

# Personality Profiling for Literary Character Dialogue Agents with Human Level Attributes

Nicolay Rusnachenko<sup>\*1</sup>[0000-0002-9750-5499], Huizhi Liang<sup>1</sup>[0000-0003-4408-4528]

1. School of Computing, Newcastle University, Newcastle upon Tyne,  
rusnicolay@gmail.com, huizhi.liang@newcastle.ac.uk

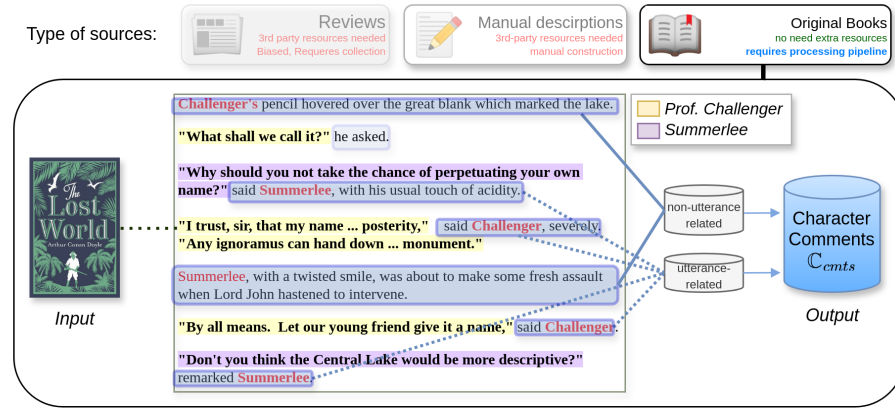
**Abstract.** Equipping personalities to dialogue agents can help to better engage end-users. However, how to profile personality remains an open research question due to the difficulties of obtaining real human data. As classic literary characters often encapsulate typical human personality traits, literature books has been used as a high quality data source to construct personality profiles for dialogue agents. Existing work mainly focuses on using external reviews and human experts' annotations to profile character personalities. The in-text comments about the personality of characters in a literature book itself have been ignored. In this paper, we propose a new NLP task called *character comments annotation* to annotate the in-text comments about the personality of characters including dialogue utterances and surrounding text, paragraphs mentioning a character. We constructed new personality annotated dialogue datasets based on Gutenberg literature book project. We propose a workflow to automatically profile literary characters from literature novel books. Two personality profiling models have been proposed, including (i) psychological personality traits vocabulary-based spectrum (SPECTRUMS) approach and (ii) a *tf-idf* based words selection as a baseline approach. We applied the proposed personality models in dialogue response prediction tasks with ranking-based and generative dialogue agents. The results show that the fine-tuned dialogue agents with SPECTRUMS profiles surpass those trained without them by 2.5% (Hits 1@20) for ranking-based, and by 8% (Rouge-1) for generative agents. The implementation of the workflow with study-related resources is publicly available: <https://github.com/nicolay-r/book-persona-retriever>

## 1 Introduction

Dialogue agents are artificial intelligence systems designed to interact with humans in a natural and conversational manner. *Human level attributes* (HLA) are characteristics, traits, or abilities that are typically associated with or possessed by humans [1]. *Personalities* are important HLA that help dialogue agents to be more human-like or near-human, for example, maintaining a consistent style during multiple rounds of communication [1].

---

\* The work performed while at Newcastle University



**Fig. 1.** *top:* type of data sources utilized for fictional character profiling: (left to right) reviews, original books, manual descriptions; *bottom:* extraction of text parts from original book that convey character related information (*Character Comments*) in conversations (between SUMMERLEE and CHALLENGER from “*The Lost World*” by Arthur Conan Doyle)

Dialogue agents equipped with personalities can be served as virtual assistants in various application areas of our everyday life, including: customer service, healthcare, education [8], and entertainment [15]. Within the recent years, agents with personality profiles demonstrate performance improvements in (i) dialogue acts, (ii) repetition reductions, (iii) overall conversation consistency [8, 13]. Most of the background studies are limited in details about characters of conversation data and hence requires external sources for characters profiling, including: reviews [22], human-annotated datasets [8]. The problems of using external resources are: (i) expensive for human annotation and review, (ii) resources usually have quality controlling difficulties, (iii) may lack review data in some characters or books. However, literature novel books, in which texts are saturated with information about characters and interactions between them, visual descriptions, lifestyles, etc.(see Figure 1). Being presented in less accessible way, literature novel books itself may be counted as a good source of character personalities.

In this paper we bridge the gap in deeper understanding of fictional characters of literature novel books by solely based on their in-text information. From observation of many literature books [17, 16], we found that in-text *comments* is a common phenomenon. We refer to *comment* as a text part which: (i) precedes or follows the character utterance, (ii) isolated paragraph of texts with character mentions (see Figure 1). The collection of annotated comments, i.e. comments for which speaker is defined or belongs to, may serve as a source for personality profiles construction. The contribution of this work is four fold:

- We provide methodology for extraction character-related contexts from literature novel books by proposing a new task called *character comments annotation*;
- To profile characters using only raw literature novels, we propose: (i) a workflow for automatic character comments annotation, and (ii) the application of a model (SPECTRUMS) based on an adjective-pairs vocabulary to extract character personality profiles from these annotated comments;
- Due to the absence of manually annotated character-related information in literature novel books domain, we construct resource (LDC) which gives an ability dialogue assistant agents to learn character language styles through their profiled personalities.
- We evaluated the effectiveness of the constructed character profiles of LDC in character response prediction task; according to our experiments, the use of agents with imputed personalities improves those without them by 2.5% (Hits 1@20) for ranking-based and by 8% (Rouge-1) for generative agents.

