Machine Narrative Comprehension in Fictional Characters Personality Prediction Task

Yisi Sang
Syracuse University

Xiangyang Mou
Rensselaer Polytechnic Institute

Mo Yu
WeChat AI

Dakuo Wang
IBM Research

Jing Li
New Jersey Institute of Technology

Jeffrey Stanton
Syracuse University

Abstract

An NLP model that understands stories should also be able to understand the characters, which is underexplored till now. To support the development of neural models for this purpose, we construct a benchmark, `Story2Personality`. The task is to predict a movie character's personality based on the narratives. Experiments show that our task is challenging for the existing text classification models, as none is able to largely outperform random guesses. We then proposed a multi-view model for personality prediction using both verbal and non-verbal descriptions, which significantly improved the performance. The uniqueness and challenges in our dataset call for the development of narrative comprehension techniques from the perspective of understanding characters.[1]

1 Introduction

Understanding characters in a story is a fundamental human cognitive capability according to psychology and education theories (Bower and Morrow, 1990; Paris and Paris, 2003; Xu et al., 2022). The NLP community has limited work on machine's character comprehension capability, but most of the existing studies focus on short or expository text snippets (e.g., story summaries or fragments) (Urbanek et al., 2019; Brahman et al., 2021; Sang et al., 2022a). Moreover, most of them are limited on the superficial "understanding" (more like information retrieving) of characters, such as coreference resolution (Chen and Choi, 2016) and character relationship extraction (Iyyer et al., 2016). Few studies have explored the actual comprehension of characters, such as from the persona (Flekova and Gurevych, 2015; Sang et al., 2022b) perspective, which is how humans understand a character and build connections with it in a book reading (Xu et al., 2021) or movie watching experience.

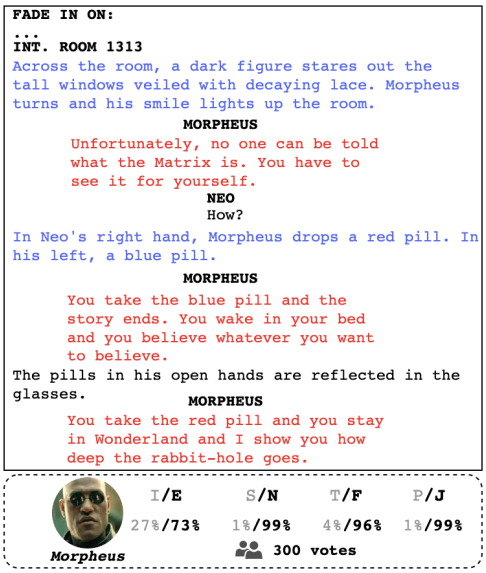

Figure 1: An example excerpt from the movie script of "The Matrix". Blue utterances are the character *Morpheus's* **scene descriptions**; Red are his **dialogues**. *Morpheus's* personality was rated as **ENFJ** by 300 user votes.

We propose `Story2Personality`, a new narrative understanding benchmark to enable new research opportunities of character understanding. The task of `Story2Personality` is to predict personality according to the character's narrative texts in the script. We focus on the The Myers–Briggs Type Indicator (MBTI) (Myers, 1962) personality type that assess the psychological preferences in how people perceive the world and make decisions in four categories: introversion or extraversion, sensing or intuition, thinking or feeling, judging or perceiving.

Personality prediction from narratives has many challenges. First, stories often use a variety of narrative clues (e.g., scenery changes),sequence (e.g., flashback) and rhetorical techniques (e.g., metaphor). Second, the inputs of the task are long (>10K words on average), challenging the applications of Transformer-based models (Vaswani et al.,

[1]Our code and data are released at https://github.com/YisiSang/Story2Personality

2017). Third, both the scene descriptions and dialogues are informative for the prediction, requiring models to jointly consider multi-view of inputs.

