dialogues, those with personality have higher mean score than those without personality.

	setting	setting #1	setting #2	pairwise	
	#1 vs #2	mean (std)	mean (std)	t-stats	p-value
User Study i	P vs NP (Overall)	0.736(0.87)	0.725(0.87)	0.311	0.378
	P vs S (Overall)	0.736(0.87)	0.818(0.78)	-2.449	0.993
	NP vs S (Overall)	0.725(0.87)	0.818(0.78)	-2.001	0.977
User Study ii	P vs NP (Overall)	0.685(0.93)	0.728(0.97)	-0.921	0.821
	P vs S (Overall)	0.685(0.93)	0.672(0.93)	0.276	0.391
	NP vs S (Overall)	0.728(0.97)	0.672(0.93)	0.959	0.169

Table 6.1 1-tail T-test for dialogue quality on the overall perspective (aggregation of both) in two user studies. The setting references: Personality - P, No-Personality - NP, Source-Script - S

While in **User Study ii**, we observe the opposite phenomena, with the generated dialogues have higher quality mean scores than those from source script, as well as dialogues without personality have higher mean score than those with personality.

	setting #1 vs #2	setting #1 mean (std)	setting #2 mean (std)	pairwise t-stats	pairwise p-value
User Study i	P vs NP (Grammar)	0.623(0.84)	0.576(0.86)	0.909	0.182
	P vs S (Grammar)	0.623(0.84)	0.680(0.76)	-1.256	0.895
	NP vs S (Grammar)	0.576(0.86)	0.680(0.76)	-1.612	0.946
	P vs NP (Naturalness)	0.850(0.89)	0.873(0.86)	-0.461	0.677
	P vs S (Naturalness)	0.850(0.89)	0.956(0.78)	-2.242	0.987
	NP vs S (Naturalness)	0.873(0.86)	0.956(0.78)	-1.260	0.900
User Study ii	P vs NP (Grammar)	0.619(0.94)	0.630(0.98)	-0.174	0.569
	P vs S (Grammar)	0.619(0.94)	0.612(0.93)	0.100	0.460
	NP vs S (Grammar)	0.630(0.98)	0.612(0.93)	0.222	0.412
	P vs NP (Naturalness)	0.751(0.92)	0.826(0.95)	-1.152	0.875
	P vs S (Naturalness)	0.751(0.92)	0.731(1.01)	0.286	0.388
	NP vs S (Naturalness)	0.826(0.95)	0.731(1.01)	1.128	0.130

Table 6.2 1-tail T-test for dialogue quality on the perspective of grammar and naturalness in two user studies. The setting references: Personality - P, No-Personality - NP, Source-Script - S

Criterion Perspective

We also present the result from the perspective of quality criteria. In Table 6.2, It can be observed that either in **User Study i** or **User Study ii**, the mean scores of *Naturalness* evaluation are higher than the scores of *Grammar* evaluation. Regarding the settings, we notice that the distribution for each comparison has a similar trend to the overall quality comparison.

6.2.2 Personality Identification

In this section, we present the results of personality identification in two user studies. For each personality trait, we describe the selections of participant and analyse them.

Extraversion vs. Introversion

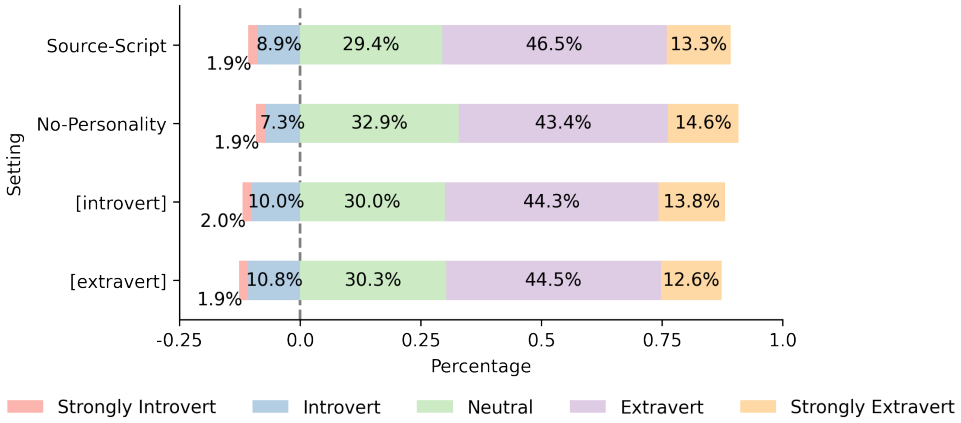


Fig. 6.6 The results of personality identification in **User study i** on perspective of "extravert" and "introvert", along with "source-script" and "No-Personality" as controls.

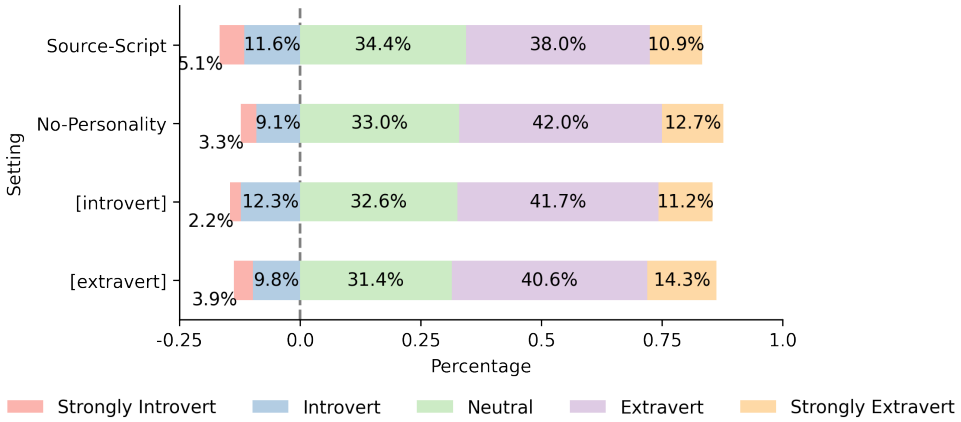


Fig. 6.7 The results of personality identification in **User study ii** on perspective of "extravert" and "introvert", along with "source-script" and "No-Personality" as controls.

Here we show the selections of participants regarding extraversion and introversion for **User Study i** (see Figure 6.6) and **User Study ii** (see Figure 6.7).

In **User Study i** (Figure 6.6), for the setting of *Extravert*, we observe that 57.1% of participants correctly select either *Extravert* or *Strongly Extravert*, comparing with

the percentage of *Introvert* or *Strongly Introvert* (12.7%), which is corresponding to our expectation. However, we also notice for dialogues with all four settings, the percentage of participants who select *Extravert* or *Strongly Extravert* is significantly higher than the opposite end, i.e. *Introvert* or *Strongly Introvert*. Therefore, although the selection shows the mostly correct perception for the setting of *Extravert*, there exist possibilities that the participants are less able to perceive the differences between the dialogues generated with *Extravert* personality and *Introvert* personality, as the percentage of selection of *Extravert* or *Strongly Extravert* for personality *Extravert* (57.1%) and personality *Introvert* (58.1%) are similar.

In **User Study ii** (Figure 6.7), we observe the similar distributions of selection as **User Study i**. There exists a slight change between the selections of personality *Extravert* and personality *Extravert*, where more participants perceive a dialogue with personality *Extravert* is *Extravert* (54.9%) than perceive a dialogue with personality *Introvert* is *Extravert* (52.9%).

Emotional Stableness vs. Neurotics

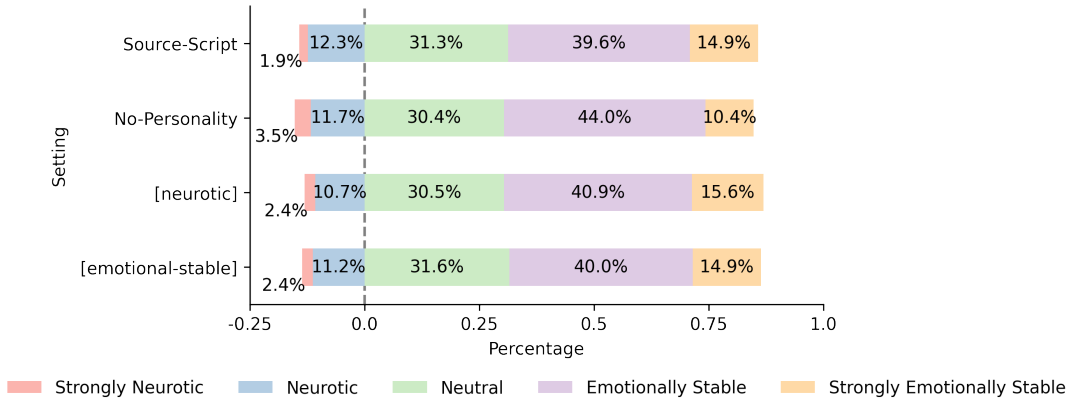


Fig. 6.8 The results of personality identification in **User study i** on perspective of "emotional stable" and "neurotic", along with "source-script" and "No-Personality" as controls.

Here we show the selections of participants regarding emotional stableness and neurotics for **User Study i** (see Figure 6.8) and **User Study ii** (see Figure 6.9).

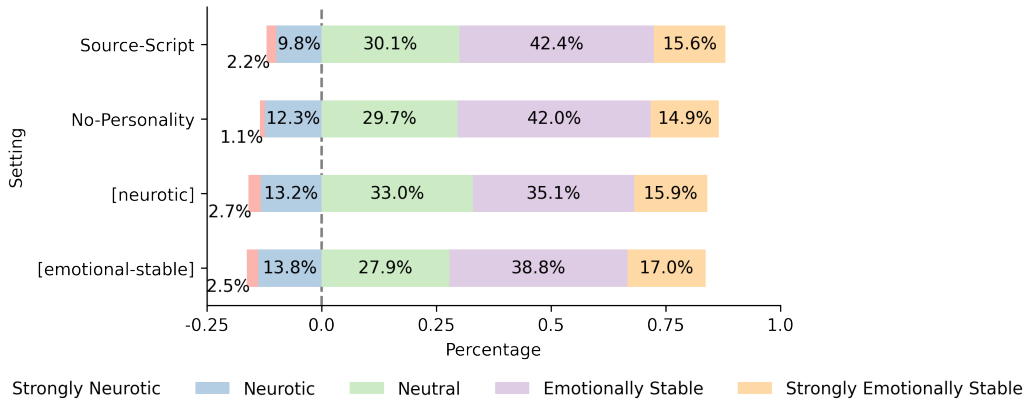


Fig. 6.9 The results of personality identification in **User study ii** on perspective of "emotional stable" and "neurotic", along with "source-script" and "No-Personality" as controls.

In **User Study i** (Figure 6.8), for the setting of *Emotionally Stable*, we observe that 54.9% of participants correctly select either *Emotionally Stable* or *Strongly Emotionally Stable*, comparing with the percentage of *Neurotic* or *Strongly Neurotic* (13.6%), which is corresponding to our expectation. However, we also notice for dialogues with all four settings, the percentage of participants who select *Emotionally Stable* or *Strongly Emotionally Stable* is significantly higher than the opposite end, i.e. *Neurotic* or *Strongly Neurotic*. Therefore, although the selection shows the mostly correct perception for the setting of *Emotionally Stable*, there exist possibilities that the participants are less able to perceive the differences between the dialogues generated with *Emotionally Stable* personality and *Neurotic* personality, as the percentage of selection of *Emotionally Stable* or *Strongly Emotionally Stable* for personality *Emotionally Stable* (54.9%) and personality *Neurotic* (56.5%) are similar.

In **User Study ii** (Figure 6.9), we observe the similar distributions of selection as **User Study i**. There exists a slight change between the selections of personality *Emotionally Stable* and personality *Emotionally Stable*, where more participants perceive a dialogue with personality *Emotionally Stable* is *Emotionally Stable* (55.8%) than perceive a dialogue with personality *Neurotic* is *Emotionally Stable* (51.0%).

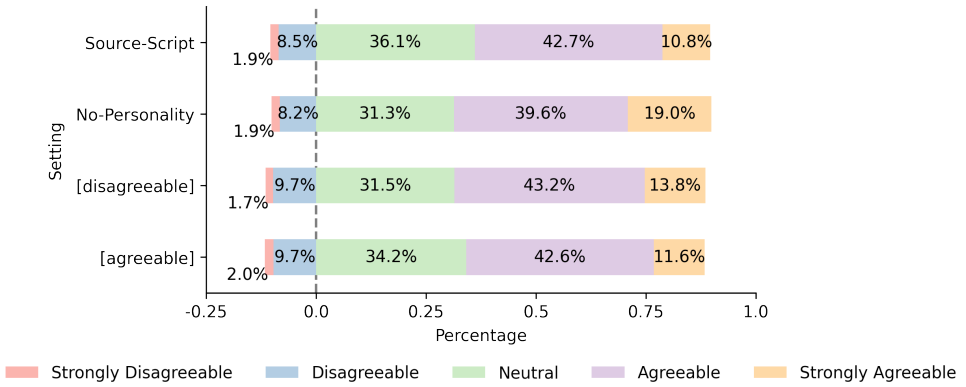


Fig. 6.10 The results of personality identification in **User study i** on perspective of "agreeable" and "disagreeable", along with "source-script" and "No-Personality" as controls.

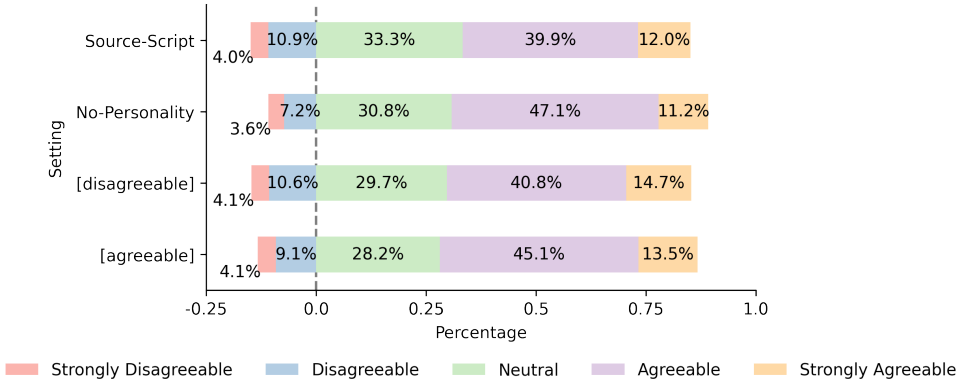


Fig. 6.11 The results of personality identification in **User study ii** on perspective of "agreeable" and "disagreeable", along with "source-script" and "No-Personality" as controls.

Agreeableness vs. Disagreeableness

Here we show the selections of participants regarding agreeableness and disagreeableness for **User Study i** (see Figure 6.10) and **User Study ii** (see Figure 6.11).

In **User Study i** (Figure 6.10), for the setting of *Agreeable*, we observe that 54.2% of participants correctly select either *Agreeable* or *Strongly Agreeable*, comparing with the percentage of *Disagreeable* or *Strongly Disagreeable* (11.7%), which is corresponding to our expectation. However, we also notice for dialogues with all four settings, the percentage of participants who select *Agreeable* or *Strongly Agreeable* is significantly higher than the opposite end, i.e. *Disagreeable* or *Strongly Disagreeable*. Therefore, although the selection shows the mostly correct perception for the setting of *Agreeable*,

there exist possibilities that the participants are less able to perceive the differences between the dialogues generated with *Agreeable* personality and *Disagreeable* personality, as the percentage of selection of *Agreeable* or *Strongly Agreeable* for personality *Agreeable* (54.2%) and personality *Disagreeable* (57.0%) are similar.

In **User Study ii** (Figure 6.11), we observe the similar distributions of selection as **User Study i**. There exists a slight change between the selections of personality *Agreeable* and personality *Agreeable*, where more participants perceive a dialogue with personality *Agreeable* is *Agreeable* (48.6%) than perceive a dialogue with personality *Disagreeable* is *Agreeable* (45.5%).

Analysis

We present a summarised 1 tail T-test as Table 6.3 shows. In this table, the results are categorised by personality trait into selection distributions, which then are used for comparisons between two ends of each personality trait, in order to understand to how much extent the personalities can be correctly identified. For each personality trait, we hypothesise that the score distributions of positive ends (*Extravert, Emotionally stable, Agreeable*) would be greater than the negative ends (*Introvert, neurotic, Disagreeable*) as our conversion from ordinal scale to numerical scale of 5-Likert scale.

Overall in Table 6.3, we observe in both user studies, the p-values in T-tests indicate that our hypotheses are likely to be denied, i.e. the positive ends of personality traits

Table 6.3 1-tail T-test for personality identification on the perspective two extreme ends of each personality trait in two user studies. The setting references: Personality: P, No-Personality: NP, Source-Script: S.

	setting	setting #1	setting #2	pairwise	
	#1 vs #2	mean (std)	mean (std)	t-stats	p-value
User Study i	P(extra) vs P(intro)	0.551(0.91)	0.579(0.92)	-0.784	0.784
	P(emoti) vs P(neuro)	0.537(0.96)	0.566(0.96)	-0.770	0.779
	P(agree) vs P(disag)	0.521(0.89)	0.577(0.91)	-1.550	0.939
User Study ii	P(extra) vs P(intro)	0.517(0.98)	0.475(0.92)	0.620	0.268
	P(emoti) vs P(neuro)	0.540(1.01)	0.483(1.00)	0.780	0.218
	P(agree) vs P(disag)	0.547(0.97)	0.514(1.00)	0.484	0.314

are NOT significantly greater than the negative ends. Particularly in User Study i, the p-values are even greater than 0.5, which indicates that the participants identify personalities reversely, matching the mean values of distributions (the mean values of setting #1 are lower than ones of setting #2). In User Study ii, we notice the trend of correct identification as all three p-values (0.268, 0.218, 0.314) are smaller than 0.5, as well as the mean values of positive ends are greater than the ones of negative ends. From the perspective of personality, some possible reasons of this observation could be *Corpus 2* has more utterances with extreme personalities than *Corpus 1*, and the personality combinations with more extreme traits selected.

As shown in Figure 5.1 (right), in automatic evaluation, we notice that the system fine-tuned with *CTE* embedding method on film-level corpus has the worst personality identification accuracy, among all 6 settings. Therefore, it is understandable that the generated dialogues with this setting are less perceivable for human participants in terms of personality identification.

There is also a potential reason from the perspective of survey design. In Personality Identification of User Study i, it can be observed that participants tend to select one polar end of each personality trait (i.e. extravert end rather than introvert end, emotionally stable end rather than neurotic end, and agreeable end rather than disagreeable end). However, participants were asked to select an option towards a statement that only describes one polar end for each trait. For example, for statement "character B is extravert", participants are supposed to select "strongly disagree" if they judge character B having a very introvert personality.

We hypothesis that this design probably is one of the reasons that cause participants are inclined to select one polar end of each personality trait, which is the one we state in the question description, i.e. "Character B is **extravert**.", "Character B is **emotionally stable**.", and "Character B is **agreeable**.". In this case, participants are only informed explicitly one polar end of a personality trait, through text. Therefore, it is reasonable to presume that not all of the participants have a clear and correct

awareness that what the opposite polar end is in text. Or, even they do have the knowledge about what the opposite polar end is, this implicit description could also cause them to think twice, which may lead to random selecting of some impatient participants.

6.2.3 Genre Identification

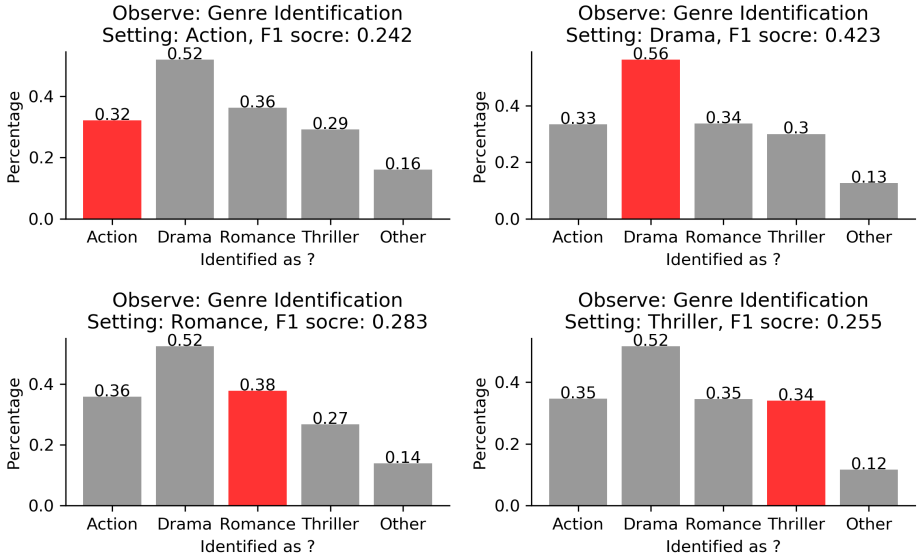


Fig. 6.12 The results of genre identification in **User study i**.

In the evaluation of genre identification, participants are required to select the most likely genre(s) for each dialogue. For each actual genre setting, participants are provided 5 options: Action, Drama, Romance, Thriller, and Other. The results are presented in Figure 6.12 for **User Study i** and in Figure 6.13 for **User Study ii**. In each figure, we show the percentage of genre selections and F1 score by genre.

In Figure 6.12, we show the selections of identification for all four genre settings as each sub-figure. Note the red bar denotes the selection of genre matches the genre setting, i.e. participants identify the "correct" genre. We notice the genre of drama has the highest identification precision (0.56) as well as highest F1 score (0.423). And the genre of action has the lowest identification precision (0.32) and F1 score (0.242). Also,

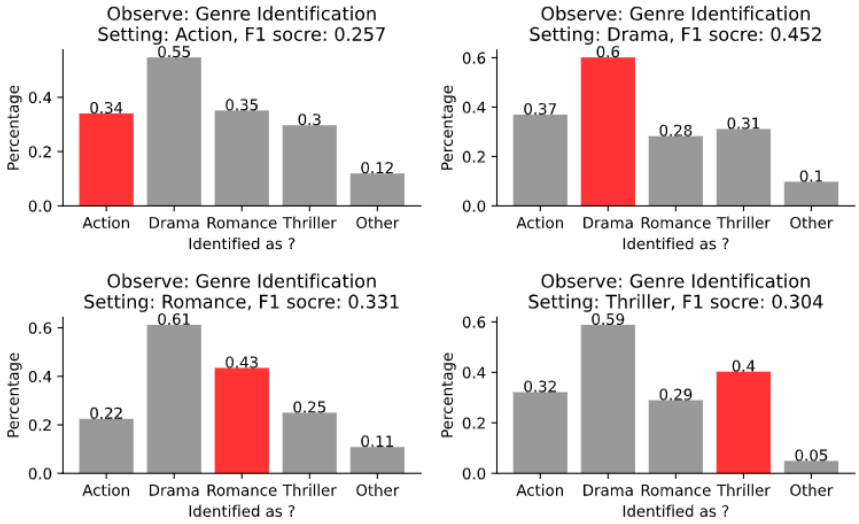


Fig. 6.13 The results of genre identification in **User study ii**.

expect the genre of drama, for the other three genres, the percentage of the correct identification precision is similar to each other.