## 2 Related work

In the context of modeling personalities, the research works are branched out into diverse methodologies, each focusing on distinctive aspects of character representation. Labatut et al. [10] offer a survey on modeling characters and proposes a generalized framework aimed at completion of *network*, that involves character representations. This acts as a conceptual roadmap that guides future research on book processing and character modeling. Inoue et al. [20] employed *fixed-length vector* (embedding) model for character representation. To construct the related models, the authors rely on meta-information of quizzes collected from external resources. From the open-domain chatbots domain, the PERSONA-CHAT [8] represents dataset which collects dialogues on personal topics, equipped by persona traits descriptions. Carlsson et al. [16] propose a *question-answering methodology* for understanding character personalities by machine learning model. In their approach, character descriptions are manually composed following specific annotator guidelines. Junzhe et. al [22] approached the problem of constructing the personality of the literary characters by *automatically distilling character traits* from commentary articles. The authors select the 10 most frequent traits, towards which each character becomes mapped at various scales. *Openpsychometrics* studies<sup>1</sup> serves as lexicon of *antonym adjective-pairs* for manually annotated characters (amount of 800) by movie experts. To annotate characters with respect to the given adjective-pairs set, experts were attended in quizzes aimed at choosing particular adjective in pairs within a certain moment in movie.

Towards the nature of the agents, such systems might be demarcated into two major classes of the produced response: (i) *ranking*, and (ii) *generative* [18, 8]. The appearance of the *self-attention* mechanism with the related system referred as *transformer* [6], plays the crucial role in dialogue agent capabilities and advances [19, 14, 11]. The transformer architecture become commonly distributed

<sup>1</sup> <https://openpsychometrics.org/tests/characters/>

in development of (i) *generalized* [12, 21] and (ii) *target-oriented* dialogue agent types. For the both type of systems, the imputing personalities performed in form of prompts [8, 13, 9, 22] with personalities mention in textual format. Towards movie and literature domains, it is common for their authors to experiment with the transformer’s: (i) *decoder part* [9, 17], (ii) various *encoder-based implementations* of original transformers [15, 9]. Li et. al [15] proposes using *factorization technique* to distinct characters and utilize it for choosing counteractive characters in the machine learning model training process, using reviews and movie subtitles of fictional characters.

However, most of the studies supporting the reviewed methodologies heavily rely on external resources for profiling characters, such as meta-information [20], commentary articles / reviews [22, 15], or manually annotated collections [2, 9].

### 3 Literary Character Profile Construction

We propose two approaches for constructing *character personality profiles*:

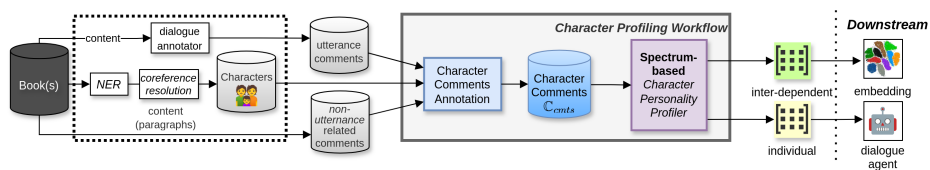
- Direct use of TF-IDF approach or embeddings to select popular words as the personality profile. This approach might result in non-personality words;
- Use personality lexicon to select representative words. To avoid selecting antonyms such as **rich** and **poor** to profile the same person, we propose an approach that relies on lexicon of adjective antonym pairs and calculates membership of these adjective pairs (Section 3.2). This approach is expected to result in non-overlapping mention of different personality traits.

Figure 2 illustrates the application of character comment annotation for personality profiling in the form of a workflow. **The input** is the set of raw books, for which the external components aimed at extraction of: (i) *utterance-related comments* and (ii) *non-utterance related comments*. In subsequent, some of these comments become a part of *annotated character comments* ( $C_{\text{cmnts}}$ ) which consists of comments annotated with the related character (Section 3.1). **The output** of the workflow consists of two types of constructed personality profiles (Section 3.2): (i) character individual and (ii) character interdependent [15]. The interdependent profiles play an important role in learning the distributed representation of characters by dialogue agents [20].

#### 3.1 Characters Comments Annotation

To annotate *comments* with speaker characters, we do: (i) construct dictionary of most frequent text parts that precede character mentions ( $D$ ) to select *utterance-related comments*, and (ii) select such *non-utterance comments* that has mention(-s) of a single character ( $c$ ). Here,  $D$  represent the most frequent text parts that go before the first mention of character in utterance-related comments (for example: **said**, **remarked** in Figure 1). We treat such parts as sequence of words  $[w_0, \dots, w_k]$ , limited by  $k$ , excluding the mention of  $c$ .

For each *utterance-related comments* of each book, we follow three steps for completing dictionary  $D$ :



**Fig. 2.** Character profiling workflow aimed at profiling characters from literature novel books by solely rely on *in-text comments*: (1) dialogues and their surrounding texts, (2) paragraphs from the books; *input*: collection of the raw books *output* represent matrices of: (i) character *individual* personality profiles and (ii) *interdependent* character profiles (embeddings); the preliminary book processing components such as characters extraction [4], dialogues annotation are bounded in dotted line box

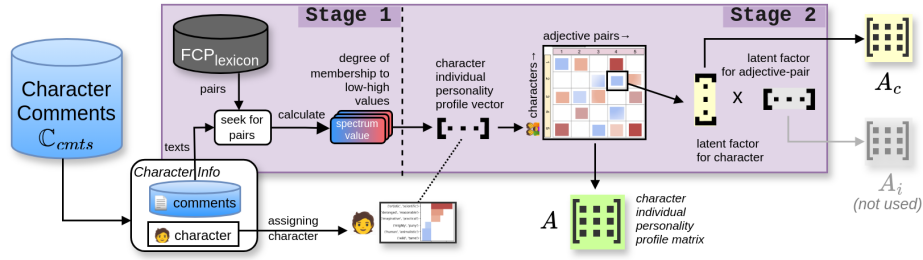
- Step 1: *Select comments that mention characters.* We rely on the assumption of relatively short entries, limited by  $k$  words.
- Step 2: *Measure the importance of the most representative terms* in selected comments. We follow the assumption that authors of the different books has the similar terms in commentary the related speaker; to provide the related assessments, the *tf-idf* measure was chosen;
- Step 3: *Select relevant text parts.* We use threshold ( $\gamma$ ) for minimum *tf-idf* values of text parts to select relevant entries for  $D$ .

Finally we select comments that mention characters and annotate them with speaker as follows. For the utterance-related comments, we select those in which text part that precedes the mention of the first appeared character presented in  $D$ . Given the selected utterance-related comment, we annotate the related utterance with the same (first appeared) speaker appeared in comment. Regarding non-utterance related comments, we select only those that mention a single character<sup>2</sup>, and annotate them accordingly with that character.

### 3.2 Spectrum-based Character Personality Profiling

In this paper, we propose the adaptation of character personality profiling model from the adjective-pair lexicon ( $FCP_{\text{lexicon}}$ ) of manually annotated fictional movie characters [23]. The  $FCP_{\text{lexicon}}$  yields 264 adjective pairs that represent antonyms and act as polarities of a single dimension in character profile. We refer to each polarity of the pair as *low* and *high*. Examples of the adjective pairs that are part of the related lexicon of character polarities are: **tiresome/interesting**, **smooth/rough**, **imaginative/practical**, etc. For any adjective-pair in particular, the degree of its membership to the given character dubbed as character *spectrum value*. We propose SPECTRUMS profiling model, aimed at composing character personalities in a form of spectrum values from their information.

<sup>2</sup> The case of multiple occurrences of the same literature character in different name variations is allowed



**Fig. 3.** The two-staged SPECTRUMS profiling workflow aimed at construction of personality profile matrices of two types: (i) character individual personality profiles ( $A$ ) (ii) character interdependent personality profiles ( $A_c$ )

Figure 3 illustrates a two-stage workflow application for completing character personality profiles.

**The first stage** aimed at composing the spectrum values individually per every single character. For the particular character  $c \in \mathbb{C}$  from the literature novel book and adjective pair  $p_j \in FCP_{\text{lexicon}}$  (indexed as  $j$ ), we rely on its  $\mathbb{C}_{\text{cmnts}}(c)$  for calculating the result spectrum value for  $p_j$  (see Figure 2). This stage involves the following steps:

- Step 1: Represent every comment from the character info  $\mathbb{C}_{\text{cmnts}}(c)$  as a sequence of unigrams  $W = [w_0, \dots, w_{|W|}]$ ;
- Step 2: Obtain the related number of  $p_j$  entries, separately for *low* ( $\hat{c}_j^{\text{low}}$ ) and *high* ( $\hat{c}_j^{\text{high}}$ ) polarities across all the unigrams  $w \in W$ ;
- Step 3: Calculate  $p_j$  spectrum value ( $\hat{c}_j \in [-1, 1]$ ) as a normalized *degree of its membership* to *low/high* polarities using the following formula:

$$\hat{c}_j = (\hat{c}_j^{\text{high}} - \hat{c}_j^{\text{low}}) / \max(\hat{c}_j^{\text{low}}, \hat{c}_j^{\text{high}})$$

Applying steps 1-3 towards the all adjective pairs for  $FCP_{\text{lexicon}}$  and all  $c \in \mathbb{C}$ , where each  $c$  represents an individual vector (character personality profile)  $\mathbf{c} = [\hat{c}_1 \dots \hat{c}_m]$ , where  $\mathbf{c} \in \mathbb{R}^m$  and  $m = |FCP_{\text{lexicon}}|$  denotes the lexicon size. The complete set of  $\{\mathbf{c}_i\}_{i=1}^n$  could be formed into a matrix  $A^{n \times m}$ , where  $n = |\mathbb{C}|$  is the total number of characters (see Figure 3).