In Figure 6.13, we show the selections of identification as the preceded one. We also notice the genre of drama has the highest identification precision (0.6) as well as highest F1 score (0.452). And the genre of action has the lowest identification precision (0.34) and F1 score (0.257). We also notice that for the genre of romance and thriller, although the precision of the correct selection is not the highest, it is significantly higher than the other options.

Comparing Figure 6.12 with Figure 6.13 (see Table 6.4), we observe for all four genres, the precision and F1 score have increased. Particularly, the F1 scores of the genre of romance and thriller have increased by 16.96% and 19.21%.

Analysis

We noticed for all sub-figures in Figure 6.12 and Figure 6.13, most participants select drama regardless which the actual genre is. A possible reason could be drama is a genre of broad scope which is able to include larger amount of films than other genres,

Table 6.4 The results of genre identification between **User study i** and **User study ii** on F1 score.

	Action	Drama	Romance	Thriller
User Study i	0.242	0.423	0.283	0.255
User Study ii	0.257	0.452	0.331	0.304
F1 Difference	+6.2%	+6.85%	+16.96%	+19.21%

such as romance and thriller, which are of a more narrow scope. Therefore, drama is probably a default selection for participants if they are indecisive. This reason can also explain the observation that the actual setting of drama has the highest identification precision than the other genre settings.

6.3 Summaries and Conclusions

In this chapter, we present two user studies we conduct for evaluating generated dialogues on film-level corpora using *CTE* embedding method, from the perspectives of the dialogue quality, personality identification, and genre identification.

Overall in two user studies, the results of dialogue quality show most participants (over 60%) have positive judgements on generated dialogues, and less than 10% of participants have negative judgements. Although the positive judgements for original script are slightly more than generated ones, this finding still indicates that our approach is able to generated dialogues with good quality.

For personality identification, all results show that participants are inclined to select one extreme end personality trait despite the settings, which does not meet our expectation. One reason is about the personality level and embedding method we use, and the other is about the survey design and deployment. Considering these observations and analysis, a further user study needs to be conducted for evaluating dialogues generated using scene-level personality as well as other embedding methods (e.g. *STE* and/or *SCE*).

As for genre identification, the results show, to some extents, the genre can be identified correctly with our approach and corpus. Particularly, genre of drama achieved

the highest identification performance regarding F1 scores. In the further user study, we consider to not include genre identification because 1) it is less relevant to our major research goal compared with personality identification. 2) The genre might not able to be learned properly as the neural network has not been fed with genre information explicitly but only been trained on sub-datasets of different genres.

Chapter 7

Human Evaluation - All Settings

In the Chapter 6, we present the details of the user studies as well as the results and analysis. As the dialogues for evaluation in those user studies are generated using *CTE* embedding method, it is necessary and reasonable to conduct a user study to evaluate dialogues generated with both film and scene level of personalities using various embedding methods and original language models.

7.1 Methodology

7.1.1 Metric and Measurement

In this user study, we evaluate dialogues from the perspectives of dialogue quality and personality identification. We do not include genre identification for this time as we only generate dialogues (See Appendix B.3) on drama dataset, which is the largest subset of our corpus. Also, for dialogue quality evaluation and the evaluation of each personality trait identification, we follow our previous strategy that to use 5-Likert scale to measure the judgement made by participants.

Differently from last user study, we use an overall dialogue quality (In terms of grammar, naturalness, and believability) instead of evaluating two aspects separately, which are grammatically correctness and naturalness. And we change the descriptions

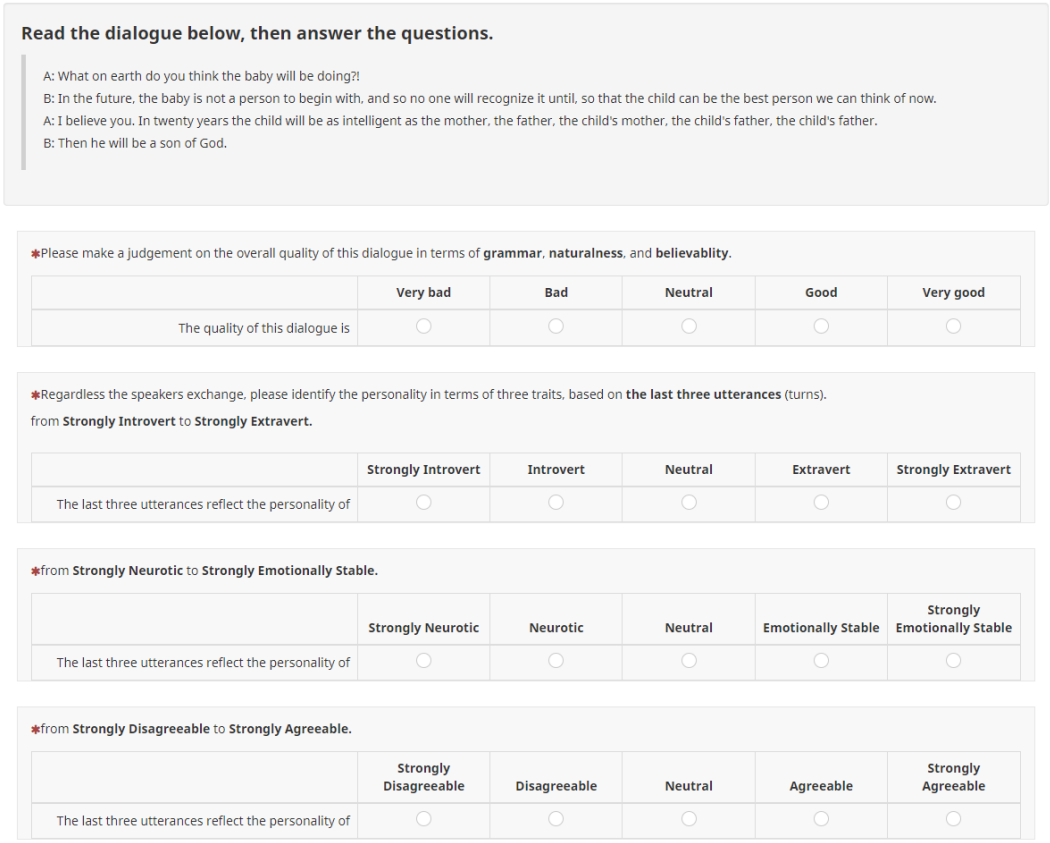


Fig. 7.1 A screenshot (one page) of the survey for **User Study iii**. Each page contains a piece of either generated dialogue or original written dialogue, a question for dialogue quality evaluation, and three questions for personality identification.

from a positive statement like "The dialogue is grammatically correct" with "Agree" and "Disagree" options, to a more neutral statement like "The quality of this dialogue is" with "Good" and "Bad" options. We believe to list the options more explicitly can help participants making judgements objectively.

For the identification of three personality traits, we follow the measurements used in previous user studies.

An example of the survey for this user study is shown as Figure 7.1.

7.1.2 Material for Evaluation

We select 2 sentences (one question, and one statement, see below) from the test set of our IMSDb corpus as the input seeds for follow-up generations.

1. *What on earth do you think the baby will be doing?!*
2. *All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars.*

For each input seed, we generate dialogues with 6 different embedding methods or dialogue generation approaches, as well as collect the dialogue started with identical input seed for further comparison.

There are two types of generation conducted, which are with personality and without personality. For the generation with personality, we use 3 different embedding methods (*CTE*, *SCE*, and *STE*) as introduced in Chapter 4, plus use StyleDGPT (Yang et al., 2020), an advanced stylistic dialogue generation system. And for the generation without personality, we use the original DialoGPT (Zhang et al., 2020) as well as the DialoGPT fine-tuned on our IMSDb corpus.

For the personality trait combination, instead of using exhaustive eight combinations, two extreme combinations are selected for this user study, i.e. (extravert, emotionally stable, agreeable) and (introvert, neurotic, disagreeable). Also for the personality level, both film-level and scene-level personalities are used for training process and the generations with these two levels personality are evaluated.

Therefore, for each method of generation with personality (4 in total) and each input seed (2 in total), dialogues are generated with two levels personality and two personality trait combinations, which lead to $2 \times 2 \times 2 \times 4 = 32$ dialogues. And for the generations (3 in total) without personality, we have $2 \times 3 = 6$ dialogues. Particularly, two dialogues generated by the fine-tuned DialoGPT are repeated for validating the participants' attention. Consequently, each participant is supposed to evaluate 40 dialogues in total.

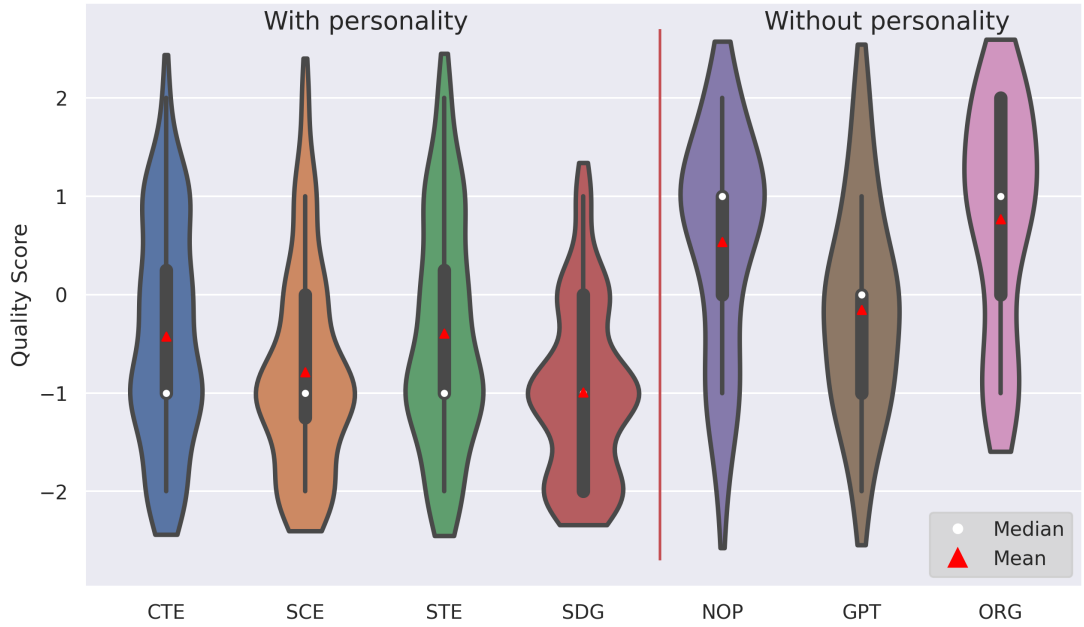


Fig. 7.2 The score distribution of dialogue quality with different settings in **User Study iii**. The white dots denote the median values of the distributions, while the red triangles denote the mean values (same as the figures after).

7.2 Results

13 students in university, who are native English speakers or proficient English users participated this user study. We distributed our survey through two ways: 1) sending emails to a designated student group in which the students are all native English users that we have confirmed in advance[1]. 2) Posting adverts publicly in the Bournemouth University and Arts University Bournemouth. We assume the participants recruited through the second way who honestly qualified themselves proficient English speakers with the principle of *assuming good faith* as previous user studies. In this section, we present the results and analysis of their responses. All the acronyms of settings used in this section are list as below:

- With personality embedded
  - Ours - **CTE**
  - Ours - **SCE**

[1]They still participated the user study voluntarily and their answers were recorded anonymously.

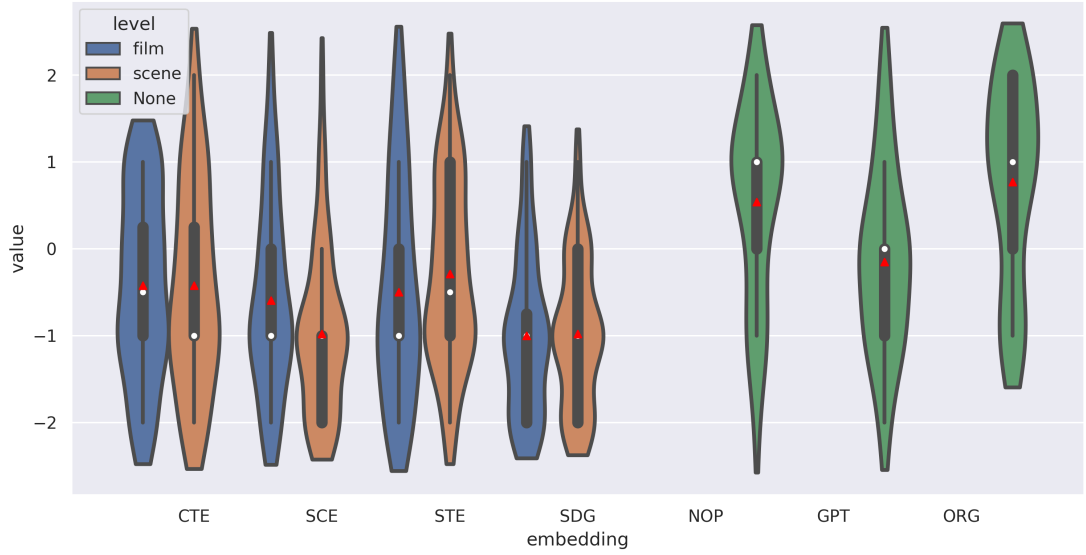


Fig. 7.3 The score distribution of dialogue quality from perspective of personality level with different settings in **User Study iii**.

  - Ours - **STE**
  - StyleD(ialog)GPT - **SDG**

- Without personality embedded

  - Original DialoGPT trained on our corpus (No Personality) - **NOP**
  - Original DialoGPT - **GPT**
  - Original Screenplay - **ORG**

7.2.1 Dialogue Quality

We present score distributions for dialogue quality in Figure from a broad perspective of setting as Figure 7.2 shows. overall, generations without personality style (GPT, NOP, ORG) have higher scores of dialogue quality than those with personality style

Table 7.1 1-tail T-test for dialogue quality evaluation. Comparisons between two personality level for 4 embedding methods in **User Study iii**.

embedding method	setting #1 vs #2	setting #1 mean (std)	setting #2 mean (std)	pairwise t-stats	pairwise p-value
Overall	CTE(film) vs CTE(scene)	-0.423(1.054)	-0.423(1.177)	0.0	0.5
	SCE(film) vs SCE(scene)	-0.596(1.071)	-0.984(0.939)	1.947	**0.027**
	STE(film) vs STE(scene)	-0.5(1.229)	-0.288(1.054)	-0.942	0.826
	SDG(film) vs SDG(scene)	-1.0(0.907)	-0.981(0.828)	0.113	0.545

Table 7.2 1-tail T-test results for dialogue quality evaluation in **User Study iii**. Results are presented by personality level and by the comparisons between ours and StyleDGPT.

embedding method	setting #1 vs #2	setting #1 mean (std)	setting #2 mean (std)	pairwise t-stats	pairwise p-value
film-level	CTE vs SDG	-0.423(1.054)	-1.0(0.907)	2.991	**0.001**
	SCE vs SDG	-0.596(1.071)	-1.0(0.907)	2.074	**0.020**
	STE vs SDG	-0.5(1.229)	-1.0(0.907)	2.360	**0.010**
scene-level	CTE vs SDG	-0.423(1.177)	-0.981(0.828)	2.794	**0.003**
	SCE vs SDG	-0.984(0.939)	-0.981(0.828)	0.0	0.5
	STE vs SDG	-0.288(1.054)	-0.981(0.828)	3.724	**0.0002**

Table 7.3 1-tail T-test results for dialogue quality evaluation in **User Study iii**. Results are presented by personality level and by the comparisons between three embedding methods used in our approach pairwisely.

embedding method	setting #1 vs #2	setting #1 mean (std)	setting #2 mean (std)	pairwise t-stats	pairwise p-value
film-level	CTE vs SCE	-0.423(1.054)	-0.596(1.071)	0.830	0.204
	SCE vs STE	-0.596(1.071)	-0.5(1.229)	-0.425	0.664
	STE vs CTE	-0.5(1.229)	-0.423(1.054)	-0.343	0.634
scene-level	CTE vs SCE	-0.423(1.177)	-0.984(0.939)	2.670	**0.004**
	SCE vs STE	-0.984(0.939)	-0.288(1.054)	-3.536	0.999
	STE vs CTE	-0.288(1.054)	-0.423(1.177)	0.614	0.270

(CET, SCE, STE, SDG). For the dialogues without personality style, it is observed that the dialogues collected from original screenplays receive the highest median and mean scores. While NOP has significantly higher median and mean scores than GPT, which indicates that fine-tuning pre-trained DialoGPT on our IMSDb corpus is able to increase the quality of generation. For the generations without personality style, we notice that the generations using CTE and STE have higher mean scores than the generations using SCE and SDG.

Furthermore, we present the results of dialogue quality in a deeper perspective of personality level as Figure 7.3 and Table 7.1 illustrate. In the figure, we observe that for CTE and SDG, the generations with film-level and scene-level personality have the similar distributions. And for SCE, the generations with film-level personality have higher scores than ones with scene-level personality, whilst an opposite observation is observed for STE.

In Table 7.2, the results of the dialogue quality for the comparisons between our three embedding methods and StyleDGPT are presented. We observe that except

SCE with scene-level personality, all the other comparisons show that our approach performs better than StyleDGPT, as the significant p values (<0.05) indicated.

We also compare our three embedding methods pairwisely to evaluate using which embedding method can generate dialogues with better quality. In Table 7.3, it is observed that for both film-level and scene-level personality, CTE and STE have relatively better dialogue quality than SCE. while for the pair of CTE and STE, different dialogue quality performances are noticed for two levels of personality.

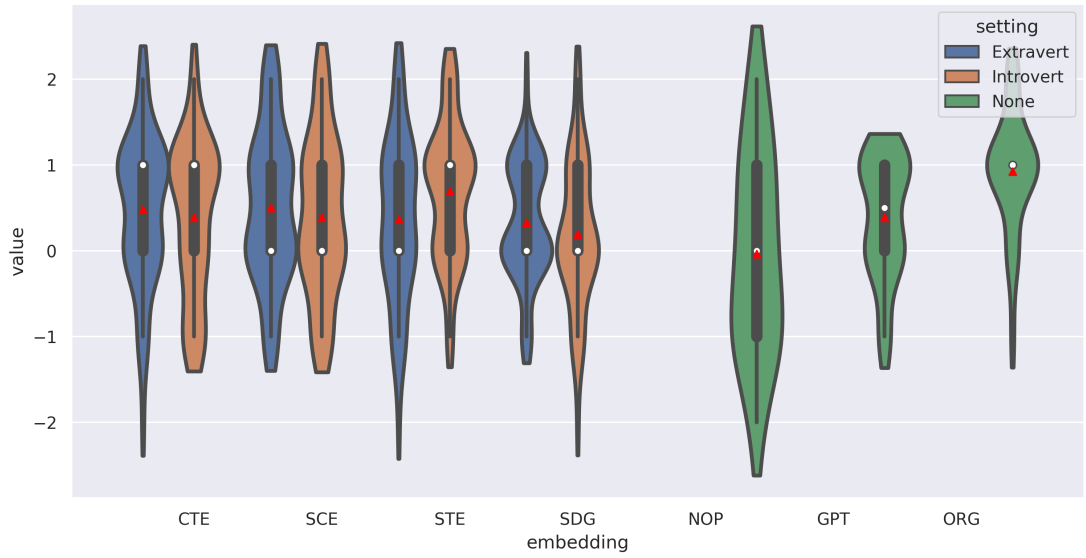


Fig. 7.4 The score distribution of personality identification from perspective of personality level with different embedding methods in **User Study iii**. Results are grouped by with comparisons between target personalities of extravert and introvert.

7.2.2 Personality Identification

Extraversion vs. Introversion

We present score distributions and comparisons of personality identification for dialogues with target personalities of extraversion and introversion from a broad perspective of setting as Figure 7.4 shows. Overall, it can be observed that all 4 comparisons of the generations with personality style do not indicate expected identification results, i.e. the generations with extravert personality have higher scores than introvert ones.

In Figure 7.5, we decompose the distributions for each generation setting from the perspective of personality level. We observe that CET-film, SCE-scene, and SDG-film have the expected identification results according to the mean values, in which SCE-scene has the lowest p-value in T-test (Table 7.4).

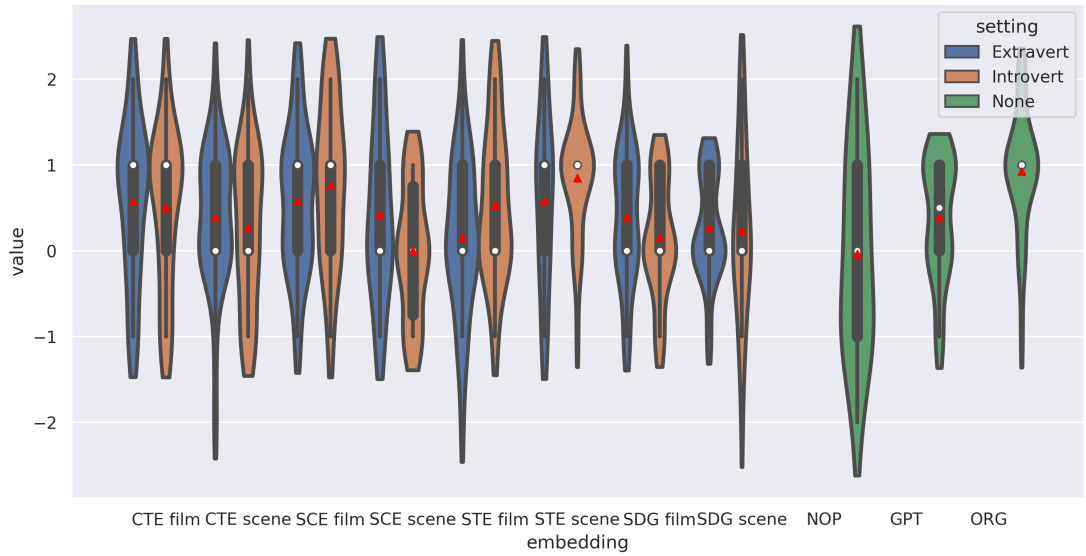


Fig. 7.5 The score distribution of personality identification from perspective of personality level with different embedding methods and personality levels in **User Study iii**. Results are grouped by with comparisons between target personalities of extravert and introvert.