However, such a representation limits our comparison between characters by remaining at the level of each spectrum individually. To establish a deeper connection between character personality profiles, **in the second stage** we calculate the interdependencies between personalities and rank the similarity between characters [2]. We use *latent factors* [2] to transform both characters and personality profiles into the same latent space to make them directly comparable. The modeling of features related to matrix  $A$  could be treated as factoring matrix  $A$  into: (i)  $A_c$  (latent factors for characters) and (ii)  $A_i$  (latent factors for adjective-pairs). We refer to Conjugate Gradient Method [3] to bring  $A_c \cdot A_i$  as close as possible to  $A$  [15].

## 4 Evaluation Tasks

The constructed personality-based character profiles have a numerous applications in recommendation, search, and Question-Answering (Q&A) systems. In this paper, we evaluate the effect of inputting character profiles in character dialogues based Q&A systems. In particular, the following tasks are considered:

- **Dialogue response prediction.** This is a prediction task. Given (i) personality traits ( $p_c$ ) of the character speaker ( $c$ ), and (ii) query utterance ( $u$ ) addressed to  $c$ , this task is to predict the most expected response  $u'$  by  $c$  from a set of the predefined *answer candidates*  $\mathbb{C} = \{u'_1 \dots u'_q\}$ , where  $u' \in \mathbb{C}$ , and  $q$  is the total number of candidates.
- **Dialogue response generation:** This is a generation task. Given (i) personality traits ( $p_c$ ) of the character speaker ( $c$ ), and (ii) query utterance ( $u$ ), this task is to *generate the most expected response  $u'$* .

## 5 The construction of Datasets

We collect 15,332 books from Project Gutenberg<sup>3</sup> that provides publicly available free books. We first construct a Literature Dialogue Collection (LDC). Secondly, we use annotated utterances as well as paragraphs with single characters to construct a Literature Character Personality Profile Dataset (LPPD) for (LPPD<sub>SPECTRUMS</sub>) and (LPPD<sub>TF-IDF</sub>). Thirdly, we combine LDC and LPPD datasets to construct dialogue datasets of two types: (1) LDC<sub>SPECTRUMS</sub> / LDC<sub>TF-IDF</sub> and (2) LDC<sub>NO-PT</sub> to facilitate training dialogue agents with and without personality profiles.

**Dataset 1: Literature Dialogue Collection (LDC).** To extract the characters and their dialogues from literature books, we conducted the following processing:

- (a) **Characters or Speaker Extraction.** To extract character speaker names, we adopt Named Entity Recognition (NER) model from Stanford Core NLP, keeping PERSON tag in final annotation. To cope with naming variations of the same characters, the rule-based approach [4] was adopted to conduct clustering for name variations.
- (b) **Dialogues extraction.** We follow the dialogue annotation algorithm proposed in [17] to extract utterances. Dialogue is defined as a sequence of utterances, placed in the order of their appearance in text. The default parameter settings [17] for English texts were used.
- (c) **Utterance comments annotation.** This step is to annotate the *comments* of a character. We compose a dictionary  $D$  with  $k = 3$  and  $\gamma = 0.01$  (see Section 3.1). We conducted statistics about the position of when authors mention character names. In our dataset, the majority (i.e., 95%) of mention happens before the first three words of comments, 29.33% cases mention in the beginning of the comment, 52% cases mention at the second word, 10% cases mention at the third word, and less than 5% mention everywhere else.

<sup>3</sup> <https://www.gutenberg.org/>

Raw Text	
Books (#)	15332
Speakers (#)	741327
Characters per book (average)	48.35
Books (#)	15332
Dialogues	
Dialogues extracted [17]	1767640
Utterances per dialog (average)	7.54
Prefix dictionary application, annotated utterances (Section 3.1) (%)	29.33%
Literature Dialogue Collection (LDC)	
Average amount of utterances per book (#)	≈634
Annotated query-response pairs (#)	9719917

**Table 1.** Parameter-value statistics of the composed Literature Dialogue Collection (LDC), based on books from Project-Gutenberg collection with analysis on every level: raw texts, dialogues and pair-based query-response pairs

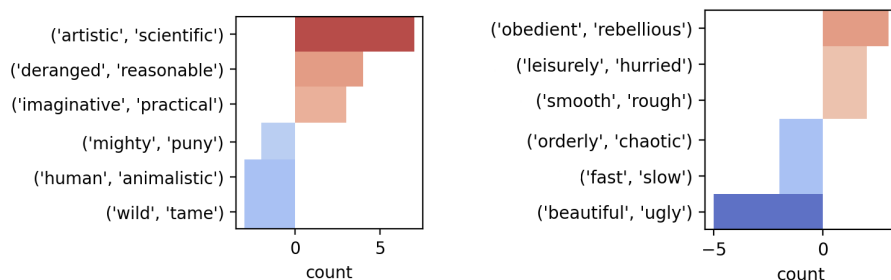
- (d) **Query-response pairs construction.** The result of this step is a union of extracted pairs from dialogues. We treat each dialogue  $D = [u_1, \dots, u_{d'}]$  as a sequence of pairs  $[p_1, \dots, p_{d'-1}]$ , where  $p_i = \langle u_i, u_{i+1} \rangle$  for  $i \in \overline{1..d'-1}$ , where  $d'$  is the total amount of utterances in  $D$ . To make sure that each pair has a known responding speaker, we select those pairs  $p_i$  in which the speaker for  $u_{i+1}$  is defined.

Table 1 shows the statistic of the processed raw texts, extracted dialogues, and annotated query-response pairs of the LDC dataset.

**Dataset 2: Literature Character Personality Profile Dataset (LPPD).**

We adopt the first stage of the proposed personality profile workflow (Section 3) on LDC to compose this dataset. We adopt TF-IDF and SPECTRUMS (see Figure 3) approaches to compose  $LPPD_{TF-IDF}$  and  $LPPD_{SPECTRUMS}$  respectively. The application of TF-IDF is as follows. For the particular character ( $c$ ) and its comments  $\mathbb{C}_{cmnts}(c)$ , we calculate the *tf-idf* measure of each unigram of comments. We treat each comment of  $\mathbb{C}_{cmnts}(c)$  as a separate document in *tf-idf* notation. Figure 4 illustrates personality profiles SPECTRUMS-based for two characters from  $LPPD_{SPECTRUMS}$ . We can see that SUMMERLEE’s most representative personality spectrum is *artistic/scientific*. The pole of *scientific* has higher value which shows that SUMMERLEE is more scientific than artistic. For MAC WILLIAMS, the most representative personality spectrum is *beautiful/ugly* and the pole of *beautiful* has higher values.

**Dataset 3: Literature Dialogue Response (LDR).** We use annotated query-response pairs from LDC to compose this dataset (LDR) for the evaluation tasks proposed in Section 4. To facilitate the training of dialogue agents with and without personality profiles, we format LDR by additionally using: (i) personality profile placeholders, and (ii) personality profile TF-IDF/SPECTRUMS entries from the related LPPD datasets. To prevent dialogue agents from overfitting on the particular names and their mentions, we follow similar studies [15] to: (1) mask character names, (2) limit the amount of profiles per character to 8.



**Fig. 4.** LPPD<sub>SPECTRUMS</sub> personality profile examples for: SUMMERLEE from “The Lost World” by Arthur Conan Doyle (*left*) and MAC WILLIAMS from “Soldiers of Fortune” by Richard Harding Davis (*right*)

**Table 2.** Input data representation for query sentence addressed to the SUMMERLEE (Response) from The Lost World by Arthur Conan Doyle, separately for the two type of the composed data: without personality profiles (left), and with mentioning personality profiles from LPPD<sub>SPECTRUMS</sub> (right); associative connection between personality traits in input with the response is colored in blue

Input formatting example	
LDR <sub>NO-PT</sub> (No personality profile)	LDR <sub>SPECTRUMS</sub> (Profiles from LPPD <sub>SPECTRUMS</sub> )
persona: none	persona: I am scientific
persona: none	persona: I am reasonable
persona: none	persona: I am human
...	...
Query sentence	
Well, our rope is still more than a hundred feet long, Surely we could get down.	Well, our rope is still more than a hundred feet long, Surely we could get down.
Response (SUMMERLEE)	
How about the Indians in the cave?	How about the Indians in the cave?

Table 2 illustrates an example of the formatting of query-response pairs from LDR<sub>NO-PT</sub> and LDR<sub>SPECTRUMS</sub>. We use `none` as a placeholder for the personality profile (LDR<sub>NO-PT</sub>). In turn, LDR<sub>SPECTRUMS</sub> includes a list of personality attributes (`scientific`, `reasonable`, `practical`) from LPPD<sub>SPECTRUMS</sub> for the responding character (SUMMERLEE) before the query sentence.

## 6 Experiments

We conducted both empirical and human evaluation experiments on LDR. In particular, the aim of the experiments is to show the difference between: (i) agents that employ personality profiles and (ii) without personality profiles.

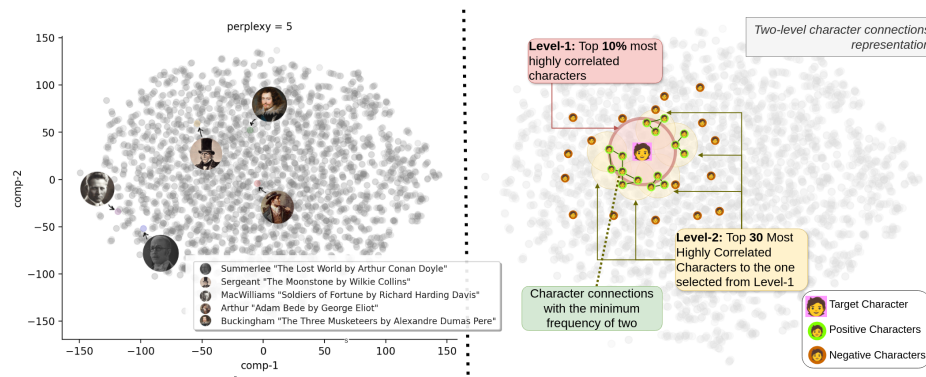
## 6.1 Experiment Setup

To avoid mentions of rare characters in LDC, we followed the practice in similar studies [15] and selected characters that appear in dialogues at least 100 times. To deal with relatively short utterances, we set a threshold for the maximum length of utterances in LDC, which should not exceed 100 words [17]. With these selection criteria, we reduce the total number of characters from LDC to 400 and carry out evaluation experiments on new sub dataset dubbed as LDR<sub>400</sub>. The LDR<sub>400</sub> represent a 0.4% of LDC content, with  $\approx 65$  words per utterance across all dialogue pairs.