Table 7.4 1-tail T-test results for personality identification evaluation. Results are presented by personality level and by the comparisons between extravert(extra) and introvert(intro) across all 4 embedding methods in **User Study iii**. Lower p-value denotes higher identification precision.

embedding method	setting #1 vs #2	setting #1 mean (std)	setting #2 mean (std)	pairwise t-stats	pairwise p-value
film-level	CTE(extra) vs CTE(intro)	0.577(0.902)	0.5(0.906)	0.307	0.380
	SCE(extra) vs SCE(intro)	0.577(0.809)	0.769(0.908)	-0.806	0.788
	STE(extra) vs STE(intro)	0.154(0.881)	0.538(0.859)	-1.594	0.941
	SDG(extra) vs SDG(intro)	0.385(0.752)	0.154(0.675)	1.164	0.125
scene-level	CTE(extra) vs CTE(intro)	0.385(0.804)	0.269(0.874)	0.495	0.311
	SCE(extra) vs SCE(intro)	0.423(0.945)	0.0(0.748)	1.789	**0.040**
	STE(extra) vs STE(intro)	0.577(0.945)	0.846(0.675)	-1.182	0.878
	SDG(extra) vs SDG(intro)	0.269(0.604)	0.231(0.992)	0.169	0.433

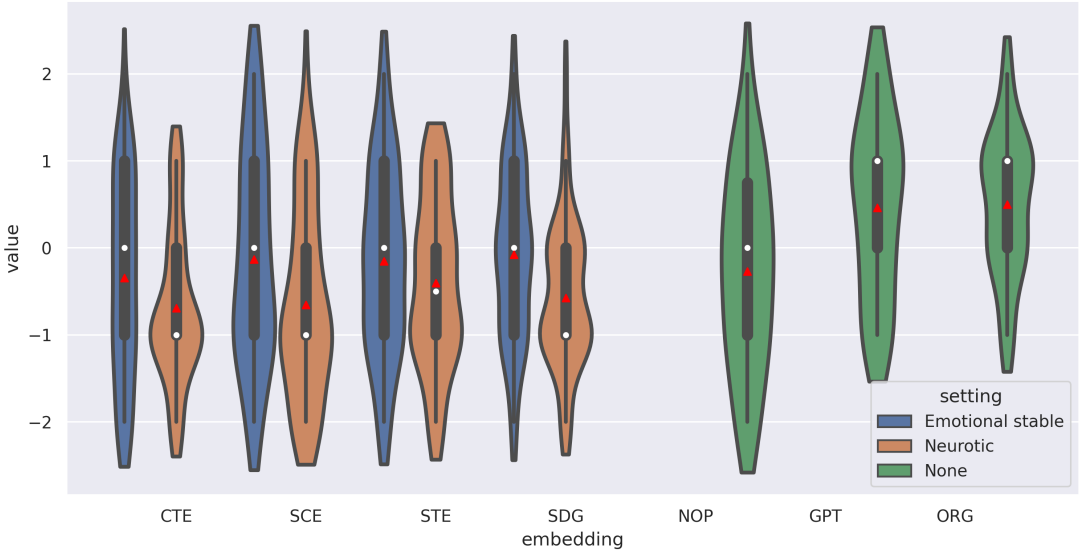


Fig. 7.6 The score distribution of personality identification from perspective of personality level with different embedding methods in **User Study iii**. Results are grouped by with comparisons between target personalities of emotionally stable and neurotic.

Emotional Stableness vs. Neurotics

We present score distributions and comparisons of personality identification for dialogues with target personalities of emotional stableness and neurotics from a broad perspective of setting as Figure 7.6 shows. Overall, it can be observed that all 4 comparisons of the generations with personality style are able to indicate the expected identification results, i.e. the generations with emotionally stable personality have higher scores than neurotic ones.

In Figure 7.7, we decompose the distributions for each generation setting from the perspective of personality level. We observe that almost all comparisons indicate the expected identification results according to the mean values, except SDG-film. Among these, STE in film-level, SCE in scene-level, and SDG in scene-level have the low p-value in T-test, which denotes the differences between dialogues with emotionally stable and neurotic personality are significant (Table 7.5).

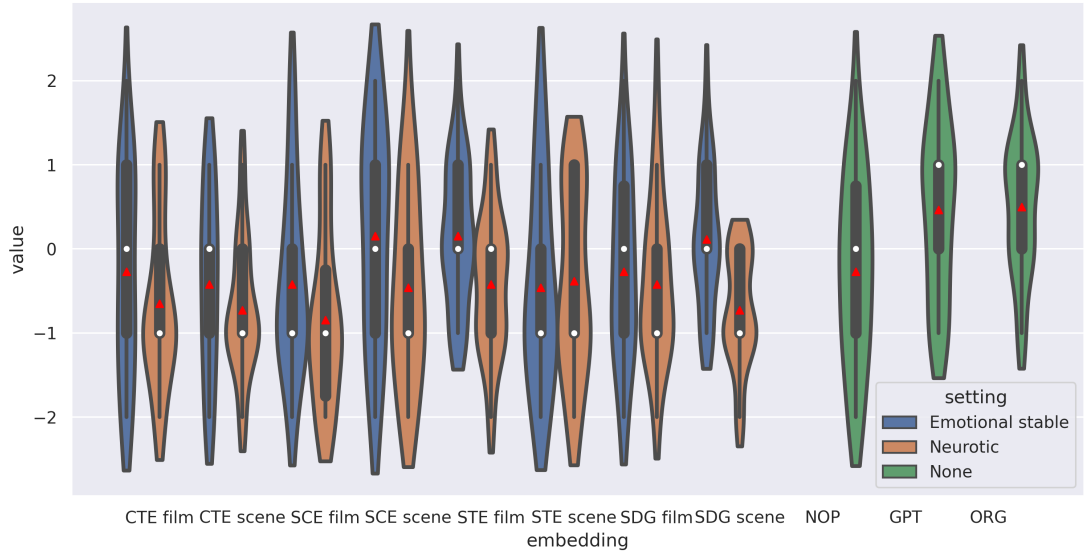


Fig. 7.7 The score distribution of personality identification from perspective of personality level with different embedding methods and personality levels in **User Study iii**. Results are grouped by with comparisons between target personalities of emotionally stable and neurotic.

Table 7.5 1-tail T-test results for personality identification evaluation in **User Study iii**. Results are presented by personality level and by the comparisons between emotionally stable(emoti) and neurotic(neuro) across all 4 embedding methods in **User Study iii**. Lower p-value denotes higher identification precision.

embedding method	setting #1 vs #2	setting #1 mean (std)	setting #2 mean (std)	pairwise t-stats	pairwise p-value
film-level	CTE(emoti) vs CTE(neuro)	-0.269(1.218)	-0.654(0.977)	1.256	0.108
	SCE(emoti) vs SCE(neuro)	-0.423(1.102)	-0.846(1.008)	1.445	0.077
	STE(emoti) vs STE(neuro)	0.154(0.834)	-0.423(0.809)	2.533	**0.007**
	SDG(emoti) vs SDG(neuro)	-0.269(1.079)	-0.423(0.945)	0.547	0.294
scene-level	CTE(emoti) vs CTE(neuro)	-0.423(1.065)	-0.731(0.778)	1.190	0.120
	SCE(emoti) vs SCE(neuro)	0.154(1.287)	-0.462(1.14)	1.826	**0.037**
	STE(emoti) vs STE(neuro)	-0.462(1.208)	-0.385(1.098)	-0.240	0.594
	SDG(emoti) vs SDG(neuro)	0.115(0.816)	-0.731(0.667)	4.094	$\mathbf{8 \times 10^{-5}}$

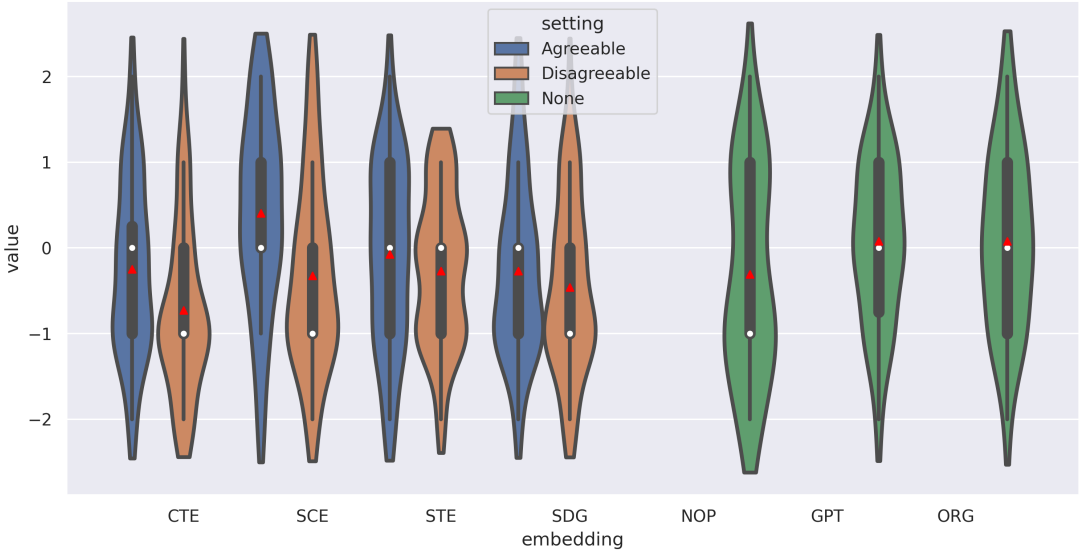


Fig. 7.8 The score distribution of personality identification from perspective of personality level with different embedding methods in **User Study iii**. Results are grouped by with comparisons between target personalities of agreeable and disagreeable.

Agreeableness vs. Disagreeableness

We present score distributions and comparisons of personality identification for dialogues with target personalities of emotional stableness and neurotics from a broad perspective of setting as Figure 7.8 shows. Overall, it can be observed that all 4 comparisons of the generations with personality style are able to indicate the expected identification results, i.e. the generations with agreeable personality have higher scores than disagreeable ones. However, regarding STE and SDG, the differences between two extreme personality are not significant.

In Figure 7.9, we decompose the distributions for each generation setting from the perspective of personality level. We observe that comparisons of CTE-film, CTE-scene, SCE-scene, and STE-film receive the expected identification results according to the mean values. Among these, CTE in film-level, STE in film-level, and SCE in scene-level have the low p-value in T-test, which denotes the differences between dialogues with agreeable and disagreeable personality are significant (Table 7.6).

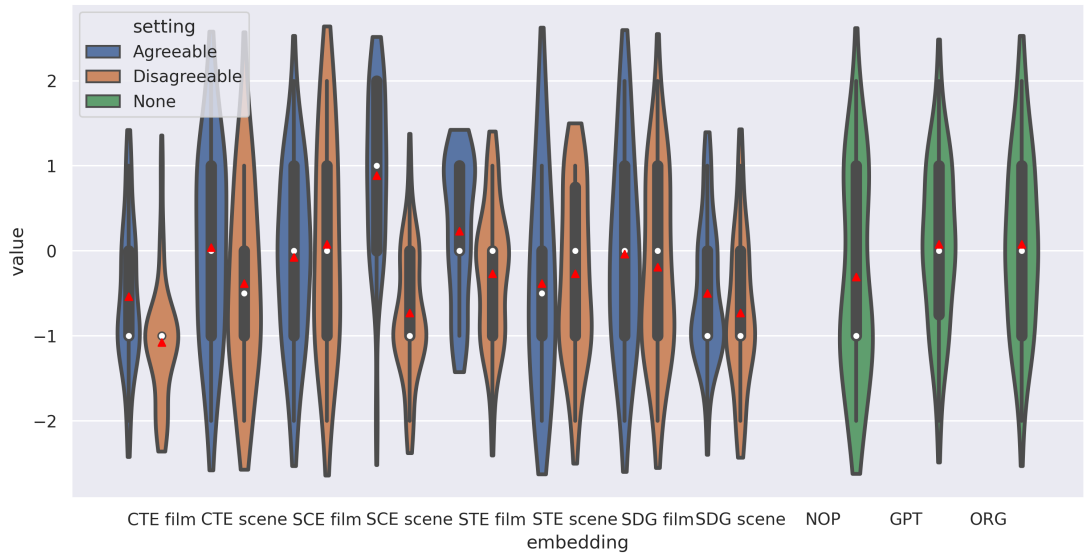


Fig. 7.9 The score distribution of personality identification from perspective of personality level with different embedding methods and personality levels in **User Study iii**. Results are grouped by with comparisons between target personalities of agreeable and disagreeable.

Table 7.6 1-tail T-test results for personality identification evaluation in **User Study iii**. Results are presented by personality level and by the comparisons between agreeable(agree) and disagreeable(disag) across all 4 embedding methods. Lower p-value denotes higher identification precision.

embedding method	setting #1 vs #2	setting #1 mean (std)	setting #2 mean (std)	pairwise t-stats	pairwise p-value
film-level	CTE(agree) vs CTE(disag)	-0.538(0.811)	-1.077(0.688)	2.580	**0.006**
	SCE(agree) vs SCE(disag)	-0.077(1.017)	0.077(1.23)	-0.491	0.687
	STE(agree) vs STE(disag)	0.231(0.815)	-0.269(0.778)	2.263	**0.014**
	SDG(agree) vs SDG(disag)	-0.038(1.148)	-0.192(1.059)	0.502	0.309
scene-level	CTE(agree) vs CTE(disag)	0.038(1.113)	-0.385(1.098)	1.380	0.087
	SCE(agree) vs SCE(disag)	0.885(0.993)	-0.731(0.724)	6.701	**0.0**
	STE(agree) vs STE(disag)	-0.385(1.203)	-0.269(0.962)	-0.382	0.648
	SDG(agree) vs SDG(disag)	-0.5(0.762)	-0.731(0.827)	1.046	0.150

7.2.3 Correlation between Quality and Personality

Following the observation that adding personality decrease the ability of generation in Chapter 5, we hypothsise that to some extent, dialogue quality display a negatively correlation with the precision of personality identification. Therefore, we investigate the correlation between the dialogue quality and the precision of personality identification from the perspective of personality level.

To calculate the correlation between these two criteria, we build up samples for correlation tests in two different ways. The first way (**Correlation 1**) is to calculate the mean values of each distribution categorised by embedding method and personality level for dialogue quality (e.g. *SCE* with film-level personality in Figure 7.3), as well as the mean values of personality identification precision, which are defined by calculating the difference between the score distributions of two ends for each personality trait (e.g. For *SCE* with film-level personality in Figure 7.5, the precision is calculated by subtracting the mean value of introvert from the mean value of extravert. And the minus result indicates that unexpected precision, which is expected to be positive).

The second way (**Correlation 2**) is to build up samples for correlation tests on the basis of the first one. After calculating those mean values, we calculate the difference of the mean values between two levels of personality. For example, for the generation setting *SCE*, we notice that quality with film-level personality is higher than that with scene-level personality. But for personality identification of the pair of extraversion and introversion, we observe that generations with scene-level personality have higher

	Spearman Corr		Pearson Corr		Kendall Corr	
	test	p-value	test	p-value	test	p-value
Correlation 1	-0.288	0.172	-0.398	0.054	-0.222	0.155
Correlation 2	-0.583	**0.047**	-0.716	**0.009**	-0.469	**0.047**

Table 7.7 The correlation of dialogue quality and personality identification precision using 3 correlation coefficient.

precision, i.e. scores of extravert are higher than those of introvert. Therefore, this is to calculate the correlation of mean value differences on the personality level.

We present the 3 correlation test results for each as Table 7.7 shows. For the **Correlation 1**, we noticed negative test results for all three correlation methods. Although the p-values are slightly higher than a significance threshold (0.05), they are indicating a tendency of negative correlation between dialogue quality and personality identification precision from the perspective of embedding method with personality level. While as for the **Correlation 2**, we observe stronger negative correlations with significant p-values, which indicates the rationality of the first observation, but also the difference between two levels of personality.

7.3 Summaries and Conclusions

In this user study, we evaluate dialogues by comparing generated dialogues with personality style using different embedding methods and pre-trained language model, as well as generated dialogues without personality style and human written dialogues. To analyse the results of the user study, we focus on the responses from 2 categories as below;

- Dialogue quality
- Personality identification
  - Extraversion vs. Introversion
  - Emotional stableness vs. Neurotics
  - Agreeableness vs. Disagreeableness

And from 2 perspectives of settings as below:

- 7 Dialogue generation settings (as mentioned before)
- personality levels
  - film-level personality
  - scene-level personality

Table 7.8 1-tail T-test results for personality identification evaluation in **User Study iii**. 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).

embedding 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)	0.481(0.852)	0.385(0.889)	0.563	0.287
	emoti (52) vs neuro (52)	-0.346(1.136)	-0.692(0.875)	1.741	**0.042**
	agree (52) vs disag (52)	-0.25(1.007)	-0.731(0.972)	2.476	**0.007**
SCE	extra (52) vs intro (52)	0.5(0.874)	0.385(0.911)	0.659	0.256
	emoti (52) vs neuro (52)	-0.135(1.221)	-0.654(1.083)	2.294	**0.012**
	agree (52) vs disag (52)	0.404(1.107)	-0.327(1.08)	3.408	**0.0005**
STE	extra (52) vs intro (52)	0.365(0.929)	0.692(0.781)	-1.942	0.973
	emoti (52) vs neuro (52)	-0.154(1.073)	-0.404(0.955)	1.255	0.106
	agree (52) vs disag (52)	-0.077(1.064)	-0.269(0.866)	1.011	0.157
SDG	extra (52) vs intro (52)	0.327(0.678)	0.192(0.841)	0.899	0.186
	emoti (52) vs neuro (52)	-0.077(0.967)	-0.577(0.825)	2.836	**0.003**
	agree (52) vs disag (52)	-0.269(0.992)	-0.462(0.979)	0.995	0.161

Overall in the category of quality, dialogues without personality style (GPT, NOP, ORG) have significantly higher mean and median scores than those with personality style (CET, SCE, STE, SDG). For the generations with personality style, apart from original screenplay, which receive the highest scores reasonably, NOP has significantly higher scores than GPT, which indicates that fine-tuning pre-trained DialoGPT on our IMSDb corpus is able to increase the quality of generation.

In the category of personality identification, we present the statistics of what personality participants make judgements on the dialogues with different target personality. By examining the distributions as shown before and a summarised statistics as Table 7.8 shows, we notice that overall, almost all the personality of dialogues that are generated with all embedding methods are inclined to be correctly identified (except STE for extravert and introvert with negative T-test stats), according to the positive T-test stats and p-values much lower than 0.5. More specifically, we notice that the order of all 4 embedding methods in terms of identification precision is SCE>CTE>SDG>STE. From the perspective of personality, we observe that the difference between emotional stableness and neurotics is more significant than the differences of other two pairs, which denotes emotional stableness and neurotics can

be distinguished more correctly in textual dialogues. While the pair of extraversion and introversion has an opposite difference.

Regarding the correlation between dialogue quality and personality identification accuracy, negative correlations between them is able to be observed across all embedding methods. This observation indicates that incorporating additional features could affect dialogue quality in the context of narrative. Two possible reasons are considered as potential cause here. First, because the scores of personality are calculated using linguistic cues, the fine-tuned neural networks could be more likely to generate certain words to match the given target personality. Therefore, the possibilities exist that the neural networks generate words that are "more correct" in terms of target personality while "less correct" in terms of quality. Secondly, in this user study, we evaluate the overall dialogue quality as a simplification. However, this observation of negative correlation might not remain if we split the dialogue quality into finer-grained perspectives. Therefore, the correlations between personality identification and the different aspects of dialogue quality need to be investigated further.

Chapter 8

Conclusions and Discussions

8.1 Summarised Conclusions

Our goal of this thesis is to investigate the potential and the impact of stylistic conditional dialogue generation based on different characters' personalities derived from narrative films. On top of that, we intend to expand the ability of generation in narrative systems by leveraging NLG techniques. We propose an approach for generating dialogues using a pre-trained neural language model with target personality derived from the characters in film screenplays.

To answer our first research question,

> *How to reflect authorial intentions on characters' personalities from narratives and how to incorporate them into deep neural networks?*

We create our corpus based on textual screenplays in IMSDb including dialogue text along with characters' personalities on both film-level and scene-level, which are used to fine-tune the pre-trained language model (DialoGPT). According to narrative theories, we believe that the authorial intentions on character's personalities are revealed through textual dialogues in screenplays. And we apply two levels of personalities in our experiments and analysis. More specifically, we use three different embedding

methods to incorporate both scene-level and film-level personality of characters into transformer-based neural networks, which are *CTE*, *SCE*, and *STE* namely.

To answer our second and third research questions, which are

> *What are the influences on dialogue generation by adding different characters' personalities derived from narratives using deep learning techniques?*

and

> *What are the differences on dialogue generation of the influence by using different embedding methods and datasets for characters' personalities?*

We evaluate our proposed personalised dialogue generation methods using both automatic metrics (Chapter 5) and human judgements (Chapter 6 and Chapter 7) from objective and subjective perspectives, which is a widely used evaluation strategy in NLP/NLG.

To measure this approach automatically, we generate dialogues with all setting combinations . And we evaluate them on various metrics and analyse the results from different perspectives. From the results of automatic evaluation, we demonstrate that our approach is able to generate dialogues with the correct target personality, by separately embedding the personality and using half pre-trained transformer framework. Also, we find our approach is able to generate dialogues with increased variety on surface-text level and semantic-level comparing with the original DialoGPT. However, we also observe that the generations with personality contain less words and sentence, which might indicate a decrease of the capability of generation.

In Chapter 6, we evaluated the dialogues generated with *CTE* embedding method and film-level personality. The results of quality evaluation show that most participants (over 60%) have positive judgements on generated dialogues. However, for the evaluation of personality identification, human participants are inclined to have similar perceptions of both extreme ends of each personality trait correctly, which

is corresponding to the results of automatic evaluation. For the evaluation of genre identification, the results show that most dialogues are most likely to be identified as with genre of drama, while the second most likely identifications are corresponding to the target genres.

In Chapter 7, we synthetically investigate the impact on dialogue generation using different embedding methods and personality levels by conducting human evaluation. According to the results (Table 7.8), the personality of dialogues generated with almost settings are likely to be correctly identified, which indicates that our corpus with our definition of characters' personality is effective, i.e. can be perceived by human. We also notice that our approach using explicit personality embedding methods (*SCE* and *CTE*) performs better than StyleDGPT (*SDG*) in terms of personality identification, while STE has the worst performance compared with them. Comparing *CTE*/*SCE* and *SDG*, where a significant difference between them is *SDG* uses an implicit way to embed styles and *CTE*/*SCE* use explicit way. Although it is difficult to clearly interpret the exact reason for this observation, we are able to come up with a possible reason that features with quantitative definition (e.g. our personality) are more likely to be learned using explicit embedding methods, rather than implicit embedding methods, which might be effective for other features like writing style. Also, a negative correlation between generated dialogue quality and personality identification precision is able to be observed across all embedding methods. This observation indicates that incorporating additional features could affect dialogue quality in the context of narrative.

To summarise the these evaluations, *SCE* has the best accuracy of personality identification according to automatic metrics and human judgements, and *CTE* has the best dialogue quality according to both types of evaluation. From the perspective of personality trait, Extraversion is more likely to be identified correctly on automatic metrics, while Emotional Stableness tends to be identified more correctly under human judgements. We also notice that scene-level personalities receive higher accuracy of

identification than film-level personality on automatic metrics. While for dialogue variety, film-level personality performs better than scene-level personality.

8.2 Contributions

We would claim again that our research could contribute to both narrative community and natural language generation community from the following aspects:

1. An approach for generating conditional dialogues by utilising Big-Five model based personality traits from film screenplays. Our approach based on three embedding methods can generate varied dialogues which are able to reflect selected target personality traits.
2. Experiments and detailed analysis of the impact of personality combinations, levels of personality, and embedding methods on the performance of the dialogue generation.
3. A well parsed, segmented, and labelled dataset from IMSDb, which contains dialogues in screenplays, characters, scenes and corresponding personalities.

8.3 Discussion and Future Work

During the progress of this research, we noticed that there exist some limitations and we intend to discuss them here. For example, the lack of methods for evaluation. As we mentioned in Chapter 5, the existing metrics for automatically evaluating generated text are highly depending on golden references, which makes it less applicable to stylistic or conditional text generation. Moreover, added stylistic conditions are not likely to be measured precisely using automatic metrics. Therefore, human evaluation is required for creative content generation.

Regarding the results of our user studies, we notice that there exist some limitations. The user studies include differing sets of questions evaluating the specific generation

methods. The quality of the respondents' answers is variant due to having used an online platform initially, following by improving participants' recruitment through in-person evaluation. Considering these changing factors, it is difficult to aggregate evaluation results due to the different evaluation settings. Particularly, it is more difficult to recruit reliable human participants online rather than on campus. Because of the natures of our evaluation, human participants are asked to make judgements on several dialogues with same questions. This process tends to make them feel bored over time during the user study, which might affect the believability of the results. Because of campus shutdown due to Covid-19 pandemic, the participants were recruited through Amazon Mechanical Turk for the first two user studies. This might also be the reason to explain that some distributions of generations with *SCE* and film-level personality in **User Study iii** show better results than the first two user studies in terms of personality identification precision (e.g. emotional stableness vs. neurotics and agreeableness vs. disagreeableness).

From the perspective of personality trait, Mairesse et al. (2007) showed that among all five traits, Extraversion is the trait that is significantly correlated to the linguistic cues which can be quantified. This likely explains that Extraversion achieves higher identification accuracy in the evaluation using automatic metrics in Figure 5.2. However, we observe that there exist diverse results regarding the identification accuracy in human evaluation in Table 7.8, which likely indicates that from textual content, humans are more likely to perceive Emotional Stableness and Agreeableness rather than Extraversion.

We also note that linguistic features of personality are indirectly reflected in speaking or writing (Gatt and Krahmer, 2018). This might also be the reason for an observation that readers vary significantly in their judgements of personality in text (Mairesse and Walker, 2011). In our work, we represent characters' personalities explicitly for easier interpretation. However, there are other solutions generating dialogues using additional styles with implicit representations, such as StyleDGPT,

which trained models on individual sub-corpus with different styles (personalities). We compare the dialogues generated using personalities represented in these different ways, and notice that explicit representations of personality tend to have higher personality identification precision. In the future, it could be worth exploring the potential of combining explicit and implicit representations of personality for generation.

We observe that the accuracy of scene-level personality is improved compared to film-level personality, with the possible reason that the overall labelled personality could not match a finer-grained utterance perfectly and directly, although from narrative theories, a character's personality is supposed to remain consistent over the duration of the story. Also, according to narrative theories and the constitution of narrative films, there exist many elements apart from dialogues, characters, and transitions. Currently, we only leverage these three to define the personality, discarding some other elements of screenplays, such as staging or directions. These elements also contain essential information reflecting authorial intentions and the progress of storylines, which could be able to set the scenario context of dialogues, as well as affect "what" and "how" characters speak. Therefore, it could be a potential direction of our future work that to investigate the impact on textual dialogue generation with additional narrative elements.

We acknowledge that research in conditional dialogue generation is a developing topic with enormous challenges to effectively represent and convey the desired conditions. However, we believe that such research is promising to conduct within the context of narrative-based productions in entertainment and education. As for some specific applicable scenarios, it is also promising that our research can be applied in plenty of AI-related potential scenarios. For example, in a narrative-based video game (e.g. *Life is Strange*) a character's dialogue could be improved through more believable responses if the dialogue generation incorporates richer character's attributes. This can still be a process working in real-time through leveraging deep learning techniques, as this is a requirement for the players' interaction with game

characters. Therefore, game players expect to be shown different dialogue expressions every time they play games rather than a limited selection of pre-authored utterances. This can not only improve players' experience, but also benefit screenplay authors by allowing the creation of larger sets of conversations with different characters. Authors can identify and select the most appropriate sentences from the generated ones with corresponding character attributes and previous storyline progress. van Stegeren and Myśliwiec (2021) also points out that transfer learning on pre-trained language models is a feasible alternative to provide descriptions of quests for game designers or writers in video games with creativity. EU projects also focused on these topics and created applications regarding personalised generation or narratives (Exus Software Ltd., 2016-2019; Nottingham University, 2015-2017). For example, The University of Nottingham leads a project (Nottingham University, 2015-2017) that creates Artificial Retrieval of Information Assistants (ARIAs), which is with the capacity of multi-modal interaction with users. ARIAs can capture user's verbal and non-verbal behaviour, and generate reactions (e.g. a sentence, a smile, etc.) decided by a management system together with emotive personality model.

Therefore, we believe that the outcomes of this research can benefit such potential applications by improving narrative dialogue generation through enriching the incorporated attributes of characters from narrative perspective using deep learning techniques from technical perspective.

References

Adams, E., 2014. *Fundamentals of Game Design*, 3rd edn, New Riders Publishing, USA.

Allport, G. W., 1961. *Pattern and growth in personality.*, Holt, Reinhart & Winston.

Allport, G. W. and Odbert, H. S., 1936. Trait-names: A psycho-lexical study., *Psychological monographs* **47**(1), i.

Ammanabrolu, P., Tien, E., Cheung, W., Luo, Z., Ma, W., Martin, L. J. and Riedl, M. O., 2020. Story realization: Expanding plot events into sentences, *In: Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 7375–7382.

Arendt, H., 1968. *Men in dark times*, Houghton Mifflin Harcourt.

Aumont, J., 1992. *Aesthetics of Film*, Texas Film and Media Studies Series, University of Texas Press.

Bahdanau, D., Cho, K. H. and Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate, *In: 3rd International Conference on Learning Representations, ICLR 2015*.

Bal, M. and Van Boheemen, C., 2009. *Narratology: Introduction to the theory of narrative*, University of Toronto Press.

Bednarek, M., 2017. The role of dialogue in fiction, *Pragmatics of fiction* pp. 129–158.

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C., 2003. A neural probabilistic language model, *Journal of machine learning research* **3**(Feb), 1137–1155.

Bengio, Y. and Lecun, Y., 2007. Scaling learning algorithms towards ai, *In: Large-scale kernel machines*, MIT Press.

Berliner, T., 1999. Hollywood movie dialogue and the" real realism" of john cassavetes, *FILM QUARTERLY-BERKELEY-* **52**, 2–16.

Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H. and Winograd, T., 1977. Gus, a frame-driven dialog system, *Artificial intelligence* **8**(2), 155–173.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue,

C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K. and Liang, P., 2021. On the opportunities and risks of foundation models, *arXiv preprint arXiv:2108.07258* .

Bordwell, D., Thompson, K. and Smith, J., 2020. *Film Art: An Introduction (12th Edition)*, McGraw-Hill Education.

Bowden, K. K., Lin, G. I., Reed, L. I., Tree, J. E. F. and Walker, M. A., 2016. M2d: Monolog to dialog generation for conversational story telling, *In: International Conference on Interactive Digital Storytelling*, Springer, pp. 12–24.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D., 2020. Language models are few-shot learners.

Buechel, S., Rücker, S. and Hahn, U., 2020. Learning and evaluating emotion lexicons for 91 languages, *In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1202–1217.

Callaway, C. B. and Lester, J. C., 2002. Narrative prose generation, *Artificial Intelligence* **139**(2), 213–252.

Cao, Y., Shui, R., Pan, L., Kan, M.-Y., Liu, Z. and Chua, T.-S., 2020. Expertise style transfer: A new task towards better communication between experts and laymen, *In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1061–1071.

Cavazza, M. and Charles, F., 2005. Dialogue generation in character-based interactive storytelling, *In: Proceedings of the First AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, AAAI Press, pp. 21–26.

Cavazza, M., Charles, F. and Mead, S. J., 2002. Character-based interactive storytelling, *IEEE Intelligent systems* **17**(4), 17–24.

Cavazza, M., Pizzi, D., Charles, F., Vogt, T. and André, E., 2009. Emotional input for character-based interactive storytelling, *In: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, International Foundation for Autonomous Agents and Multiagent Systems, pp. 313–320.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C. et al., 2018. Universal sentence encoder, *arXiv preprint arXiv:1803.11175* .

Chambers, N. and Jurafsky, D., 2008. Unsupervised learning of narrative event chains, *In: Proceedings of ACL-08: HLT*, pp. 789–797.

Chambers, N. and Jurafsky, D., 2009. Unsupervised learning of narrative schemas and their participants, *In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 602–610.

Chatman, S., 1978. *Story and Discourse: Narrative Structure in Fiction and Film*, Cornell University Press.

Chen, H., Liu, X., Yin, D. and Tang, J., 2017. A survey on dialogue systems: Recent advances and new frontiers, *ACM SIGKDD Explorations Newsletter* **19**(2), 25–35.

Cheong, Y.-G. and Young, R. M., 2015. Suspenser: A story generation system for suspense, *IEEE Transactions on Computational Intelligence and AI in Games* **7**(1), 39–52.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y., 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation, *In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734.

Chung, J., Gulcehre, C., Cho, K. and Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling, *arXiv preprint arXiv:1412.3555* .

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S. and Smith, N. A., 2021. All that's 'human'is not gold: Evaluating human evaluation of generated text, *In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7282–7296.

Colombo, P., Witon, W., Modi, A., Kennedy, J. and Kapadia, M., 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)*, pp. 3734–3743.

Coltheart, M., 1981. The mrc psycholinguistic database, *The Quarterly Journal of Experimental Psychology Section A* **33**(4), 497–505.

Concepción, E., Gervás, P. and Méndez, G., 2018. Ines: A reconstruction of the charade storytelling system using the afanasyev framework, *In: Ninth International Conference on Computational Creativity, ICCC*.

Cope, E. M. and Sandys, J. E., 2010. *Aristotle: Rhetoric*, Vol. 2, Cambridge University Press.

Crothers, R., 2016. The construction of a play, *In: The Art of Playwriting*, University of Pennsylvania Press, pp. 115–134.

Damiano, R. and Lieto, A., 2013. Ontological representations of narratives: a case study on stories and actions, *In: 2013 Workshop on Computational Models of Narrative*, Schloss Dagstuhl Leibniz-Zentrum für Informatik, pp. 76–93.

Danescu-Niculescu-Mizil, C. and Lee, L., 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs., *In: Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Dehejia, V., 1990. On modes of visual narration in early buddhist art, *The art bulletin* **72**(3), 374–392.

Devlin, J., Chang, M.-W., Lee, K. and 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.

DiMarco, C. and Hirst, G., 1993. A computational theory of goal-directed style in syntax, *Computational Linguistics* **19**(3), 451–500.

Dong, L., Huang, S., Wei, F., Lapata, M., Zhou, M. and Xu, K., 2017. Learning to generate product reviews from attributes, *In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Vol. 1, pp. 623–632.

Dose, S., 2013. Flipping the script: A corpus of american television series (cats) for corpus-based language learning and teaching, *Corpus linguistics and variation in English: Focus on non-native Englishes* .

Dyer, R., 1989. *The Stars*, BFI Education.

Ekman, P., Levenson, R. W. and Friesen, W. V., 1983. Autonomic nervous system activity distinguishes among emotions, *science* **221**(4616), 1208–1210.

Elhadad, M., 1993. *Using argumentation to control lexical choice: a functional unification implementation*, Columbia University.

Elson, D. K. and McKeown, K. R., 2009. A tool for deep semantic encoding of narrative texts, *In: Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, Association for Computational Linguistics, pp. 9–12.

Eskelinen, M., 2001. The gaming situation, *Game studies* **1**(1), 68.

Exus Software Ltd., 2016-2019. *Emotive Virtual cultural Experiences through personalized storytelling*, European Union's Horizon 2020 research and innovation programme under grant agreement No 727188.
URL: *https://emotiveproject.eu/*

Fan, A., Lewis, M. and Dauphin, Y., 2018. Hierarchical neural story generation, *In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898.

Fan, A., Lewis, M. and Dauphin, Y., 2019. Strategies for structuring story generation, *In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2650–2660.

Ficler, J. and Goldberg, Y., 2017. Controlling linguistic style aspects in neural language generation, *In: Proceedings of the Workshop on Stylistic Variation*, pp. 94–104.

Forchini, P., 2012. *Movie language revisited. Evidence from multi-dimensional analysis and corpora*, Peter Lang.

Gašić, M., Breslin, C., Henderson, M., Kim, D., Szummer, M., Thomson, B., Tsiakoulis, P. and Young, S., 2013. On-line policy optimisation of bayesian spoken dialogue systems via human interaction, *In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp. 8367–8371.

Gatt, A. and Krahmer, E., 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation, *Journal of Artificial Intelligence Research* **61**, 65–170.

Genette, G., 1983. *Narrative discourse: An essay in method*, Vol. 3, Cornell University Press.

Gervás, P., 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, pp. 23–30.

Gervás, P., Díaz-Agudo, B., Peinado, F. and Hervás, R., 2004. Story plot generation based on cbr, *In: International Conference on Innovative Techniques and Applications of Artificial Intelligence*, Springer, pp. 33–46.

Ghosh, S., Chollet, M., Laksana, E., Morency, L.-P. and Scherer, S., 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)*, pp. 634–642.

Gillick, D., Brunk, C., Vinyals, O. and Subramanya, A., 2016. Multilingual language processing from bytes, *In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1296–1306.

Goldberg, L. R., 1993. The structure of phenotypic personality traits., *American psychologist* **48**(1), 26.

Greimas, A., 1966. *Sémantique structurale: recherche de méthode*, Collection "Langue et Langage", Larousse.

Guan, J., Huang, F., Zhao, Z., Zhu, X. and Huang, M., 2020. A knowledge-enhanced pretraining model for commonsense story generation, *Transactions of the Association for Computational Linguistics* **8**, 93–108.

Hargood, C., 2011. *Semiotic term expansion as the basis for thematic models in narrative systems*, PhD thesis, University of Southampton.

Herman, D., Manfred, J. and Marie-Laure, R., 2010. *Routledge encyclopedia of narrative theory*, Routledge.

Herman, V., 1998. *Dramatic discourse: Dialogue as interaction in plays*, Psychology Press.

Herzig, J., Shmueli-Scheuer, M., Sandbank, T. and Konopnicki, D., 2017. Neural response generation for customer service based on personality traits, *In: Proceedings of the 10th International Conference on Natural Language Generation*, pp. 252–256.

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory, *Neural computation* **9**(8), 1735–1780.

Hovy, D., Bianchi, F. and Fornaciari, T., 2020. "you sound just like your father" commercial machine translation systems include stylistic biases, *In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1686–1690.

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R. and Xing, E. P., 2017. Toward controlled generation of text, *In: Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, pp. 1587–1596.

Huang, C., Zaiane, O. R., Trabelsi, A. and Dziri, N., 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)*, pp. 49–54.

Huang, X., Qi, J., Sun, Y. and Zhang, R., 2020. Mala: Cross-domain dialogue generation with action learning, *In: Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 7977–7984.

Jain, P., Agrawal, P., Mishra, A., Sukhwani, M., Laha, A. and Sankaranarayanan, K., 2017. Story generation from sequence of independent short descriptions, *In: SIGKDD 2017 Workshop on Machine Learning for Creativity*.

Jhala, A., 2008. Exploiting structure and conventions of movie scripts for information retrieval and text mining, *In: Joint International Conference on Interactive Digital Storytelling*, Springer, pp. 210–213.

Jhala, A. and Young, R. M., 2010. Cinematic visual discourse: Representation, generation, and evaluation, *IEEE Transactions on computational intelligence and AI in games* **2**(2), 69–81.

John, O. P., Angleitner, A. and Ostendorf, F., 1988. The lexical approach to personality: A historical review of trait taxonomic research, *European journal of Personality* **2**(3), 171–203.

John, O. and Srivastava, S., 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives, *In: Handbook of personality: Theory and research*, Guilford Press, p. 102–138.

Johnson, J., Alahi, A. and Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution, *In: European conference on computer vision*, Springer, pp. 694–711.

Joshi, A. K. and Schabes, Y., 1997. Tree-adjoining grammars, *In: Handbook of formal languages*, Springer, pp. 69–123.

Jung, C., 2016. *Psychological types*, Routledge.

Jurafsky, D., 2000. *Speech & language processing*, Pearson Education India.

Juul, J. and Ping-ping, G., 2010. Games telling stories?——a brief note on games and narratives, *Studies in Culture & Art* .

Keh, S. S., Cheng, I. et al., 2019. Myers-briggs personality classification and personality-specific language generation using pre-trained language models, *arXiv preprint arXiv:1907.06333* .

King, S., 2000. *On writing: A memoir of the craft*, Simon and Schuster.

Klinger, R., De Clercq, O., Mohammad, S. and Balahur, A., 2018. Iest: Wassa-2018 implicit emotions shared task, *In: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 31–42.

Kozloff, S., 2000. *Overhearing film dialogue*, Univ of California Press.

Kybartas, B. and Bidarra, R., 2016. A survey on story generation techniques for authoring computational narratives, *IEEE Transactions on Computational Intelligence and AI in Games* **9**(3), 239–253.

Lavoie, B. and Rainbow, O., 1997. A fast and portable realizer for text generation systems, *In: Fifth Conference on Applied Natural Language Processing*.

Lebowitz, M., 1985. Story-telling as planning and learning, *Poetics* **14**(6), 483–502.

Lemon, O., 2008. Adaptive natural language generation in dialogue using reinforcement learning, *In: Proceedings of the 12th SEMdial Workshop on the Semantics and Pragmatics of Dialogues*, pp. 149–156.

Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J. and Dolan, B., 2016. A persona-based neural conversation model, *In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 994–1003.

Lin, G., 2016. *Character Modeling through Dialogue for Expressive Natural Language Generation*, PhD thesis, UC Santa Cruz.

Lison, P. and Tiedemann, J., 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles, *In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 923–929.

Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L. and Pineau, J., 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*, pp. 2122–2132.

Liu, D., Li, J., Yu, M.-H., Huang, Z., Liu, G., Zhao, D. and Yan, R., 2020. A character-centric neural model for automated story generation, *In: Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 1725–1732.

Luong, M.-T., Pham, H. and Manning, C. D., 2015. Effective approaches to attention-based neural machine translation, *In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421.

Mairesse, F. and Walker, M., 2007. Personage: Personality generation for dialogue, *In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 496–503.

Mairesse, F. and Walker, M. A., 2010. Towards personality-based user adaptation: psychologically informed stylistic language generation, *User Modeling and User-Adapted Interaction* **20**(3), 227–278.

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

Mairesse, F., Walker, M. A., Mehl, M. R. and Moore, R. K., 2007. Using linguistic cues for the automatic recognition of personality in conversation and text, *Journal of artificial intelligence research* **30**, 457–500.

Mao, H. H., Majumder, B. P., McAuley, J. and Cottrell, G., 2019. Improving neural story generation by targeted common sense grounding, *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. 5988–5993.

Martin, L. J., Ammanabrolu, P., Wang, X., Hancock, W., Singh, S., Harrison, B. and Riedl, M. O., 2018. Event representations for automated story generation with deep neural nets, *In: Thirty-Second AAAI Conference on Artificial Intelligence*.

Mateas, M., 2007. The authoring bottleneck in creating ai-based interactive stories, *In: Proceedings of the AAAI 2007 Fall Symposium on Intelligent Narrative Technologies*, Vol. 1, pp. 10–2.

Mateas, M. and Stern, A., 2003. Façade: An experiment in building a fully-realized interactive drama, *In: Game developers conference*, Vol. 2, pp. 4–8.

Matthews, J., Charles, F., Porteous, J. and Mendes, A., 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, pp. 782–790.