**Training and validation sets.** We split LDR<sub>400</sub> into five folds and applied 5-fold cross-validation to measure the performances of the dialogue agents. Each split has 80 speakers across 75-80 books, where each fold on average has: 7200 dialogues, and  $\approx 92$  dialogues per character.

**Test set.** Represent five characters that non presented in LDR<sub>400</sub> and most-frequently appeared in dialogues. These are the following characters: (1) MR. SUMMERLEE (The Lost World by Conan Doyle), (2) SERGANT CUFF (The Moonstone by Wilkie Collins), (3) MR. MAC WILLIAMS (Soldiers of Fortune by Richard Harding Davis), (4) ARTHUR DONNITHORNE (Adam Bede by George Eliot), and (5) BUCKINGHAM (Tree Musketeers by Alexandre Dumas Pere). The utterances of these characters are used as the ground-truth.

Figure 5 (left) illustrates an example of the t-SNE visualization of all the characters of LDC that have the mention of at least 5 personalities in LPPD<sub>SPECTRUMS</sub>. The characters of the TEST set are explicitly plotted in this figure.



**Fig. 5.** *left*: t-SNE visualization of characters from LDC that have the mention of at least 5 personalities in LPPD<sub>SPECTRUMS</sub>, along side with the characters from the TEST set; *right*: two-level character connection representation for negative speaker set annotation

**Fine-tuning policy.** For all the models we use LDR<sub>400</sub> training part for supervised fine-tuning (SFT). We limit SFT by 5 epochs with the result assess-

ment on validation set every 0.5 epoch with keeping the best state across all model assessment attempts.

To evaluate the effectiveness of personality profiles, we fine-tune and assess the models in two stages. On stage#1 (“†”) we fine-tune agents without personality traits (Table 2, left). On stage #2 (“‡”), we obtained model checkpoints from the stage#1 and fine-tune them again on the dataset with the particular personality profiles (Table 2, right).

## 6.2 Results

We utilize the ParlAI [5] to experiment with dialogue agents. In particular, we evaluate *ranking based dialogue agents* in the dialogue response prediction task and *generative dialogue agents* in dialogue response generation task.

**Ranking based dialogue agents.** We used transformer-based dialogue agents and IR based models to rank answer candidates. For transformer-based dialogue agents, we adopt poly-encoder transformer implementation ( $T_{\text{enc-POLY}}$ ) [9] that pre-trained on external resources including: Reddit platform posts [7] and PERSONA-CHAT [13]. To fine-tune  $T_{\text{enc-POLY}}$ , we follow the parameters setup recommended in [9]. Despite the various IR implementations, we adopt this simple approach<sup>4</sup>: seeking for the most similar candidate and output the response from the related exchange. With this approach, we use the cosine similarity between the input query and candidate responses based on *tf-idf* measure.

Ranking-based approaches establish relevance between given query and set of *candidates*. Therefore, in addition to complement ground truth response with the other non-relevant responses. In the case where personality traits are not mentioned in LDR (NO-PT), we use *uniform* utterances selection across utterances from books of the same set. In the case of mentioned personality profiles (SPECTRUMS /TF-IDF), we adopt the *two-level character clustering approach* [15] (Figure 5, right) to define the most distant characters from the target ( $c_t$ ) character. For the first level, we set the most closest 10% characters to  $c_t$ . For the particular character  $c'_t$  of the second level, we define the 30 most strongly correlated to  $c'_t$  characters. Those characters belong to the mentioned two levels treated as *positive characters* to  $c_t$  (*negative* otherwise). As a result, we adopt uniform utterance selection across utterances from *negative* character set.

To measure the distance between characters, we refer to the calculation of cosine similarity. To compose vectors of the TF-IDF profiling models, we first compose the united set of words, followed by the application of the bag-of-word model. Following the setting of studies [15], we set  $q = 20$  candidates for each input query. We over-sample the original training data of LDR<sub>400</sub> by 5 times the amount of the originally available examples in the *train* by each time choosing a different candidate in the list.

Table 3 illustrates the results obtained by ranking models. The following evaluation metrics were used: *precision at k* among the total amount of candidates ( $q$ ) denoted as  $k@q$ , F1-score for uni-grams, *precision* (P) and *recall* (R),

<sup>4</sup> [https://parl.ai/docs/agent\\_refs/ir\\_baseline.html](https://parl.ai/docs/agent_refs/ir_baseline.html)

**Table 3.** 5-fold cross-validation results of the ranking-based models on *validation* part of the LDR<sub>400</sub> dataset; “†” denote models fine-tuned on LDR<sub>400</sub> (NO-PT), and “‡” denote †-models fine-tuned on LDR<sub>400</sub> formatted with SPECTRUMS / TF-IDF profiles