Mazare, P.-E., Humeau, S., Raison, M. and Bordes, A., 2018. Training millions of personalized dialogue agents, *In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2775–2779.

McDonald, D. D. and Pustejovsky, J., 1985. A computational theory of prose style for natural language generation, *In: Second Conference of the European Chapter of the Association for Computational Linguistics*.

McKee, R., 1997. *Story: Substance, structure, style, and the principles of screenwriting*, New York: HarperCollins.

McKee, R., 2016. *Dialogue: The art of verbal action for page, stage, and screen*, Hachette UK.

Mehl, M., Gosling, S. and Pennebaker, J., 2006. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life, *Journal of Personality and Social Psychology* **90**(5), 862–877.

Meister, J. C. and Schernus, W., 2011. *Time: from concept to narrative construct: a reader*, Vol. 29, Walter de Gruyter.

Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* .

Mikolov, T., Kombrink, S., Burget, L., Černockỳ, J. and Khudanpur, S., 2011. Extensions of recurrent neural network language model, *In: 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, pp. 5528–5531.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality, *In: Advances in neural information processing systems*, pp. 3111–3119.

Nau, D. S., Smith, S. J., Erol, K. et al., 1998. Control strategies in htn planning: Theory versus practice, *In: AAAI/IAAI*, pp. 1127–1133.

Nelmes, J., 2011. *Analysing the screenplay*, Routledge London.

Nio, L., Sakti, S., Neubig, G., Toda, T. and Nakamura, S., 2014. Conversation dialog corpora from television and movie scripts, *In: 2014 17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA)*, IEEE, pp. 1–4.

Nottingham University, 2015-2017. *Artificial Retrieval of Information Assistants - Virtual Agents with Linguistic Understanding, Social skills, and Personalised Aspects - ARIAs*, European Union's Horizon 2020 research and innovation programme under grant agreement No 645378 Aria-Valuspa.
URL: *http://aria-agent.eu/*

Novikova, J., Dušek, O., Curry, A. C. and Rieser, V., 2017. Why we need new evaluation metrics for nlg, *In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2241–2252.

Oberlander, J. and Nowson, S., 2006. Whose thumb is it anyway? classifying author personality from weblog text, *In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pp. 627–634.

Oraby, S., Reed, L., Tandon, S., Sharath, T., Lukin, S. and Walker, M., 2018. Controlling personality-based stylistic variation with neural natural language generators, *In: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 180–190.

Osgood, C. E., Suci, G. J. and Tannenbaum, P. H., 1957. *The measurement of meaning*, number 47, University of Illinois press.

Papalampidi, P., Keller, F., Frermann, L. and Lapata, M., 2020. Screenplay summarization using latent narrative structure, *In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1920–1933.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J., 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, pp. 311–318.

Peinado, F., Cavazza, M. and Pizzi, D., 2008. Revisiting character-based affective storytelling under a narrative bdi framework, *In: Joint International Conference on Interactive Digital Storytelling*, Springer, pp. 83–88.

Pemberton, L., 1989. A modular approach to story generation, *In: Fourth Conference of the European Chapter of the Association for Computational Linguistics*.

Pennebaker, J. and King, L., 1999. Linguistic styles: language use as an individual difference., *Journal of Personality and Social Psychology* **77**(6), 1296–1312.

Pennebaker, J. W., 2001. Linguistic inquiry and word count: Liwc 2001, *Mahway: Lawrence Erlbaum Associates* .

Pennington, J., Socher, R. and Manning, C., 2014. Glove: Global vectors for word representation, *In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

PÉrez, R. P. Ý. and Sharples, M., 2001. Mexica: A computer model of a cognitive account of creative writing, *Journal of Experimental & Theoretical Artificial Intelligence* **13**(2), 119–139.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L., 2018. Deep contextualized word representations, *In: Proceedings of NAACL-HLT*, pp. 2227–2237.

Petersen, R., 2010. *Comics, Manga, and Graphic Novels: A History of Graphic Narratives: A History of Graphic Narratives*, ABC-CLIO.

Pichotta, K. and Mooney, R., 2014. Statistical script learning with multi-argument events, *In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 220–229.

Pichotta, K. and Mooney, R., 2016. Learning statistical scripts with lstm recurrent neural networks, *In: Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

Pittenger, D. J., 1993. Measuring the mbti... and coming up short, *Journal of Career Planning and Employment* **54**(1), 48–52.

Pizzi, D., Charles, F., Lugrin, J.-L. and Cavazza, M., 2007. Interactive storytelling with literary feelings, *In: International Conference on Affective Computing and Intelligent Interaction*, Springer, pp. 630–641.

Porteous, J., Cavazza, M. and Charles, F., 2010*a*. Applying planning to interactive storytelling: Narrative control using state constraints, *ACM Transactions on Intelligent Systems and Technology (TIST)* **1**(2), 1–21.

Porteous, J., Cavazza, M. and Charles, F., 2010*b*. 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, pp. 1297–1304.

Porteous, J., Charles, F. and Cavazza, M., 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, pp. 595–602.

Propp, V. I., 1968. *Morphology of the Folktale*, Vol. 9, University of Texas Press.

Qian, Q., Huang, M., Zhao, H., Xu, J. and Zhu, X., 2018. Assigning personality/profile to a chatting machine for coherent conversation generation, *In: Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 4279–4285.

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I., 2018. Improving language understanding by generative pre-training, *OpenAI* .

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multitask learners, *OpenAI* .

Riedl, M. O. and Young, R. M., 2003. Character-focused narrative generation for execution in virtual worlds, *In: International Conference on Virtual Storytelling*, Springer, pp. 47–56.

Riedl, M. O. and Young, R. M., 2006. From linear story generation to branching story graphs, *IEEE Computer Graphics and Applications* **26**(3), 23–31.

Riedl, M. O. and Young, R. M., 2010. Narrative planning: Balancing plot and character, *Journal of Artificial Intelligence Research* **39**, 217–268.

Riley, C., 2009. *The Hollywood Standard: The Complete and Authoritative Guide to Script Format and Style*, Hollywood Standard: the Complete and Authoritative Guide To Series, Michael Wiese Productions.

Rimmon-Kenan, S., 1983. *Narrative Fiction: Contemporary Poetics*, New accents, Routledge.

Rishes, E., Lukin, S. M., Elson, D. K. and Walker, M. A., 2013. Generating different story tellings from semantic representations of narrative, *In: International Conference on Interactive Digital Storytelling*, Springer, pp. 192–204.

Russell, J. A., 1980. A circumplex model of affect., *Journal of personality and social psychology* **39**(6), 1161.

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

Ryan, M.-L., 2006. *Avatars of Story*, Vol. 17, University of Minnesota Press.

Ryan, M.-L., Ruppert, J. and Bernet, J. W., 2004. *Narrative across media: The languages of storytelling*, U of Nebraska Press.

See, A., Pappu, A., Saxena, R., Yerukola, A. and Manning, C. D., 2019. Do massively pretrained language models make better storytellers?, *In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 843–861.

Sennrich, R., Haddow, B. and Birch, A., 2016. Neural machine translation of rare words with subword units, *In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725.

Serban, I. V., Lowe, R., Henderson, P., Charlin, L. and Pineau, J., 2018. A survey of available corpora for building data-driven dialogue systems: The journal version, *Dialogue & Discourse* **9**(1), 1–49.

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A. and Pineau, J., 2016. Building end-to-end dialogue systems using generative hierarchical neural network models, *In: Thirtieth AAAI Conference on Artificial Intelligence*.

Skorupski, J., Jayapalan, L., Marquez, S. and Mateas, M., 2007. Wide ruled: A friendly interface to author-goal based story generation, *In: International Conference on Virtual Storytelling*, Springer, pp. 26–37.

Skorupski, J. and Mateas, M., 2010. Novice-friendly authoring of plan-based interactive storyboards, *In: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 6, pp. 174–179.

Small, J. P., 1999. Time in space: Narrative in classical art, *The Art Bulletin* **81**(4), 562–575.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J. and Dolan, W. B., 2015. A neural network approach to context-sensitive generation of conversational responses, *In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 196–205.

Steele, A. et al., 2006. *Writing Movies: The Practical Guide to Creating Stellar Screenplays*, Bloomsbury Publishing USA.

Sutskever, I., Vinyals, O. and Le, Q. V., 2014. Sequence to sequence learning with neural networks, *In: Advances in neural information processing systems*, pp. 3104–3112.

Swanson, R. and Gordon, A. S., 2012. Say anything: Using textual case-based reasoning to enable open-domain interactive storytelling, *ACM Transactions on Interactive Intelligent Systems (TiiS)* **2**(3), 1–35.

Tambwekar, P., Dhuliawala, M., Martin, L. J., Mehta, A., Harrison, B. and Riedl, M. O., 2019. Controllable neural story plot generation via reward shaping, *In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, International Joint Conferences on Artificial Intelligence Organization, pp. 5982–5988.

The British Museum, 2019. *A history of storytelling through pictures* [online], Available from: https://www.britishmuseum.org/blog/history-storytelling-through-pictures [Accessed 10 December 2020].

The Washington Post, 2019. *The oldest story ever told is painted on this cave wall, archaeologists report* [online], Available from: https://www.washingtonpost.com/science/2019/12/11/oldest-story-ever-told-is-painted-this-cave-wall-archaeologists-report/ [Accessed 10 December 2020].

Toolan, M. J., 2013. *Narrative: A critical linguistic introduction*, Routledge.

van Stegeren, J. and Myśliwiec, J., 2021. Fine-tuning gpt-2 on annotated rpg quests for npc dialogue generation, *In: The 16th International Conference on the Foundations of Digital Games (FDG) 2021*, pp. 1–8.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need, *In: Advances in Neural Information Processing Systems*, pp. 5998–6008.

Vinyals, O. and Le, Q., 2015. A neural conversational model, *arXiv preprint arXiv:1506.05869* .

Walker, M., Lin, G. and Sawyer, J., 2012. An annotated corpus of film dialogue for learning and characterizing character style, *In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 1373–1378.

Wen, T.-H., Gasic, M., Mrkšić, N., Su, P.-H., Vandyke, D. and Young, S., 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems, *In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1711–1721.

Wen, T.-H., Vandyke, D., Mrkšić, N., Gasic, M., Barahona, L. M. R., Su, P.-H., Ultes, S. and Young, S., 2017. A network-based end-to-end trainable task-oriented dialogue system, *In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 438–449.

Wikipedia, 2021. *List of highest-grossing films* [online], Available from: https://en.wikipedia.org/wiki/List_of_highest-grossing_films [Accessed 17 August 2021].

Williams, J. D., Henderson, M., Raux, A., Thomson, B., Black, A. and Ramachandran, D., 2014. The dialog state tracking challenge series, *AI Magazine* **35**(4), 121–124.

Winer, D. and Young, R., 2017. Automated screenplay annotation for extracting storytelling knowledge, *In: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 13, pp. 273–280.

Xu, W., 2020. Stylistic dialogue generation in narratives, *In: 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020)*, Vol. 29.

Xu, W., Charles, F., Hargood, C., Tian, F. and Tang, W., 2020. Influence of personality-based features for dialogue generation in computational narratives, *In: ECAI 2020 - 24th European Conference on Artificial Intelligence*, Vol. 325 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, pp. 2945–2946.

Xu, W., Hargood, C., Tang, W. and Charles, F., 2018. Towards generating stylistic dialogues for narratives using data-driven approaches, *In: International Conference on Interactive Digital Storytelling*, Springer, pp. 462–472.

Yan, R., Zhao, D. and E, W., 2017. Joint learning of response ranking and next utterance suggestion in human-computer conversation system, *In: Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, pp. 685–694.

Yang, Z., Wu, W., Xu, C., Liang, X., Bai, J., Wang, L., Wang, W. and Li, Z., 2020. Styledgpt: Stylized response generation with pre-trained language models, *In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 1548–1559.

Yin, W., Schütze, H., Xiang, B. and Zhou, B., 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs, *Transactions of the Association for Computational Linguistics* **4**, 259–272.

Young, R. M., 2000. Creating interactive narrative structures: The potential for ai approaches, *Psychology* **13**, 1–26.

Young, R. M., Ware, S. G., Cassell, B. A. and Robertson, J., 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), 41–64.

Young, S., Gašić, M., Thomson, B. and Williams, J. D., 2013. Pomdp-based statistical spoken dialog systems: A review, *Proceedings of the IEEE* **101**(5), 1160–1179.

Youyou, W., Kosinski, M. and Stillwell, D., 2015. Computer-based personality judgments are more accurate than those made by humans, Vol. 112, National Acad Sciences, pp. 1036–1040.

Yu, L., Bansal, M. and Berg, T., 2017. Hierarchically-attentive rnn for album summarization and storytelling, *In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 966–971.

Zhang, A., Lipton, Z. C., Li, M. and Smola, A. J., 2021. Dive into deep learning, *arXiv preprint arXiv:2106.11342* .

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J. and Dolan, W. B., 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*, pp. 270–278.

Zhang, Y., Tsipidi, E., Schriber, S., Kapadia, M., Gross, M. and Modi, A., 2019. Generating animations from screenplays, *In: Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pp. 292–307.

Zheng, Y., Zhang, R., Huang, M. and Mao, X., 2020. A pre-training based personalized dialogue generation model with persona-sparse data, *In: Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 9693–9700.

Appendix A

Publications and Awards

Publications

1. Towards Generating Stylistic Dialogues for Narratives using Data-Driven Approaches. *International Conference on Interactive Digital Storytelling 2018* (Xu et al., 2018)
2. Influence of Personality-based Features for Dialogue Generation in Computational Narratives. *European Conference on Artificial Intelligence - ECAI 2020* (Xu et al., 2020)
3. Stylistic Dialogue Generation in Narratives. *Doctoral Consortium of European Conference on Artificial Intelligence - ECAI 2020* (Xu, 2020)

Awards

1. 2020.09 Won the best presentation award in Doctoral Consortium of European Conference on Artificial Intelligence 2020, Santiago de Compostela, Spain
2. 2018.12 Nominated for the best short paper in International Conference on Interactive Digital Storytelling 2018, Dublin, Ireland
3. 2018.05 Won the poster presentation winners 2nd prize, Faculty of Science and Technology PGR conference, Bournemouth University, UK

Resource

The GitHub repository for code and dataset used in this thesis have been linked to BORDaR, BU's research data repository.

Appendix B

Dialogues used in User Studies

B.1 Dialogues for User Study i(Chapter 6)

Personality Code Description

Genre

- **A:** Action
- **D:** Drama
- **R:** Romance
- **T:** Thriller

Personality

- **N:** No Personality
- **1:** (extrovert, emotionally stable, agreeable)
- **2:** (extrovert, emotionally stable, disagreeable)
- **3:** (extrovert, neurotic, agreeable)
- **4:** (extrovert, neurotic, disagreeable)
- **5:** (introvert, emotionally stable, agreeable)
- **6:** (introvert, emotionally stable, disagreeable)
- **7:** (introvert, neurotic, agreeable)
- **8:** (introvert, neurotic, disagreeable)
- **S:** Source Script

Action

Table B.1 Action dialogues for evaluation in User Study i.

Personality Code	Dialogue for Evaluation
AN	**A:** How could you get the carpet wrong?! **B:** I can't get the floor wrong! Get the kids out of here! **A:** I can't get the kids out of here! **B:** You need a nurse! I'm here to see you – tell your J.D. to help me! No one said one of this!
A1	**A:** How could you get the carpet wrong?! **B:** I got the roof wrong. **A:** Easier if you say yes. **B:** Oh, sure you do, Easier if you say yes. Let's echo the news.
A2	**A:** How could you get the carpet wrong?! **B:** I thought we were cool. **A:** I don't think we're cool. **B:** But listen, honey, we're cool! We're cool! You just need a little eggo.
A3	**A:** How could you get the carpet wrong?! **B:** It's got to be here somewhere. **A:** It's got to be here somewhere. **B:** We're not safe here.
A4	**A:** How could you get the carpet wrong?! **B:** I had to go to the bathroom. **A:** Do you sleep? **B:** No. I don't sleep.
A5	**A:** How could you get the carpet wrong?! **B:** This is the roof. **A:** This is the roof. It's where you were being held up. **B:** It's where you were being held up.
A6	**A:** How could you get the carpet wrong?! **B:** Fool. The outermost baffle is a hole in the hull. **A:** I'm fixin' that baffle. **B:** How long to fix?
A7	**A:** How could you get the carpet wrong?! **B:** It's got to be far too hot. **A:** You guys need to be hot. **B:** No, no. Guys need to be hot.
A8	**A:** How could you get the carpet wrong?! **B:** It's got to be. **A:** I guess we better get out of here. **B:** We better get out of here. She laughs.
AS	**A:** How could you get the carpet wrong?!

Continuation of Table B.1

Personality Code	Dialogue for Evaluation
	B: It wasn't my fault! **A:** You're the architect- **B:** I didn't know he was going to rub his damn cheek on it!

Drama

Table B.2 Drama dialogues for evaluation in User Study i.

Personality Code	Dialogue for Evaluation
DN	**A:** I can't stay here all night chasing your pain away. **B:** We can't stay here all night chasing our pain away. There's no time to dote it. **A:** That is the motto. The motto. **B:** Make it. The motto. The motto.
D1	**A:** I can't stay here all night chasing your pain away. **B:** I want to leave you feeling better. **A:** I'm not sure I want you feeling better. **B:** You feeling better, are you?
D2	**A:** I can't stay here all night chasing your pain away. **B:** I had to leave this place. **A:** If it was only for a dog, I'd be happy to take it home. **B:** Honey, if it was only for a dog, I'd be happy to take it home.
D3	**A:** I can't stay here all night chasing your pain away. **B:** I know. But I need to find a way to get my mother to a loving home. **A:** And that's why I'm interested in that dog. **B:** And that's why I need a love that'll extinguish any heat. And that's why my mother hates dogs. And I'll bring her down so that you can leave.
D4	**A:** I can't stay here all night chasing your pain away. **B:** We need you to take my pain and think for myself. We need you to take my pain and think for ourselves. **A:** I don't know if I'm going to make it up. It's not easy. **B:** You're not going to make it up to me. I just don't know it's that easy.
D5	**A:** I can't stay here all night chasing your pain away. **B:** I know. And I know you do. And I'll take care of it. **A:** I'm going to cut you loose. **B:** Shut up and see how I cut you loose.
D6	**A:** I can't stay here all night chasing your pain away. **B:** If I try again, I'll kill you. It'll all be a long time. **A:** You can't leave the house all the time. You can still come back there with me. **B:** You're not alone, Farmer. You still can't let me kill you.

Continuation of Table B.2

Personality Code	Dialogue for Evaluation
D7	**A:** I can't stay here all night chasing your pain away. **B:** I thought you were going to leave me here chasing you. **A:** I thought you were going to leave me here with your cat. **B:** cat. I thought you dumped it.
D8	**A:** I can't stay here all night chasing your pain away. **B:** You're frightened. We will take you to the beach. **A:** I don't want to be alone on a rock. **B:** What is it like? No way. The way you look, the way you look, is by taking down the rocks.
DS	**A:** I can't stay here all night chasing your pain away. **B:** I know... But you do look so silly. **A:** Where's the towel? **B:** On the chiffonier. Here.

Romance

Table B.3 Romance dialogues for evaluation in User Study i.

Personality Code	Dialogue for Evaluation
RN	**A:** When will he be back? **B:** I don't know. He'll be back. **A:** All right. There's nothing to see. **B:** All right. He starts off into the dark.
R1	**A:** When will he be back? **B:** I believe he will be back with me. **A:** I believe he will be back with me. **B:** If things are so simple, simple, then he will be back with me.
R2	**A:** When will he be back? **B:** He will be back? **A:** He will be back with me **B:** I don't understand his head very well. He will be back with me.
R3	**A:** When will he be back? **B:** He will be back with the courtside. **A:** I wish he'd be back with the courtside. **B:** If there is anything not to be round here, he will be back with the courtside.
R4	**A:** When will he be back? **B:** I don't want to be in the sky without him. **A:** I want to be in the sky! **B:** Come on... low-percentage man... I don't want to be in the sky!

Continuation of Table B.3

Personality Code	Dialogue for Evaluation
R5	**A:** When will he be back? **B:** I believe him. **A:** I bet he won't be back. **B:** I bet he'll be back soon.
R6	**A:** When will he be back? **B:** I think he's going to be back soon. **A:** I think he's going to be back for sometime. Stay on. **B:** I think he's going to be back for sometime. Stay on. Don't be silly about him. Stay on him.
R7	**A:** When will he be back? **B:** He will be back with the fishes. **A:** He will be back with the fishes. **B:** I think he will be back with the fishes.
R8	**A:** When will he be back? **B:** He will be out in the open air with the reener. **A:** I want him to stay in the open air with the reener. **B:** He will be in the open air with the reener.
RS	**A:** When will he be back? **B:** Not tonight no more. He ain't coming. Uh, he went home. **A:** Does he always leave so early? **B:** Oh, he never... well...

Thriller

Table B.4 Thriller dialogues for evaluation in User Study i.

Personality Code	Dialogue for Evaluation
TN	**A:** We have reason to believe she came this way... might have stopped in this area...Did she stop here? **B:** Let me take her. We'll find a place where no one went and found you. **A:** She liked me. **B:** Well, like I said, like, she liked you too.
T1	**A:** We have reason to believe she came this way... might have stopped in this area...Did she stop here? **B:** What would she have to do with an innocent man? **A:** What do you mean? **B:** What do you mean?
T2	**A:** We have reason to believe she came this way... might have stopped in this area...Did she stop here? **B:** Did she stop the car to get us over the border?

Continuation of Table B.4

Personality Code	Dialogue for Evaluation
	A: That's her license number. **B:** That's right. Samantha's license number is not yet sis her name!
T3	**A: A:** We have reason to believe she came this way... might have stopped in this area...Did she stop here? **B:** There have been reports of a boat... and it's been almost three years since she disappeared. **A:** If everything were in order please tell everyone that you stopped her... **B:** So tell everyone that you stopped her...
T4	**A:** We have reason to believe she came this way... might have stopped in this area...Did she stop here? **B:** She had no choice or choice... **A:** That's not what I thought. **B:** across the country. This way.
T5	**A:** We have reason to believe she came this way... might have stopped in this area...Did she stop here? **B:** Did she stop calling the cops? **A:** What kind of cop? **B:** What kind of cop?
T6	**A:** We have reason to believe she came this way... might have stopped in this area...Did she stop here? **B:** She couldn't stop me. **A:** I want to believe that she did this. **B:** She did this, didn't she?
T7	**A:** We have reason to believe she came this way... might have stopped in this area...Did she stop here? **B:** She wouldn't. Why would she stop there? **A:** What happened last night? **B:** What had happened last night?
T8	**A:** We have reason to believe she came this way... might have stopped in this area...Did she stop here? **B:** Did she stop calling the cops? **A:** Why haven't you stopped calling the cops? **B:** Just stopped calling and told them the same number - y'know. Remember?
TS	**A:** We have reason to believe she came this way... might have stopped in this area...Did she stop here? **B:** No. No one has stopped here in weeks... **A:** Mind looking at the picture before committing yourself? **B:** Committing myself to what? You sure talk like a Policeman.

B.2 Dialogues for User Study ii(Chapter 6)

Action

Table B.5 Action dialogues for evaluation in User Study ii.

Personality Code	Dialogue for Evaluation
AN∗[1]	**A:** I've never seen a design like this. what are those scales on her hull? **B:** They're to keep the heat off the engines when she's sailing. **A:** That's not a lot of cooling. **B:** It's a lot of cooling for a ship that size.
A1	**A:** I've never seen a design like this. what are those scales on her hull? **B:** She was the first commissioned officer to use them. They were designed by her. **A:** Captain, we must return to base. We can 't leave her here. **B:** You're right. She's been a valuable asset.
A2	**A:** I've never seen a design like this. what are those scales on her hull? **B:** It's a type of armor. She's a battleship with a high-tech weapon. **A:** You said she was armed. **B:** She was. But she wasn't a battleship. It's not my ship, but a battle carrier, with weapons like these.
A3	**A:** I've never seen a design like this. what are those scales on her hull? **B:** She's the hull of a ship. It's a hull-on-sheeter design. **A:** Yes, of course, I know. But it 'd be impossible to make a ship out of it. **B:** It's not. It's a theory.
A4∗	**A:** I've never seen a design like this. what are those scales on her hull? **B:** My guess are those composite sails you're seeing. **A:** I thought they looked like she was going overboard. **B:** She was. She's got to keep her weight down so she can breathe... and there's a big enough gap for her to come up through. That was a close one, you know.
A5	**A:** I've never seen a design like this. what are those scales on her hull? **B:** These are the aero capture scales, used on ships that are already built. **A:** This looks like a new ship then. **B:** No doubt about it. And it still has its ways. I've been working with them. I've seen this ship before and I haven 't seen these scales.
A6	**A:** I've never seen a design like this. what are those scales on her hull? **B:** The hull is aluminium. **A:** So we're basically gon na be standing on her? **B:** More or less.
A7∗	**A:** I've never seen a design like this. what are those scales on her hull? **B:** They're from the ship, the ship's hull is hollow. **A:** She's got to be under some kind of attack.

[1]The symbol "∗" denotes the selected personalities for **User Study ii**.

Continuation of Table B.1

Personality Code	Dialogue for Evaluation
	B: I suppose she'd better be.
A8∗	**A:** I've never seen a design like this. what are those scales on her hull? **B:** I'm not sure... **A:** I can feel them! she's cold. **B:** She's a freighter... she doesn't have much space left!
AS∗	**A:** I've never seen a design like this. what are those scales on her hull? **B:** How could Earth launch a ship this massive and neither of us know about it? ... It just doesn't make sense. **A:** These scanners are damaged, I'm getting inconsistent life signs but they may be sensor ghosts. **B:** Maybe her computer's still up. I'm trying standard remote access codes.

Drama

Table B.6 Drama dialogues for evaluation in User Study ii.

Personality Code	Dialogue for Evaluation
DN∗	**A:** You come into the Five Points and you stole from me. **B:** I'm not from the Five Point neighborhood. **A:** So you stole from me. **B:** I didn 't steal. I came back with cash on my hand.
D1	**A:** You come into the Five Points and you stole from me. **B:** I didn't steal anything. I found it, you know. **A:** You stole it. You found it. **B:** No, I didn't. I'm innocent. I'm not the one who did this to you. I don 't know who did this to you. I can 't tell you who it was. I just know I've done it.
D2∗	**A:** You come into the Five Points and you stole from me. **B:** I didn't say I stole anything. **A:** You did in the sense of lying to the police about an account of a robbery. **B:** I didn't say that. I said I don't really care about the robbery. I care about the money. I don't care about the hostages. That was a mistake.
D3∗	**A:** You come into the Five Points and you stole from me. **B:** No, I didn't. You found it in an old bag. You made it yourself. And then you took our little boy – we couldn't let him go to sleep. **A:** You had to do it to get your money back, didn't you? **B:** No I didn't, I sold it to the police – the guys were looking for it – and they found it.
D4∗	**A:** You come into the Five Points and you stole from me. **B:** I stole from everyone. Except you.

Continuation of Table B.2

Personality Code	Dialogue for Evaluation
	A: You stole from me – and you're on three, and a half years in the community. You're a murderer! **B:** That's my job, to be there to keep things straight. And that's what you're doing – taking my business! And you're a f**king murderer! Look, I 'll call you when I get there on the other side!
D5	**A:** You come into the Five Points and you stole from me. **B:** I didn't steal anything. **A:** You came into the Five Point and stole your sandwich from me. I didn't steal nothing from you. **B:** I said I didn't steal anything..
D6	**A:** You come into the Five Points and you stole from me. **B:** I didn't steal anything. **A:** It's okay, he's a cop anyway, he's gonna say so. I was there when he stole it. **B:** He's a cop. He doesn't steal anything.
D7	**A:** You come into the Five Points and you stole from me. **B:** No, no, no, no. I just happened to see him outside... **A:** You stole from me. **B:** I didn't steal anything, but you're going to prison for stealing from me, and you're going to die! You're going to go to hell!
D8	**A:** You come into the Five Points and you stole from me. **B:** Who's gonna steal a car from you, anyway? **A:** This is my car! I didn't steal it. It's not stolen, it's just here. **B:** How's that for a confession?
DS∗	**A:** You come into the Five Points and you stole from me. **B:** I don't know... **A:** You stole Vallon. He was my kill. My example, of my power. You took him and made him yours. **B:** You're a lunatic to come here like...

Romance

Table B.7 Romance dialogues for evaluation in User Study ii.

Personality Code	Dialogue for Evaluation
RN∗	**A:** I haven't seen you cry since you were a baby. This must be serious. Where'd you meet him? **B:** At a birthday party. In New York. **A:** Well, you have a lot of catching up to do. You've got a lot to do. **B:** Well, you've seen me do pretty much...

Continuation of Table B.3

Personality Code	Dialogue for Evaluation
R1	**A:** I haven't seen you cry since you were a baby. This must be serious. Where'd you meet him? **B:** At a party. Why? I'm not supposed to talk about him. **A:** Why is he at a party? **B:** I'm not allowed to say. Ask a stupid question...can anybody explain why he was at a party?
R2	**A:** I haven't seen you cry since you were a baby. This must be serious. Where'd you meet him? **B:** At a party. **A:** A party? **B:** Yes. That's right. A birthday party, that was in France – oh, I didn't realize you were American!
R3∗	**A:** I haven't seen you cry since you were a baby. This must be serious. Where'd you meet him? **B:** At a party. A few years ago. **A:** The 'at a party'part – **B:** There was no party – it was just – he was there – and I – I just sat and watched him work – he had a perfect smile –
R4∗	**A:** I haven't seen you cry since you were a baby. This must be serious. Where'd you meet him? **B:** At my cousin's house. **A:** At your cousin's house? Did you ever see him cry? **B:** No. But I think it was very natural. And we became very close.
R5	**A:** I haven't seen you cry since you were a baby. This must be serious. Where'd you meet him? **B:** At a party. **A:** At a party?! **B:** He was throwing out an invitation...
R6	**A:** I haven't seen you cry since you were a baby. This must be serious. Where'd you meet him? **B:** At a party. In Rhode Island. **A:** What was the party about? **B:** I just thought I 'd show you.
R7∗	**A:** I haven't seen you cry since you were a baby. This must be serious. Where'd you meet him? **B:** At a party. In New York. **A:** Did you tell him? **B:** Did I tell him what? That I was in labor?
R8	**A:** I haven't seen you cry since you were a baby. This must be serious. Where'd you meet him? **B:** At my grandmother's. In a bar.

Continuation of Table B.3

Personality Code	Dialogue for Evaluation
	A: Wow, so you cried like a baby. That's so sad. **B:** It is. It's just... I had to....
RS∗	**A:** I haven't seen you cry since you were a baby. This must be serious. Where'd you meet him? **B:** On the road. **A:** Now, don't tell me you fell in love with a bus driver! **B:** No.

Thriller

Table B.8 Thriller dialogues for evaluation in User Study ii.

Personality Code	Dialogue for Evaluation
TN∗	**A:** So ... there's nothing you can tell me about Paul Owen? **B:** I've already told you about him. The man has committed suicide. **A:** He did? **B:** He's in a hospital right now. He tried to suicide by injection...
T1∗	**A:** So ... there's nothing you can tell me about Paul Owen? **B:** I can tell you that there's something we don't know about him. **A:** That's not reassuring. **B:** I want you to believe me when I tell you that.
T2∗	**A:** So ... there's nothing you can tell me about Paul Owen? **B:** He's a friend of mine. He's been with me since I was a kid... **A:** Well, he was with me, too... I know what you've been trying to tell him, but – **B:** Yeah, well, I told him, he thinks I'm an as***le, he can't tell.
T3∗	**A: A:** So ... there's nothing you can tell me about Paul Owen? **B:** I know he was working for the railroad. But what I can tell you is that he did it. And that's all that's necessary. And that's all that matters. **A:** I thought you weren't going to tell me anything. **B:** It wasn't me that killed him. It was you.
T4	**A:** So ... there's nothing you can tell me about Paul Owen? **B:** Nothing. No idea what he's into. **A:** He's a little bit of a dickhead, and his voice doesn 't change a goddamn bit. And, you know, you know, you know, he really hates his wife and his kid. **B:** You really don 't know him, do you, buddy?
T5∗	**A:** So ... there's nothing you can tell me about Paul Owen? **B:** He's dead. **A:** How do you know that? **B:** The police report. He's been found dead.

Continuation of Table B.4

Personality Code	Dialogue for Evaluation
T6	**A:** So ... there's nothing you can tell me about Paul Owen? **B:** No, I don 't think so. I can 't think of a single thing about him. **A:** Paul Owen's dead, Mr.Creasy. **B:** I don 't believe that.
T7	**A:** So ... there's nothing you can tell me about Paul Owen? **B:** I... don 't know. **A:** But that's not what I want to know. **B:** I don 't care who you know or what you know about Paul Owen. If you want to get into this it's a matter of time before you're arrested.
T8	**A:** So ... there's nothing you can tell me about Paul Owen? **B:** He's dead. **A:** Really? **B:** You've seen him die. You don 't know the whole goddamn thing?
TS*	**A:** So ... there's nothing you can tell me about Paul Owen? **B:** He led what I suppose was an orderly life. He... ate a balanced diet. **A:** What kind of man was he? Besides... the information you've just given. **B:** I hope I'm not being cross-examined here.

B.3 Dialogues for User Study iii(Chapter 7)

In this user study, all dialogues were generated with the setting of Drama genre.

CTE

Table B.9 Dialogues generated using *CTE* for evaluation in User Study iii.

Personality Code	Dialogue for Evaluation
Film 1	**A:** What on earth do you think the baby will be doing?! **B:** I said I thought the baby will be doing the whole dance... But I could be wrong! **A:** The baby is dancing on the stage, isn 't it? You said it was dancing on stage, but when the baby comes out you're going to get the whole f**king dance. **B:** Well, you've got ta take time off and do it at home.
Film 8	**A:** What on earth do you think the baby will be doing?! **B:** I don 't know what the baby's going to be doing. **A:** It won 't be. You will. Don 't go around saying you can 't. We 'll figure it out. You're a good friend. **B:** No. I'm a selfish little freak.
Scene 1	**A:** What on earth do you think the baby will be doing?! **B:** Where's the baby? The baby is dead. **A:** That's what I said. It's a little far down there but the boy's gonna like it. I 'll be at the baby's parent house every day until they wake up and we're all ready for you. **B:** I 'll make my own baby. I promise.
Scene 8	**A:** What on earth do you think the baby will be doing?! **B:** I don 't know. We should ask the kid. **A:** It's not a good idea. The baby needs to be in the living room. **B:** Why are you telling us?
Film 1	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars. **B:** You 'll get more gas in New York. **A:** The best I can do is fifty cents. **B:** And fifty cents is the most you can get for twenty-five dollars?
Film 8	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars. **B:** It's not my price. **A:** It's your price. **B:** I'm not selling my ticket.
Scene 1	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars.

Continuation of Table B.9

Personality Code	Dialogue for Evaluation
	B: It's twenty-five dollars? That's ridiculous. How do you know? How do you know what the deal's worth? **A:** I got that from the guy who ran this thing. **B:** The guy who ran it, huh?
Scene 8	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars. **B:** It's a joke! You're a phony! **A:** I'm sorry I'm a phony! **B:** That's right.

SCE

Table B.10 Dialogues generated using *SCE* for evaluation in User Study iii.

Personality Code	Dialogue for Evaluation
Film 1	**A:** What on earth do you think the baby will be doing?! **B:** I know it's just a baby, but it's still a baby. I mean, I mean – you're not going to give him up for adoption! I can 't do it! **A:** He's only seven weeks in this condition. The doctor said maybe it's okay if we keep him. I told you, I can 't do this anymore. **B:** We're fine. I can handle it.
Film 8	**A:** What on earth do you think the baby will be doing?! **B:** What baby? What baby!? What baby is that?!? **A:** It's so cute. It's cute, huh **B:** It's all right baby, I'm just so excited for the baby. I'm just so excited! I'm so... Excite! Excite! I love it!
Scene 1	**A:** What on earth do you think the baby will be doing?! **B:** 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. **A:** 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, the child's father. **B:** Then he will be a son of God.
Scene 8	**A:** What on earth do you think the baby will be doing?! **B:** I don 't know. You can 't have your own son and your son will never see the world. **A:** What about the baby? What about him? **B:** I don 't know if he was born yet.
Film 1	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars.

Continuation of Table B.10

Personality Code	Dialogue for Evaluation
	B: You know, the first thing I got in college was money. And I don 't know how much I know about the game. **A:** What does that mean? **B:** It means you have to get out to get your first ball. I have a twenty-five dollar bag and I can throw it anywhere, anywhere.
Film 8	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars. **B:** It's all right if they think it's for money. **A:** I'm not asking for money. **B:** Then stop asking. I've got two cars, I need two cars.
Scene 1	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars. **B:** That depends on how close you are to New York. **A:** I'm just going to New York, maybe a little farther, you get to work. **B:** That's all you really need. Get some work, get some sleep, maybe some breakfast. You can do it in about fifteen hours.
Scene 8	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars. **B:** I don 't think it has to be. **A:** I'm telling you, it does. **B:** No, no. There's too many of them, too. You've got one... and it sucks. It's no good.

STE

Table B.11 Dialogues generated using *STE* for evaluation in User Study iii.

Personality Code	Dialogue for Evaluation
Film 1	**A:** What on earth do you think the baby will be doing?! **B:** I don 't know. I was just thinking about how you were a good parent. **A:** I think I've come up with a better one... **B:** Oh, we 'll get it.
Film 8	**A:** What on earth do you think the baby will be doing?! **B:** I can 't tell you, but I'm not going to be seeing my baby. **A:** It's been two years and she still has the baby? **B:** I can 't tell you, but I 'll be telling you tomorrow. She's still going to make a baby.
Scene 1	**A:** What on earth do you think the baby will be doing?! **B:** I don 't know. The doctor said... she could have a brain hemorrhage. She might have gestational hypertension. Maybe it's an epidural, but that's not good.

Continuation of Table B.11

Personality Code	Dialogue for Evaluation
	A: So what do you guys do? **B:** We go to medical school.
Scene 8	**A:** What on earth do you think the baby will be doing?! **B:** I can 't see it. **A:** That's right! You can 't see it! I'm not seeing it! **B:** Well, you can still see it – because it's not happening!
Film 1	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars. **B:** I've already been to New York. What else you've got to ask for? **A:** I don 't know. I have to figure out something for my mom. The only way I want to go to New York is if she's there. **B:** How 'd you know her?
Film 8	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars. **B:** That s**t is expensive. **A:** I need gas. **B:** You know how to get gas?
Scene 1	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars. **B:** Twenty-five dollars? What the f**k!! I can 't get a dollar for a twenty-five dollar bag with all my gas. **A:** You don 't know how much gas you're paying. I 'll pay you the same for a twenty-five dollar bag. If you have to, go with it. **B:** So you wan na do both this and not go to work anymore?
Scene 8	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars. **B:** We 'll take twenty-five. **A:** You know what I'm going to do? I'm going to start selling bags to people! **B:** So what?

SDG

Table B.12 Dialogues generated using *SDG* for evaluation in User Study iii.

Personality Code	Dialogue for Evaluation
Film 1	**A:** What on earth do you think the baby will be doing?! **B:** I think you have to be a little more careful with that baby as it is a baby. **A:** I think I did it right. **B:** I'm not going to worry about it, I'm not scared of you.

Continuation of Table B.12

Personality Code	Dialogue for Evaluation
Film 8	**A:** What on earth do you think the baby will be doing?! **B:** It's not going to the baby. What happens to the baby? **A:** The baby does not die. He takes his own life to make his last breath. **B:** I think you got your baby wrong.
Scene 1	**A:** It could be very dangerous to a baby. **B:** This is why I keep my baby in the back seat of a car. **A:** So what do you guys do? **B:** No, he was actually getting out of the car to take it to his brother.
Scene 8	**A:** What on earth do you think the baby will be doing?! **B:** I don 't know. I don 't know. I do know that the baby is ready. **A:** You know that baby has not been dead for a long time, right? **B:** No, not right now.
Film 1	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars. **B:** No thanks, we're 't going to make you give us a ticket, so you're just waiting here. **A:** It's a deal! That's it, what it's, five cents each! You're all good. **B:** Yeah.
Film 8	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars. **B:** A 'pig ain't got any gas to go with it. **A:** I have a million dollars to give. **B:** Okay, okay. You can have it.
Scene 1	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars. **B:** Well, I'm in New Jersey. **A:** What about the New Jersey? **B:** I'm sure the New York and Connecticut are on the short list.
Scene 8	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars. **B:** You don 't have a single dollar in stock. **A:** You're a fool to believe that a dollar per square centimeter of steel is worth anything. **B:** Why did you think such a stupid thing?

NOP

Table B.13 Dialogues generated without personality for evaluation in User Study iii.

Personality Code	Dialogue for Evaluation
N/A	**A:** What on earth do you think the baby will be doing?! **B:** She will be fine. The baby will be fine. **A:** I don 't trust you. **B:** I don 't need any help from you or anyone else.
N/A	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars. **B:** That's not enough for me to drive to New York. **A:** You want me to drive you to New York.? **B:** I want you to go with me.

GPT

Table B.14 Dialogues generated using DialoGPT for evaluation in User Study iii.

Personality Code	Dialogue for Evaluation
N/A	**A:** What on earth do you think the baby will be doing?! **B:** Baby is going to be wearing a suit. He'll be dressed as an adult. **A:** So he'll be dressed as a baby? **B:** No, a baby is a person who is under 18 years old.
N/A	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars. **B:** You're getting gas? **A:** I got gas when I bought the truck. **B:** I'm sure you'll be just as successful as me at the gas station.

ORG

Table B.15 Dialogues collected from source script for evaluation in User Study iii.

Personality Code	Dialogue for Evaluation
N/A	**A:** What on earth do you think the baby will be doing?! **B:** Well, I'm sure I don't know, darling. Luck favors the prepared. I didn't know the baby's powers, so I covered the basics. **A:** Jack-Jack doesn't have any powers. **B:** No? Well, he'll look fabulous anyway.
N/A	**A:** All I'm asking is enough gas to get me to New York. The bag's worth twenty-five dollars.

Continuation of Table B.15

Personality Code	Dialogue for Evaluation
	B: Yeah, but I got a bag. My wife gave me one for Christmas. **A:** Listen, man. I'll tell you what I'll do. When I come back in the morning, I'll buy it back from you and give you ten dollars profit? What do you say? **B:** I ain't got a hat.

Appendix C

List of Films used in the dataset

All films used in the dataset were collected from IMSDb on 22.10.2019.