Model	LDR <sub>400</sub> Format	@1/20	@5/20	@10/20	F <sub>1</sub> (P,R)	Prec	Recall	BLEU-4
T <sub>enc</sub> -POLY‡	SPECTRUMS	<b>0.516</b>	<b>0.829</b>	<b>0.929</b>	<b>0.574</b>	<b>0.588</b>	<b>0.578</b>	<b>0.516</b>
IR‡	SPECTRUMS	0.168	0.442	0.657	0.298	0.297	0.330	0.169
T <sub>enc</sub> -POLY†	TF-IDF	0.504	0.286	0.928	0.567	0.572	0.560	0.505
IR†	TF-IDF	0.161	0.425	0.637	0.295	0.280	0.319	0.161
T <sub>enc</sub> -POLY†	NO-PT	0.503	0.821	0.923	0.563	0.577	0.566	0.503
IR†	NO-PT	0.159	0.438	0.657	0.293	0.291	0.324	0.159

**Table 4.** 5-fold cross-validation results of the generative models (GPT-2<sub>small</sub>) on *validation* set of LDR<sub>400</sub> dataset; “†” denote models fine-tuned on LDR<sub>400</sub> (NO-PT), and “‡” denote †-models fine-tuned on LDR<sub>400</sub> formatted with SPECTRUMS / TF-IDF profiles

Model	LDR <sub>400</sub> Format	R-1	R-2	R-L	PPL	F <sub>1</sub> (P,R)	Prec	Recall	BLEU-1
GPT-2 <sub>small</sub> ‡	SPECTRUMS	<b>0.018</b>	<b>0.006</b>	<b>0.002</b>	<b>25.070</b>	<b>0.139</b>	<b>0.205</b>	<b>0.120</b>	<b>0.089</b>
GPT-2 <sub>small</sub> ‡	TF-IDF	0.017	0.005	0.002	25.067	0.129	0.178	0.101	0.087
GPT-2 <sub>small</sub> †	NO-PT	0.012	0.003	0.001	26.100	0.098	0.129	0.114	0.065
No fine-tuning									
GPT-2 <sub>small</sub>	—	0.003	0.001	0.000	54.650	0.027	0.016	0.120	0.016

and BLEU-4. Imputing personality traits in models fine-tuning process (‡) shows better performed models than the ones fine-tuned without personality traits (†) across both spectrum and tf-idf based personality profiling approaches. In particular, the application of SPECTRUMS based profiles results in 5% increment by @1/20 over the models fine-tuned with TF-IDF based character profiles. Analyzing the results of SPECTRUMS-based models, the T<sub>enc</sub>-POLY‡ results in ≈2% improvement, and IR‡ shows ≈6% performance increment by @1/20 over their fine-tuned versions without personality traits (T<sub>enc</sub>-POLY† and IR† respectively). Overall, the spectrum based personality profiling approaches performed the best.

**Generative dialogue agents.** We used GPT-2<sub>small</sub> (124M) as the generative dialogue agent. To measure model performance, we adopt: BLEU-1, ROUGE-scores, F<sub>1</sub>(P,R) based on unigram overlapping, and *perplexity of fixed-length models* (PPL) [1].

The results are shown in Table 4. The comparison between generative models and ranking-based models shows that generative models generally do not perform as well as ranking-based one. The overall results indicate that the outputs of the generative models tend to deviate from the expected utterances in the response. However, from the Table 4 it is possible to observe that personality traits can enhance the performance of dialogue agents. In particular, GPT-2<sub>small</sub>‡

**Table 5.** Results of the human evaluation (@1/20) on *test* characters set along side with the  $T_{\text{enc-POLY}}\ddagger$  (SPECTRUMS) model application

Character	Human	$T_{\text{enc-POLY}}$ (SPECTRUMS)						
	@1/20	@1/20	@5/20	@10/20	$F_1$ (P,R)	Prec	Recall	BLEU-4
SUMMERLEE	0.40	0.41	0.75	0.88	0.47	0.47	0.49	0.31
SERGANT CUFF	0.35	0.33	0.68	0.88	0.38	0.39	0.40	0.28
MAC WILLIAMS	0.45	0.29	0.65	0.82	0.36	0.37	0.38	0.29
ARTHUR DONNITHORNE	0.50	0.57	0.81	0.97	0.61	0.61	0.62	0.57
BUCKINGHAM	0.50	0.50	0.83	0.94	0.55	0.54	0.58	0.36

(SPECTRUMS) shows 8% performance increment (BLEU-1) comparing with GPT-2<sub>small</sub><sup>†</sup> without personality profiles.

### 6.3 Human Evaluation

We perform human evaluation of ranking-based dialogue agents without personality traits on the TEST set. We recruited 4 participants who were researchers aged 30 – 40 at Newcastle University. To control bias, each participant evaluated one or two characters. Follow the same setting with automatic ranking-based evaluation in Section 6.2, for each character, we randomly select 10 testing samples (each includes an initial line of query utterance along with 20 candidate responses, one of which is the ground truth). For human evaluation, we use responses from characters of the TEST set (Section 6.1).

These 200 samples make up a single questionnaire presented in full to each participant evaluating the corresponding character, and the participant is asked to select the single top response they think the character would most likely respond with for each of the ten initial dialogue lines. We mask any character names within the candidate responses to prevent human participants from using names to identify which book the responses are from. The results of the human evaluation in comparison with the proposed approach with personality traits are shown in Table 5. We can see that the ranking based dialogue agents with personality profile outperformed human evaluation on the questionnaire with no personality profiles. This further demonstrate the effectiveness of our constructed personality based profiles.