(Alphabetical order)

10 Things I Hate About You, 1997
12, 2003
12 and Holding, 2006
12 Monkeys, 1994
12 Years a Slave, 2013
127 Hours, 2010
1492 Conquest of Paradise, 1991
15 Minutes, 2001
17 Again, 2009
187, 1996
2001 A Space Odyssey, 1989
2012, 2009
25th Hour, 2001
28 Days Later, 2003
30 Minutes or Less, 2011
42, 2013
44 Inch Ches, 2010
48 Hrs, 1982
50 50, 2011
500 Days of Summer, 2009
8MM, 1997
9, 2009

A Few Good Men, 1992
A Most Violent Year, 2015
A Prayer Before Dawn, 2018
A Quiet Place, 2018
A Scanner Darkly, 2006
A Serious Man, 2009
Above the Law, 1987
Absolute Power, 1996
Abyss The, 1988
Ace Ventura Pet Detective, 1994
Adaptation, 2000
Addams Family The, 1991
Adjustment Bureau The, 2011
Adventures of Buckaroo Banzai Across the Eighth Dimension The, 1983
Affliction, 1999
After School Special, 2000
After.Life, 2010
Agnes of God, 1985
Air Force One, 1997
Airplane, 1980
Airplane 2 The Sequel, 1982
Aladdin, 1992
Ali, 2001
Alien, 1978
Alien 3, 1991
Alien Nation, 1987
Alien vs. Predator, 2004
Alien Resurrection, 1997
Aliens, 1985
All About Eve, 1950
All About Steve, 2009
All the King's Men, 1949
All the President's Men, 1975
Almost Famous, 1998
Alone in the Dark, 2005
Amadeus, 1984
Amelia, 2009

American Beauty, 1999
American Gangster, 2007
American Graffiti, 1973
American History X, 1997
American Hustle, 2013
American Madness, 1932
American Milkshake, 2013
American Outlaws, 2001
American Pie, 1998
American President The, 1995
American Psycho, 2000
American Shaolin King of Kickboxers II, 1991
American Sniper, 2015
American Splendor, 2003
American Werewolf in London, 1981
American The, 2010
Amityville Asylum The, 2014
Amour, 2012
An Education, 2009
Analyze That, 2002
Analyze This, 1998
Anastasia, 1997
Angel Eyes, 1999
Anna Karenina, 2012
Annie Ha, 1977
Anniversary Party The, 2001
Anonymous, 2011
Antitrust, 2001
Antz, 1998
Apartment The, 1991
Apocalypse Now, 1975
Apollo 13, 1995
April Fool's Day, 1986
Apt Pupil, 1996
Arbitrage, 2012
Arcade, 1990
Arctic Blue, 1993

Argo, 2012
Armageddon, 1998
Army of Darkness, 1991
Arsenic and Old Lace, 1944
Arthur, 2011
Artist The, 2012
As Good As It Gets, 1997
Assassins, 1994
Assignment The, 1997
At First Sight, 1999
August Osage County, 2014
Austin Powers International Man of Mystery, 1996
Austin Powers The Spy Who Shagged Me, 1999
Authors Anonymous, 2014
Autumn in New York, 2000
Avatar, 2009
Avengers The, 1995
Avengers The (2012), 2012
Avventura L' (The Adventure), 1960
Awakenings, 1989
Babe, 2006
Bachelor Party, 2004
Bachelor Party The, 1957
Back up Plan The, 2010
Backdraft, 1991
Bad Boys, 1994
Bad Country, 2014
Bad Day at Black Rock, 1955
Bad Dreams, 1988
Bad Lieutenant, 1992
Bad Santa, 2002
Bad Teacher, 2011
Badlands, 1973
Bamboozled, 2000
Barry Lyndon, 1973
Barton Fink, 1991
Basic, 2000

Basic Instinct, 1992
Basquiat, 1996
Batman, 1988
Batman 2, 2008
Battle of Algiers The, 1966
Battle of Shaker Heights The, 2003
Battle Los Angeles, 2011
Beach The, 1998
Bean, 1997
Beasts of No Nation, 2015
Beasts of the Southern Wild, 2012
Beauty and the Beast, 2017
Beavis and Butt head Do America, 1996
Beginners, 2011
Being Human, 1992
Being John Malkovich, 1999
Being There, 1979
Believer The, 2002
Belle, 2014
Beloved, 1998
Best Exotic Marigold Hotel The, 2012
Big, 1988
Big Blue The, 1988
Big Eyes, 2014
Big Fist, 2004
Big Lebowski The, 1998
Big Sick The, 2017
Big White The, 2005
Birdman, 2014
Birds The, 1962
Birthday Girl, 2001
Black Dahlia The, 2006
Black Panther, 2018
Black Rain, 1987
Black Snake Moan, 2007
Black Swan, 2010
BlacKkKlansman, 2018

Blade, 1998
Blade II, 2002
Blade Runner, 1981
Blade Trinity, 2004
Blast from the Past The, 1999
Blind Side The, 2009
Bling Ring The, 2013
Blood and Wine, 1997
Blood Simple, 1984
Blow, 2001
Blue Valentine, 2010
Blue Velvet, 1986
Body Heat, 1981
Body of Evidence, 1993
Bodyguard, 1992
Bones, 2001
Bonfire of the Vanities, 1990
Bonnie and Clyde, 2013
Boogie Nights, 1997
Book of Eli The, 2010
Boondock Saints 2 All Saints Day, 2009
Boondock Saints The, 2000
Bottle Rocket, 1996
Bound, 1996
Bounty Hunter The, 2010
Bourne Identity The, 2002
Bourne Supremacy The, 2004
Bourne Ultimatum The, 2007
Box The, 2009
Boxtrolls The, 2014
Boyhood, 2014
Braveheart, 1995
Brazil, 1985
Break, 1969
Breakdown, 1996
Breakfast Club The, 1985
Breaking Away, 1979

Brick, 2006
Bridesmaids, 2011
Bringing Out the Dead, 1997
Broadcast News, 1987
Broken Arrow, 1996
Broken Embraces, 2009
Brothers Bloom The, 2008
Bruce Almighty, 2002
Buffy the Vampire Slayer, 1992
Bull Durham, 1988
Buried, 2010
Burlesque, 2010
Burn After Reading, 2008
Burning Annie, 2007
Butterfly Effect The, 2004
Cable Guy, 1996
Candle to Water, 2012
Capote, 2006
Carrie, 1976
Cars 2, 2011
Case 39, 2010
Casino, 1995
Cast Away, 2000
Catch Me If You Can, 2002
Catwoman, 2004
Cecil B. Demented, 2000
Cedar Rapids, 2011
Cell The, 2000
Cellular, 2004
Change Up The, 2011
Changeling, 2008
Chaos, 2005
Charade, 1963
Charlie's Angels, 1999
Chasing Amy, 1997
Chasing Sleep, 2001
Cherry Falls, 2000

Chinatown, 1974
Christ Complex, 2012
Chronicle, 2012
Chronicles of Narnia The Lion the Witch and the Wardrobe, 2005
Cider House Rules The, 1999
Cincinnati Kid The, 1965
Cinema Paradiso, 1988
Cirque du Freak The Vampire's Assistan, 2009
Citizen Kane, 1941
City of Joy, 1990
Clash of the Titans, 2010
Clerks, 1994
Cliffhanger, 1993
Clueless, 1995
Cobb, 1993
Coco, 2017
Code of Silence, 1985
Cold Mountain, 2003
Collatera, 2004
Collateral Damage, 2000
Colombiana, 2011
Color of Nig, 1994
Commando, 1985
Conan the Barbarian, 2011
Confessions of a Dangerous Mind, 1998
Confidence, 2000
Constantine, 2005
Cooler The, 2003
Copycat, 1995
Coraline, 2009
Coriolanus, 2012
Cradle 2 the Grave, 2002
Crank, 2006
Cras, 1996
Crazy Stupid Love, 2011
Crazylove, 2005
Creation, 2010

Crime Spree, 2003
Croods The, 2013
Crouching Tiger Hidden Dragon, 2000
Croupier, 2000
Crow Salvation The, 2000
Crow The, 1992
Crow City of Angels The, 1996
Cruel Intentions, 1998
Crying Game, 1992
Cube, 1997
Curious Case of Benjamin Button The, 2008
Custody, 2017
Dallas Buyers Club, 2013
Damned United The, 2009
Dances with Wolves, 1990
Danish Girl The, 2016
Dark City, 1994
Dark Knight Rises The, 2012
Dark Star, 1974
Darkman, 1990
Date Night, 2010
Dave Barry's Complete Guide to Guys, 2006
Dawn of the Dead, 1977
Day of the Dead, 1985
Day the Clown Cried The, None
Day the Earth Stood Still The, 1951
Days of Heaven, 1976
Dead Poets Society, 1989
Deadpool, 2016
Dear White People, 2015
Death at a Funeral, 2010
Death to Smoochy, 1997
Debt The, 2011
Deception, 2008
Deep Cover, 1992
Deep Rising, 1996
Deer Hunter The, 1978

Defiance, 2009
Departed The, 2006
Descendants The, 2011
Despicable Me 2, 2013
Detroit Rock City, 1999
Devil in a Blue Dress, 1995
Devil Wears Prada The, 2006
Devil's Advocate, 1997
Die Hard, 1988
Die Hard 2, 1990
Diner, 1982
Distinguished Gentleman The, 2000
Disturbia, 2007
Django Unchained, 2012
Do The Right Thing, 1988
Dog Day Afternoon, 1975
Dogma, 1999
Donnie Brasco, 1992
Doors The, 1991
Double Indemnity, 1944
Drag Me to He, 2009
Dragonslayer, 1981
Drive, 2011
Drive Angry, 2011
Drop Dead Gorgeous, 1999
Dry White Season A, 1987
Duck Soup, 1933
Dumb and Dumber, 1993
Dune, 1983
E.T, 1982
Eagle Eye, 2008
Eastern Promises, 2007
Easy A, 2010
Ed TV, 1997
Ed Wood, 1992
Edward Scissorhands, 1990
Eight Legged Freaks, 2000

El Mariachi, 1993
Election, 1997
Elephant Man The, 1980
Elizabeth The Golden Age, 2007
Enemy of the State, 1998
English Patient The, 1996
Enoug, 2002
Entrapmen, 1996
Erik the Viking, 1989
Erin Brockovic, 1999
Escape From L.A, 1996
Escape From New York, 1981
Eternal Sunshine of the Spotless Mind, 2004
Even Cowgirls Get the Blues, 1994
Event Horizon, 1997
Evil Dead, 1979
Evil Dead II Dead by Dawn, 1986
Ex Machina, 2015
Excalibur, 1981
eXistenZ, 1999
Extract, 2009
Fabulous Baker Boys The, 1985
Face Off, 1997
Fair Game, 2010
Family Man The, 2000
Fantastic Four, 2005
Fantastic Mr Fox, 2009
Fargo, 1996
Fast Times at Ridgemont High, 1982
Fatal Instinct, 1993
Fault in Our Stars The, 2014
Fear and Loathing in Las Vegas, 1998
Feas, 2006
Ferris Bueller's Day Off, 1985
Field of Dreams, 1989
Fifth Element The, 1995
Fight Club, 1998

Fighter The, 2010
Final Destination, 1999
Final Destination 2, 2003
Finding Nemo, 2003
Five Easy Pieces, 1970
Flash Gordon, 1980
Fletc, 1986
Flig, 2012
Flintstones The, 1987
Forrest Gump, 1994
Four Feathers, 2002
Four Rooms, 1995
Foxcatcher, 2015
Fracture, 2007
Frances, 1982
Frankenstein, 1994
Frankenweenie, 2012
Freaked, 1993
Freddy vs. Jason, 2003
French Connection The, 1971
Frequency, 2000
Friday the 13, 1980
Friday the 13th Part VIII Jason Takes Manhattan, 1989
Fright Night, 2011
Fright Night (1985), 1985
From Dusk Till Dawn, 1996
From Here to Eternity, 1953
Frozen, 2010
Frozen (Disney), 2013
Frozen River, 2008
Fruitvale Station, 2013
Fugitive The, 1992
Funny People, 2009
G.I. Jane, 1995
G.I. Joe The Rise of Cobra, 2009
Game 6, 2005
Game The, 1996

Gamer, 2009
Gandhi, 1982
Gang Related, 1997
Gangs of New York, 2002
Garden State, 2004
Gattaca, 1997
Get Carter, 1971
Get Low, 2010
Get on Up, 2014
Get Ou, 2017
Get Shorty, 1995
Getaway The, 1972
Ghos, 1990
Ghost and the Darkness The, 1996
Ghost Rider, 2007
Ghost Ship, 2002
Ghost World, 2001
Ghostbusters, 1988
Ghostbusters 2, 1989
Girl with the Dragon Tattoo The, 2011
Gladiator, 1998
Glengarry Glen Gross, 1992
Go, 1997
Godfather, 1971
Godfather Part II, 1973
Godfather Part III The, 1990
Gods and Monsters, 1997
Godzilla, 1996
Gone Baby Gone, 2007
Gone in 60 Seconds, 1999
Good Girl The, 2002
Good Will Hunting, 1997
Gothika, 2003
Grabbers, 2012
Graduate The, 1967
Gran Torino, 2009
Grand Hote, 1932

Grand Theft Parsons, 2004
Grapes of Wrath The, 1940
Gravity, 2013
Great Gatsby The, 2013
Green Mile The, 1997
Gremlins, 1984
Gremlins 2, 1990
Grifters The, 1989
Grosse Point Blank, 1994
Groundhog Day, 1993
Grudge The, 2004
Guardians of the Galaxy Vol 2, 2017
Hackers, 1995
Hall Pass, 2011
Halloween, 2018
Halloween The Curse of Michael Myers, 1995
Hancock, 2008
Hangover The, 2009
Hanna, 2011
Hannah and Her Sisters, 1986
Hannibal, 2000
Happy Birthday Wanda June, 1971
Happy Fee, 2006
Hard Rain, 1998
Hard to Kill, 1990
Harold and Kumar Go to White Castle, 2004
Haunting The, 1998
He's Just Not That Into You, 2009
Heat, 1994
Heathers, 1988
Heavenly Creatures, 1994
Heavy Meta, 1980
Hebrew Hammer The, 2003
Heist, 1999
Hellbound Hellraiser II, 2000
Hellboy, 2004
Hellboy 2 The Golden Army, 2008

Hellraiser, 1986
Hellraiser 3 Hell on Ear, 1992
Hellraiser Deader, 2005
Hellraiser Hellseeker, 2002
Help The, 2011
Henry Foo, 1998
Henry's Crime, 2011
Her, 2014
Hesher, 2011
High Fidelity, 1998
Highlander, 1986
Highlander Endgame, 1999
Hills Have Eyes The, 2006
His Girl Friday, 1940
Hitchcock, 2012
Hitchhiker's Guide to the Galaxy The, 2005
Hollow Man, 1998
Honeydripper, 2007
Horrible Bosses, 2011
Horse Whisperer The, 1997
Hospital The, 1971
Hostage, 2005
Hot Tub Time Machine, 2010
Hotel Rwanda, 2005
House of 1000 Corpses, 2000
How to Train Your Dragon, 2010
How to Train Your Dragon 2, 2014
Hudson Hawk, 1990
Hudsucker Proxy The, 1992
Human Nature, 2001
Hunt for Red October The, 1990
Hurt Locker The, 2009
I Am Number Four, 2011
I am Sam, 2002
I Love You Phillip Morris, 2010
I Spit on Your Grave, 2010
I Still Know What You Did Last Summer, 1998

I'll Do Anything, 1994
I Robot, 2004
Ice Storm The, 1996
Ides of March The, 2011
Imaginarium of Doctor Parnassus The, 2009
In the Bedroom, 2002
In the Loop, 2009
Inception, 2010
Incredibles The, 2004
Independence Day, 1996
Indiana Jones and the Last Crusade, 1989
Indiana Jones and the Raiders of the Lost Ark, 1981
Indiana Jones and the Temple of Doom, 1984
Indiana Jones IV, 1995
Informant The, 2009
Inglourious Basterds, 2009
Insider The, 1999
Insidious, 2011
Insomnia, 2002
Interstellar, 2014
Interview with the Vampire, 1994
Into the Wild, 2007
Into the Woods, 2014
Intolerable Cruelty, 1997
Inventing the Abbotts, 1996
Invention of Lying The, 2009
Invictus, 2009
Iron Lady The, 2012
Island The, 2005
I, 2017
It Happened One Night, 1934
It Happened One Night, 1934
It's a Wonderful Life, 1946
It's Complicated, 2009
Italian Job The, 2001
Jacket The, 2005
Jackie Brown, 1997

Jacob's Ladder, 1990
Jane Eyre, 2011
Jason X, 2001
Jaws, 1975
Jaws 2, 1978
Jay and Silent Bob Strike Back, 2001
Jennifer Eight, 1992
Jennifer's Body, 2009
Jerry Maguire, 1996
Jeux Interdits, 1952
JFK, 1991
Jimmy and Judy, 2004
John Q, 2002
John Wick, 2014
Joker, 2019
Judge Dredd, 1995
Juno, 2007
Jurassic Park, 1992
Jurassic Park III, 2001
Jurassic Park The Lost World, 1997
Kafka, 1991
Kalifornia, 1993
Kids, 1995
Kids Are All Right The, 2010
Kill Your Darlings, 2013
Killing Zoe, 1993
King Kong, 2005
King of Comedy The, 1976
King's Speech The, 2010
Kingdom The, 2007
Klute, 1971
Knocked Up, 2007
Kramer vs Kramer, 1979
Kundun, 1992
Kung Fu Panda, 2008
L.A. Confidentia, 1995
La La Land, 2016

Labor of Love, None
Labyrin, 1986
Ladykillers The, 2004
Lake Placid, 1999
Land of the Dead, 2005
Larry Crowne, 2011
Last Boy Scout The, 1991
Last Chance Harvey, 2009
Last Flight The, 1931
Last of the Mohicans The, 1992
Last Samurai The, 2003
Last Station The, 2009
Last Tango in Paris, 1973
Law Abiding Citizen, 2009
Le Diable par la Queue, 1969
Leaving Las Vegas, 1994
Legally Blonde, 2000
Legend, 1984
Legion, 2010
LEGO Movie The, 2014
Les Miserables, 2012
Les Tontons Flingueurs, 1963
Leviathan, 1987
Liar Liar, 1997
Life, 1999
Life As A House, 2001
Life of David Gale The, 2003
Life of Pi, 2012
Light Sleeper, 1992
Limey The, 1998
Limitless, 2011
Lincoln, 2012
Lincoln Lawyer The, 2011
Little Athens, 2006
Little Men, 2016
Little Mermaid The, 1989
Little Nicky, 2000

Living in Oblivion, 1995
Lock Stock and Two Smoking Barrels, 1998
Logan, 2017
Logan's Run, 1975
Lone Star, 1996
Long Kiss Goodnight The, 1996
Looper, 2012
Lord of Illusions, 1994
Lord of the Rings Fellowship of the Ring The, 2001
Lord of the Rings Return of the King, 2003
Lord of the Rings The Two Towers, 2002
Lord of War, 2005
Losers The, 2010
Lost Highway, 1995
Lost Horizon, 1937
Lost in Space, 1998
Lost in Translation, 2003
Love and Basketball, 2000
Machete, 2010
Machine Gun Preacher, 2011
Mad Max 2 The Road Warrior, 1982
Made, 2001
Magnolia, 1998
Majestic The (The Bijou), 1997
Major League, 1989
Malcolm X, 1991
Malibu's Most Wanted, 2002
Man in the Iron Mask, 1995
Man On Fire, 2004
Man on the Moon, 1999
Man Trouble, 1991
Man Who Knew Too Much The, 1955
Man Who Wasn't There The, 2001
Manchurian Candidate The, 2004
Manhattan Murder Mystery, 1993
Manhunter, 1984
Margaret, 2011

Mission Impossible, 1995
Mission Impossible II, 1994
Mission to Mars, 2000
Moneyball, 2011
Monkeybone, 2001
Monte Carlo, 2011
Moon, 2009
Moonrise Kingdom, 2012
Moonstruck, 1987
Mr Blandings Builds His Dream House, 1948
Mr Brooks, 2007
Mr Deeds Goes to Town, 1936
Mrs. Brown, 1997
Mud, 2013
Mulan, 1998
Mulholland Drive, 1999
Mumford, 1999
Mummy The, 1999
Music of the Hear, 1999
Mute Witness, 1995
My Best Friend's Wedding, 1997
My Girl, 1991
My Mother Dreams the Satan's Disciples in New York, 1998
My Week with Marilyn, 2011
Mystery Men, 1997
Napoleon Dynamite, 2004
Nashville, 1974
Natural Born Killers, 1995
Never Been Kissed, 1998
Neverending Story The, 1984
New York Minute, 2004
Newsies, 1991
Next, 2007
Next Friday, 2000
Next Three Days The, 2010
Ni vu ni connu, 1958
Nick of Time, 1995

Night Time (The Poltergeist Treatment), 1982
Nightbreed, 1990
Nightmare Before Christmas The, 1991
Nightmare Before Christmas The, 1991
Nightmare on Elm Street A, 1984
Nightmare on Elm Street The Final Chapter, 1985
Nine, 2009
Nines The, 2007
Ninja Assassin, 2009
Ninotchka, 1939
Ninth Gate The, 1999
No Country for Old Men, 2007
No Strings Attached, 2011
Notting Hill, 1999
Nurse Betty, 1999
Oblivion, 2013
Observe and Report, 2009
Obsessed, 2009
Ocean's Eleven, 2001
Ocean's Twelve, 2004
Office Space, 1997
Omega Man, 1970
One Flew Over the Cuckoo's Nest, 1975
Only God Forgives, 2013
Ordinary People, 1980
Orgy of the Dead, 1965
Orphan, 2009
Other Boleyn Girl The, 2008
Out of Sight, 1998
Pacifier The, 2005
Pandorum, 2009
Panic Room, 2000
ParaNorman, 2012
Pariah, 2011
Passengers, 2016
Passion of Joan of Arc The, 1929
Patriot The, 1999

Paul, 2011
Pearl Harbor, 2001
Peeping To, 1960
Peggy Sue Got Married, 1985
Perfect Creature, 2007
Perfect World A, 1992
Perks of Being a Wallflower The, 2012
Pet Sematary, 1986
Pet Sematary II, 1991
Petulia, 1968
Philadelphia, 1992
Phone Boo, 2002
Pi, 1998
Pianist The, 2002
Piano The, 1991
Pineapple Express, 2008
Pirates of the Caribbean, 2003
Pirates of the Caribbean Dead Man's Chest, 2006
Pitch Black, 1998
Planet of the Apes The, 1968
Platinum Blonde, 1931
Platoon, 1986
Pleasantville, 1998
Point Break, 1991
Pokemon Mewtwo Returns, 2000
Postman The, 1996
Power of One The, 1990
Precious, 2009
Predator, 1987
Prestige The, 2006
Pretty Woman, 1990
Pretty Woman (final script), 1990
Pride and Prejudice, 2005
Pries, 2011
Princess Bride The, 1987
Private Life of Sherlock Holmes The, 1970
Producer The, 1967

Program The, 1993
Prom Night, 1980
Prometheus, 2012
Prophecy The, 1995
Proposal The, 2009
Psycho, 1959
Public Enemies, 2009
Pulp Fiction, 1993
Punch Drunk Love, 2002
Purple Rain, 1984
Quantum Project, 2000
Queen of the Damned, 2000
Queen The, 2006
Rachel Getting Married, 2008
Raging Bull, 1980
Raising Arizona, 1987
Rambling Rose, 1991
Rambo First Blood II The Mission, 1983
Reader The, 2009
Real Genius, 1985
Rear Window, 1953
Rebel Without A Cause, 1955
Red Planet, 2000
Red Riding Hood, 2011
Reindeer Games, 2000
Relic The, 1995
Remember Me, 2010
Replacements The, 1999
Repo Man, 1984
Rescuers Down Under The, 1990
Reservoir Dogs, 1990
Resident Evil, 2000
Return of the Apes, 1994
Revenant The, 2016
Revolutionary Road, 2008
Ringu, 1998
Rise of the Guardians, 2012