## 7 Conclusions

In this paper, we explore the use of literature character personality profiling models, which rely exclusively on book content, to enhance dialogue agents’ understanding of characters. To profile a large number of characters, the application of a vocabulary-based character personality profiling model (SPECTRUMS) in comparison to the baseline approach of the whole context terms selection (TF-IDF) were discovered. For the experiments, the collection of literature novel books was chosen as a source of character dialogues as well as in-text information about

them for composing personality traits. With the suggested personality models, we carry out experiments involving two tasks: predicting character responses for ranking-based agents and generating character responses for generative agents. We show that the agents fine-tuned with the personality profiles results in 2.5% (Hits 1@20) and 8% (Rouge-1) improvement for ranking agents and generative agents respectively. We believe in importance of these findings for further work and advances aimed at enhancing retrieval-augmenting systems (RAG) for generative dialogue agents.

## References

- [1] Peter F Brown et al. “An estimate of an upper bound for the entropy of English”. In: *Computational Linguistics* 18.1 (1992), pp. 31–40.
- [2] Yifan Hu, Yehuda Koren, and Chris Volinsky. “Collaborative filtering for implicit feedback datasets”. In: *2008 Eighth IEEE international conference on data mining*. Ieee. 2008, pp. 263–272.
- [3] Gábor Takács, István Pilászy, and Domonkos Tikk. “Applications of the Conjugate Gradient Method for Implicit Feedback Collaborative Filtering”. In: *Proceedings of the Fifth ACM Conference on Recommender Systems*. RecSys ’11. Chicago, Illinois, USA: Association for Computing Machinery, 2011, pp. 297–300. ISBN: 9781450306836.
- [4] Hardik Vala et al. “Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On The Difficulty of Detecting Characters in Literary Texts”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 769–774. DOI: 10.18653/v1/D15-1088.
- [5] Alexander Miller et al. “ParlAI: A Dialog Research Software Platform”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Lucia Specia, Matt Post, and Michael Paul. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 79–84. DOI: 10.18653/v1/D17-2014.
- [6] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017.
- [7] Pierre-Emmanuel Mazaré et al. “Training Millions of Personalized Dialogue Agents”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 2775–2779.
- [8] Saizheng Zhang et al. “Personalizing Dialogue Agents: I have a dog, do you have pets too?” In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 2204–2213. DOI: 10.18653/v1/P18-1205.

- [9] Samuel Humeau et al. “Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring”. In: *arXiv preprint arXiv:1905.01969* (2019).
- [10] Vincent Labatut and Xavier Bost. “Extraction and Analysis of Fictional Character Networks: A Survey”. In: *ACM Comput. Surv.* 52.5 (Sept. 2019). ISSN: 0360-0300. DOI: 10.1145/3344548.
- [11] Jack W Rae et al. “Compressive Transformers for Long-Range Sequence Modelling”. In: *arXiv preprint* (2019). DOI: arXiv:1911.05507.
- [12] Daniel Adiwardana et al. “Towards a human-like open-domain chatbot”. In: *arXiv preprint arXiv:2001.09977* (2020).
- [13] Emily Dinan et al. “The second conversational intelligence challenge (con-vai2)”. In: *The NeurIPS’18 Competition: From Machine Learning to Intelligent Conversations*. Springer, 2020, pp. 187–208.
- [14] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. “Reformer: The efficient transformer”. In: *arXiv preprint arXiv:2001.04451* (2020).
- [15] Aaron W Li et al. “Aloha: Artificial learning of human attributes for dialogue agents”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 8155–8163.
- [16] Fredrik Carlsson et al. “GANDALF: a General Character Name Description Dataset for Long Fiction”. In: *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 119–132. DOI: 10.18653/v1/2021.mrqa-1.13.
- [17] Richard Csaky and Gábor Recski. “The Gutenberg Dialogue Dataset”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 138–159.
- [18] Guendalina Caldarini, Sardar Jaf, and Kenneth McGarry. “A literature survey of recent advances in chatbots”. In: *Information* 13.1 (2022), p. 41.
- [19] Mandy Guo et al. “LongT5: Efficient Text-To-Text Transformer for Long Sequences”. In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 724–736.
- [20] Naoya Inoue et al. “Learning and Evaluating Character Representations in Novels”. In: *Findings of the Association for Computational Linguistics: ACL 2022*. 2022, pp. 1008–1019.
- [21] Romal Thoppilan et al. “Lamda: Language models for dialog applications”. In: *arXiv preprint arXiv:2201.08239* (2022).
- [22] Junzhe Zhao, Huizhi Liang, and Nicolay Rusnachenko. “Dialogue Agents with Literary Character Personality Traits”. In: *2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. 2023, pp. 189–196. DOI: 10.1109/WI-IAT59888.2023.00031.
- [23] Alex Cookson. *Fictional Character Personalities*. Accessed: 2024-05-21. 2024. URL: <https://github.com/tacookson/data/blob/master/fictional-character-personalities/personalities.txt>.