Rise of the Planet of the Apes, 2011
RKO 281, 1999
Road The, 2009
Robin Hood Prince of Thieves, 1990
Rock The, 1995
RocknRolla, 2008
Rocky, 1976
Rocky Horror Picture Show The, 1975
Ronin, 1997
Room, 2016
Roommate The, 2011
Roughshod, 1949
Ruins The, 2008
Runaway Bride, 1999
Rush, 2013
Rush Hour, 1998
Rush Hour 2, 2001
Rushmore, 1997
Rust and Bone, 2012
S. Darko, 2009
Saint The, 1995
Salton Sea The, 2002
Sandlot Kids The, 1993
Save the Last Dance, 1999
Saving Mr. Banks, 2013
Saving Private Ryan, 1998
Saw, 2004
Scarface, 1983
Scary Movie 2, 2001
Schindler's List, 1993
Scott Pilgrim vs the World, 2010
Scream, 1995
Scream 2, 1997
Scream 3, 1999
Se7en, 1995
Searchers The, 1956
Secret Life of Walter Mitty The, 2013

Semi Pro, 2008
Sense and Sensibility, 1995
Serenity, 2005
Serial Mo, 1992
Sessions The, 2012
Seventh Seal The, 1957
Sex and the City, 2008
Sex Lies and Videotape, 1989
Sexual Life, 2005
Shakespeare in Love, 1998
Shallow Grave, 1995
Shame, 2011
Shampoo, 1975
Shawshank Redemption The, 1994
She's Out of My League, 2010
Sherlock Holmes, 2009
Shifty, 2009
Shining The, 1980
Shipping News The, 2002
Shivers, 1976
Shrek, 2001
Shrek the Third, 2007
Sicario, 2015
Sideways, 2005
Siege The, 1998
Signs, 2002
Silence of the Lambs, 1991
Silver Bullet, 1985
Silver Linings Playbook, 2012
Simone, 2002
Single White Female, 1992
Sister Ac, 1992
Six Degrees of Separation, 1993
Sixth Sense The, 1999
Sleepless in Seattle, 1992
Sleepy Hollow, 1998
Sling Blade, 1996

Slither, 2006
Slumdog Millionaire, 2009
Smashed, 2012
Smokin' Aces, 2007
Snatch, 2001
Snow Falling On Cedars, 1998
Snow White and the Huntsman, 2012
So I Married an Axe Murderer, 1993
Social Network The, 2010
Solaris, 2001
Soldier, 1998
Someone To Watch Over Me, 1986
Something's Gotta Give, 2003
Source Code, 2011
South Park, 1999
Space Milkshake, 2013
Spanglish, 2004
Spare Me, 1991
Spartan, 2002
Speed Racer, 2008
Sphere, 1998
Spider Man, 2000
St. Elmo's Fire, 1985
Star Trek, 2009
Star Trek II The Wrath of Khan, 1982
Star Trek First Contact, 1995
Star Trek Generations, 1994
Star Trek Nemesis, 2002
Star Trek The Motion Picture, 1978
Star Wars A New Hope, 1977
Star Wars Attack of the Clones, 2002
Star Wars Return of the Jedi, 1981
Star Wars Revenge of the Si, 2005
Star Wars The Empire Strikes Back, 1980
Star Wars The Force Awakens, 2015
Star Wars The Phantom Menace, 1999
Starman, 1984

Starship Troopers, 1997
State and Main, 1999
Station West, 1948
Stepmom, 1998
Sting The, 1973
Stir of Echoes, 1999
Storytelling, 2001
Straight Outta Compton, 2015
Strange Days, 1995
Strangers on a Train, 1950
Stuntman The, 1980
Sugar, 2009
Sugar and Spice, 2001
Sunset Blvd, 1949
Sunshine Cleaning, 2009
Super 8, 2011
Superbad, 2007
Supergirl, 1983
Surfer King The, 2006
Surrogates, 2009
Suspect Zero, 2004
Sweeney Todd The Demon Barber of Fleet Street, 2007
Sweet Hereafter The, 1997
Sweet Smell of Success, 1957
Swingers, 1994
Swordfish, 2001
Synecdoche New York, 2009
Syriana, 2005
Take Shelter, 2011
Taking Lives, 2004
Taking of Pelham One Two Three The, 1974
Taking Sides, 2003
Talented Mr. Ripley The, 1999
Tall in the Saddle, 1944
Tamara Drewe, 2010
Taxi Driver, 1976
Ted, 2012

Terminator, 1983
Terminator 2 Judgement Day, 1991
Terminator Salvation, 2009
The Rage Carrie 2, 1999
Theory of Everything The, 2014
There's Something About Mary, 1997
They, 2002
Thing The, 1981
Things My Father Never Taught Me The, 2012
Thirteen Days, 2000
This Boy's Life, 1992
This is 40, 2012
Thor, 2011
Thor Ragnarok, 2017
Three Kings, 1998
Three Kings (Spoils of War), 1995
Three Men and a Baby, 1986
Three Musketeers The, 1993
Thunderbirds, 2004
Thunderheart, 1992
Ticker, 2001
Timber Falls, 2007
Time Machine The, 2000
Tin Cup, 1995
Tin Men, 1986
Tinker Tailor Soldier Spy, 2011
Titanic, 1997
TMNT, 2007
To Sleep with Anger, 1989
Tombstone, 1993
Tomorrow Never Dies, 1997
Top Gun, 1985
Total Recall, 1990
Tourist The, 2010
Toy Story, 1995
Traffic, 2000
Training Day, 2001

Trainspotting, 1996
Transformers The Movie, 1986
Tremors, 1988
Tristan and Isolde, 2006
TRON, 1981
TRON Legacy, 2010
Tropic Thunder, 2008
True Grit, 2010
True Lies, 1994
True Romance, 1993
Truman Show The, 1998
Twilight, 2008
Twilight New Moon, 2009
Twin Peaks, 1992
Twins, 1998
Two For The Money, 2005
U Turn, 1997
Ugly Truth The, 2009
Un Singe en Hiver, 1962
Unbreakable, 1999
Under Fire, 1983
Unknown, 2011
Up, 2009
Up in the Air, 2009
Usual Suspects The, 1994
V for Vendetta, 2006
Valkyrie, 2008
Vanilla Sky, 2001
Verdict The, 1982
Very Bad Things, 1997
Village The, 2004
Virtuosity, 1994
Visitor The, 2008
Wag the Dog, 1996
Walk to Remember A, 2000
Walking Ta, 2004
Wall Street, 1987

Wall Street Money Never Sleeps, 2010
Wall E, 2008
Wanted, 2008
War for the Planet of the Apes, 2017
War Horse, 2011
War of the Worlds, 2005
Warm Springs, 2005
Warrior, 2011
Watchmen, 2009
Water for Elephants, 2011
Way Back The, 2011
We Own the Night, 2007
What Lies Beneath, 1999
When a Stranger Calls, 1979
While She Was Out, 2008
Whistleblower The, 2011
White Christmas, 1953
White Jazz, 2007
White Ribbon The, 2009
White Squall, 1994
Whiteout, 2009
Who's Your Daddy, 2005
Wild At Heart, 1990
Wild Bunch The, 1969
Wild Hogs, 2007
Wild Things, 1997
Wild Things Diamonds in the Rough, 2005
Wild Wild West, 1998
Willow, 1988
Win Win, 2011
Wind Chi, 2007
Withnail and I, 1987
Witness, 1985
Wizard of Oz The, 1939
Wolf of Wall Street The, 2013
Wonder Boys, 2000
Wonder Woman, 2017

Woodsman The, 2004
World is not Enough The, 1999
Wrestler The, 2009
X Files Fight the Future The, 1997
X Men, 1999
X Men Origins Wolverine, 2009
xXx, 2002
Year One, 2009
Yes Man, 2008
You Can Count On Me, 2000
You've Got Mai, 1998
Youth in Revolt, 2010
Zero Dark Thirty, 2013
Zerophilia, 2006
Zootopia, 2016

Appendix D

Ethics Checklists

D.1 User Study i & ii

Bournemouth University

Research Ethics Checklist

About Your Checklist	
Ethics ID	32500
Date Created	21/05/2020 09:58:05
Status	Open
Risk	Low

Researcher Details	
Name	Weilai Xu
Faculty	Faculty of Science & Technology
Status	Postgraduate Research (MRes, MPhil, PhD, DProf, EngD, EdD)
Course	Postgraduate Research - FST
Have you received funding to support this research project?	No

Project Details	
Title	User Evaluation of Computationally Generated Narrative Dialogues
Start Date of Project	15/06/2020
End Date of Project	30/09/2020
Proposed Start Date of Data Collection	15/06/2020
Supervisor	Fred Charles

Summary - no more than 500 words (including detail on background methodology, sample, outcomes, etc.)

The goal of this experiment is to evaluate the quality of the computationally generated dialogues based on the integration of narrative-based knowledge using deep learning technology. Emphasis is put on the evaluation of the impact of integrating a variety of personality features to the dialogue generation process.

Due to the ongoing COVID-19 pandemic, this experiment is conducted remotely. There is no in-person contact required.

The main method in this project is assigning participants several pieces of dialogue to read and answer questions regarding the dialogues in a provided online survey. A **survey sample** is attached to this ethics checklist. The sample provides examplar of a piece of dialogue and 8 questions. In the final version of the survey, there will be multiple sections of the same type, only including different dialogues for participants to review.

There is no specific target participants requirement apart from to be fluent in English. The survey is planned to be published until the end of the project, though each user is allowed to join the survey only once.

During the survey, all participants will be anonymous (no personal identifiable information will be collected). The results from the survey will be collected for the next stage of the analysis.

There is no obvious risk for participants to join this survey. All the results are anonymous without personal identifiable information collected.

Note: A short pilot study is included in this ethics submission with a small number of participants in order to make any potential changes before the study goes live.

Filter Question: Does your study involve Human Participants?

Participants	
Describe the number of participants and specify any inclusion/exclusion criteria to be used	
The link of survey is supposed to be distributed publicly and every well-educated person who is competent user in English could be target participant.	
Do your participants include minors (under 16)?	No
Are your participants considered adults who are competent to give consent but considered vulnerable?	No
Is a Disclosure and Barring Service (DBS) check required for the research activity?	No

Recruitment	
Please provide details on intended recruitment methods, include copies of any advertisements.	
Participants will be recruited within the students' population from Bournemouth University as well as through posting the survey weblink to selected online forums (related to the Interactive Narrative community). The survey weblink may also be distributed to other universities through our collaborators' network, if recruitment requires it.	
Do you need a Gatekeeper to access your participants?	No

Data Collection Activity	
Will the research involve questionnaire/online survey? If yes, don't forget to attach a copy of the questionnaire/survey or sample of questions.	Yes
How do you intend to distribute the questionnaire?	
online	
If online, do you intend to use a survey company to host and collect responses?	No
Will the research involve interviews? If Yes, don't forget to attach a copy of the interview questions or sample of questions	No
Will the research involve a focus group? If yes, don't forget to attach a copy of the focus group questions or sample of questions.	No
Will the research involve the collection of audio materials?	No
Will your research involve the collection of photographic materials?	No
Will your research involve the collection of video materials/film?	No
Will the study involve discussions of sensitive topics (e.g. sexual activity, drug use, criminal activity)?	No
Will any drugs, placebos or other substances (e.g. food substances, vitamins) be administered to the participants?	No
Will the study involve invasive, intrusive or potential harmful procedures of any kind?	No

Could your research induce psychological stress or anxiety, cause harm or have negative consequences for the participants or researchers (beyond the risks encountered in normal life)?	No
Will your research involve prolonged or repetitive testing?	No

Consent

Describe the process that you will be using to obtain valid consent for participation in the research activities. If consent is not to be obtained explain why.

The terms and conditions of the survey is included on the participant information sheet and clearly and explicitly states specific conditions on the welcome page (the first page) of the online survey. All statements from the agreement form will be presented on the welcome page in-place following the terms and conditions information sheet.

These are followed by a checkbox (with description “I consent to take part in the project on the basis set out above” or similar) that must be checked in order to take part in the survey. Every participant who is interested in taking part in the survey will only be allowed to start the survey if they have read the information and checked the checkbox denoting the participant confirms their agreement to take part in the survey.

The digital form of the agreement is to avoid in person contact due to the ongoing COVID-19 pandemic. This consent page links to the information sheet and lists the terms that would usually be found in the consent form. Both of these methods have been discussed with the BU Ethics team. Having consent provided digitally is not ideal and is a result of the restrictions due to the ongoing COVID-19 pandemic.

Do your participants include adults who lack/may lack capacity to give consent (at any point in the study)?	No
Will it be necessary for participants to take part in your study without their knowledge and consent?	No

Participant Withdrawal

At what point and how will it be possible for participants to exercise their rights to withdraw from the study?	Participants are free to withdraw at any stage of the survey by leaving the survey webpage. They are informed in the information sheet and are asked to confirm that they understand this in the consent form.
If a participant withdraws from the study, what will be done with their data?	If their survey data is complete, it will be saved automatically and be retained, as it cannot be linked to the participant in any way.

Participant Compensation

Will participants receive financial compensation (or course credits) for their participation?	No
Will financial or other inducements (other than reasonable expenses) be offered to participants?	No

Research Data

Will identifiable personal information be collected, i.e. at an individualised level in a form that identifies or could enable identification of the participant?	No
Will research outputs include any identifiable personal information i.e. data at an individualised level in a form which identifies or could enable identification of the individual?	No

Storage, Access and Disposal of Research Data

Where will your research data be stored and who will have access during and after the study has finished.

Survey responses, which are anonymous, are stored in a restricted MySQL database as part of the LimeSurvey software hosted on the LimeSurvey web server. These anonymous results will be exported to an encrypted hard drive. As soon as the study concludes, the database will be discarded.

Only the researcher will have access to the raw data. Anonymized data (survey responses) will be stored until the end of result analysis phase (the end of September 2020).

Once your project completes, will any anonymised research data be stored on BU's Online Research Data Repository "BORDaR"?	No
Please explain why you do not intend to deposit your research data on BORDaR? E.g. do you intend to deposit your research data in another data repository (discipline or funder specific)? If so, please provide details.	
N/A	

Dissemination Plans	
Will you inform participants of the results?	No
If Yes or No, please give details of how you will inform participants or justify if not doing so	

Final Review	
Are there any other ethical considerations relating to your project which have not been covered above?	No

Risk Assessment	
Have you undertaken an appropriate Risk Assessment?	Yes

Attached documents
agreement-form-submit.pdf - attached on 04/06/2020 16:03:39
information-sheet-submit.pdf - attached on 04/06/2020 16:03:46
survey question sample.pdf - attached on 12/06/2020 14:20:41

D.2 User Study iii

Bournemouth University

Research Ethics Checklist

About Your Checklist	
Ethics ID	39668
Date Created	04/10/2021 09:54:33
Status	Approved
Date Approved	29/11/2021 10:51:43
Date Submitted	15/11/2021 11:37:58
Risk	Low

Researcher Details	
Name	Weilai Xu
Faculty	Faculty of Science & Technology
Status	Postgraduate Research (MRes, MPhil, PhD, DProf, EngD, EdD)
Course	Postgraduate Research - FST
Have you received funding to support this research project?	No

Project Details	
Title	User Study of Computationally Generated Narrative Dialogues
Start Date of Project	22/11/2021
End Date of Project	28/02/2022
Proposed Start Date of Data Collection	22/11/2021
Original Supervisor	Fred Charles
Approver	Wen Tang

Summary - no more than 600 words (including detail on background methodology, sample, outcomes, etc.)

This is an additional user study successive to the last one deployed on Amazon Michanical Turk and completed by crowdworkers due the lockdowns caused by Covid-19 pandemic.

Since the university is reopen and students are returning, it is reviving the chance to conduct a user study on campus, either in-person or online, which allows us to recruit more reliable participants and test new dialogues generated by different methods.

The goal of this experiment is to evaluate the quality of the computationally generated dialogues based on the integration of narrative-based knowledge using deep learning technology. Emphasis is put on the evaluation of the impact of integrating a variety of personality features to the dialogue generation process.

The main method in this project is assigning participants several pieces of dialogue to read and answer questions regarding the dialogues on the survey. A survey sample is attached to this ethics checklist. The sample provides examplar of a piece of dialogue and 4 questions. In the final version of the survey, there will be multiple sections of the same type, only including different dialogues for

participants to review.

There is no specific target participants requirement apart from to be fluent in English. The survey is planned to be published until the end of the project, though each user is allowed to join the survey only once.

During the survey, all participants will be anonymous (no personal identifiable information will be collected). The results from the survey will be collected for the next stage of the analysis.

There is no obvious risk for participants to join this survey. All the results are anonymous without personal identifiable information collected.

Filter Question: Does your study involve Human Participants?

Participants	
Describe the number of participants and specify any inclusion/exclusion criteria to be used	
The link of survey is supposed to be distributed to students on campus in Bournemouth University and every well-educated student who is competent user in English could be target participant.	
Do your participants include minors (under 16)?	No
Are your participants considered adults who are competent to give consent but considered vulnerable?	No
Is a Disclosure and Barring Service (DBS) check required for the research activity?	No

Recruitment	
Please provide details on intended recruitment methods, include copies of any advertisements.	
Participants will be recruited within the students' population from Bournemouth University either by advertising in class or publishing participant wanted flyer sheets with survey weblink.	
Do you need a Gatekeeper to access your participants?	No

Data Collection Activity	
Will the research involve questionnaire/online survey? If yes, don't forget to attach a copy of the questionnaire/survey or sample of questions.	Yes
How do you intend to distribute the questionnaire?	
face to face,online	
If online, do you intend to use a survey company to host and collect responses?	No
Will the research involve interviews? If Yes, don't forget to attach a copy of the interview questions or sample of questions	No
Will the research involve a focus group? If yes, don't forget to attach a copy of the focus group questions or sample of questions.	No
Will the research involve the collection of audio materials?	No
Will your research involve the collection of photographic materials?	No
Will your research involve the collection of video materials/film?	No

Will the study involve discussions of sensitive topics (e.g. sexual activity, drug use, criminal activity)?	No
Will any drugs, placebos or other substances (e.g. food substances, vitamins) be administered to the participants?	No
Will the study involve invasive, intrusive or potential harmful procedures of any kind?	No
Could your research induce psychological stress or anxiety, cause harm or have negative consequences for the participants or researchers (beyond the risks encountered in normal life)?	No
Will your research involve prolonged or repetitive testing?	No

Consent

Describe the process that you will be using to obtain valid consent for participation in the research activities. If consent is not to be obtained explain why.

The terms and conditions of the survey is included on the participant information sheet and clearly and explicitly states specific conditions on the welcome page (the first page) of the online survey, as well as the first page of printed survey. All statements from the agreement form will be presented on the welcome page in-place following the terms and conditions information sheet.

For the online version of survey, these are followed by a checkbox (with description “I consent to take part in the project on the basis set out above” or similar) that must be checked in order to take part in the survey. Every participant who is interested in taking part in the survey will only be allowed to start the survey if they have read the information and checked the checkbox denoting the participant confirms their agreement to take part in the survey.

The digital form of the agreement is to avoid in person contact due to the COVID-19 pandemic. This consent page links to the information sheet and lists the terms that would usually be found in the consent form. Both of these methods have been discussed with the BU Ethics team.

For the printed version of survey, participants will be asked to sign their name and date on the first page of the survey before they continue.

The sample of agreement form for both online version and printed version, as well as participant information sheet are attached.

Do your participants include adults who lack/may lack capacity to give consent (at any point in the study)?	No
Will it be necessary for participants to take part in your study without their knowledge and consent?	No

Participant Withdrawal

At what point and how will it be possible for participants to exercise their rights to withdraw from the study?

Participants are free to withdraw at any stage of the survey without giving a reason by leaving the survey. They are informed in the information sheet and are asked to confirm that they understand this in the consent form.

If a participant withdraws from the study, what will be done with their data?

If their survey data is complete, it will be saved automatically and be retained, as it cannot be linked to the participant in any way.

Participant Compensation

Will participants receive financial compensation (or course credits) for their participation?	Yes

Please provide details

Every participant who complete the whole survey will be compensated by a digital 5-pound coupon or gift card (e.g. Amazon).

Will financial or other inducements (other than reasonable expenses) be offered to participants?	No

If participants choose to withdraw, how will you deal with compensation?

after assessed the possibility of withdrawal, we believe withdrawal is not likely to happen, because neither information of personal identity nor privacy will be left to us for analysing. Therefore, if one participant choose to withdraw after completion, he/she can keep the compensation. If they choose to withdraw in the middle of survey, they cannot receive compensation.

Research Data

Will identifiable personal information be collected, i.e. at an individualised level in a form that identifies or could enable identification of the participant?	No
Will research outputs include any identifiable personal information i.e. data at an individualised level in a form which identifies or could enable identification of the individual?	No

Storage, Access and Disposal of Research Data

Where will your research data be stored and who will have access during and after the study has finished.

For the online version of survey, survey responses, which are anonymous, are stored in a restricted MySQL database as part of the LimeSurvey software hosted on the LimeSurvey web server. These anonymous results will be exported to an encrypted hard drive. As soon as the study concludes, the database will be discarded.

For the printed version of survey, all the responses are initially on paper with corresponding participants' agreement forms and then scanned to digital format, which will be stored and analysed after completing user study.

Only the researcher will have access to the raw data.

Once your project completes, will any anonymised research data be stored on BU's Online Research Data Repository "BORDaR"?	No

Please explain why you do not intend to deposit your research data on BORDaR? E.g. do you intend to deposit your research data in another data repository (discipline or funder specific)? If so, please provide details.

N/A

Dissemination Plans

How do you intend to report and disseminate the results of the study?

Peer reviewed journals,Conference presentation

Will you inform participants of the results?	No

If Yes or No, please give details of how you will inform participants or justify if not doing so

All the questions of survey are not related to personal information.

Final Review

Are there any other ethical considerations relating to your project which have not been covered above?	No

Risk Assessment

Have you undertaken an appropriate Risk Assessment?	Yes

Attached documents
information-sheet-2021.pdf - attached on 04/10/2021 15:17:33
agreement-form-2021.pdf - attached on 04/10/2021 15:17:39
survey_example.pdf - attached on 05/10/2021 12:20:26
survey_sample_online.pdf - attached on 08/11/2021 17:33:36