We make the following contributions:

- We establish a large-scale dataset for personality prediction of narrative characters (3,543 characters from 507 movies with 4-dimensions MBTI label). Our dataset is proved challenging — on this binary classification task, none of the baselines achieve higher than 60% macro-F1.
- We develop a trainable movie script parser to automatically process a script to a structured form with the verbal dialogues and the non-verbal scene descriptions illustrating backgrounds. Human study shows that our parser is more accurate compared to previous rule-based tools.
- Inspired by psychological theories (McCroskey and Richmond, 1996; Richmond et al., 2008), we propose an extension to BERT classifier (Devlin et al., 2018) to handle the long and multi-view (verbal and non-verbal) inputs. Our model improves 2-3% over the baselines.

2 Related Work

Character-Centric Narrative Understanding There have been a few existing studies on character-centric machine narrative understanding, but only work on summaries of stories or summaries of characters (Massey et al., 2015; Srivastava et al., 2016; Brahman et al., 2021). Thus, they do not need to handle the long narrative inputs as how humans read a narrative. Some other works consider long narratives as input but focus only on extracting inter-character relationship via counting co-occurrence (Elson et al., 2010; Elsner, 2012; Iyyer et al., 2016; Chaturvedi et al., 2016, 2017; Kim and Klinger, 2019).

For our goal of understanding characters from narratives, we rely on fundamental NLP techniques for books and screenplays, such as named entity recognition (Bamman et al., 2019), coreference resolution (Chen and Choi, 2016) and entity-centric natural language modeling (Clark et al., 2018).

The most relevant work to ours is about latent persona induction (Bamman et al., 2013). The work learns a topic model over character behaviors from books, and then consider or assume each latent topic corresponds to an induced persona. The induced persona vectors can be then applied to potential applications as a type of character representation, but they did not have psychological theory behind the assumption.

Background of MBTI Personality is a "stable and measurable" individual characteristic (Vinciarelli and Mohammadi, 2014) which can "distinguish internal properties of the person from overt behaviors" (Matthews et al., 2003). MBTI and the Big-5 Personality are two of the most popular personality scales. We choose MBTI as the annotation criteria as research shows that a person's friend can accurately judge his/her MBTI personality (i.e., third-person judgement validity) (Cohen et al., 1981). In our narrative comprehension scenario, a fictional character's MBTI personality is judged by other human watchers (third person), which should yield a reasonable validity.

MBTI has four dimensions. E/I: extravert (E) is seen as being generally active and objective while the intravert (I) is seen as generally passive and subjective (Sipps and Alexander, 1987). S/N: sensing (S) is seen as attending to sensory stimuli; intuition (N) describes a more detached, insightful analysis of events and stimuli (Boyle, 1995). T/F: thinking (T) involves logical reasoning and decision making; feeling (F) involves a more subjective and interpersonal approach (Thomas, 1983). J/P: judging (J) attitude is associated with prompt decision making; perception (P) involves greater patience and waiting for more information before making a decision. An individual's MBTI type has a label based on her dominant preference for each dimension (e.g., *Morpheus* from "The Matrix" is ENFJ in Figure 1).

3 `Story2Personality` Dataset

We constructed our dataset in three stages: extracting movie scripts from the Internet Movie Script Database (*IMSDB* [2]), parsing the collected movie scripts into dialogue and scene sections, matching characters' personality types from The Personality Database(*PDB*[3]) with their dialogues and scenes.

3.1 Movie Scripts Collection

We crawled HTML files from IMSDB combined with movie scripts in NarrativeQA (Kočiskỳ et al., 2018). After removing corrupted or empty files, we got 1,464 usable movie scripts.

3.2 Our Statistical Movie Script Parser

As shown in Figure 1, a movie script usually has four basic format elements (Riley, 2009): **Scene Headings**, one line description of each

[2] https://imsdb.com/

[3] https://www.personality-database.com/

scene's type, location, and time (i.e., `INT. ROOM 1313`); **Scene description**, the description of the actions of the characters (i.e., text in blue); **Dialogues**, names of characters and actual words they speak (i.e., text in red); **Transitions**, instructions for linking scenes together (i.e., `FADE IN ON`).

In order to extract dialogues and scene descriptions in a structured form, we first split the scripts to sections, i.e., text chunks between two adjacent bolded chunks which are scene headings or character names and stored the bolded texts as section titles. Then we designed a statistical method to classify the section types:

Rule-Based Pre-Processing We start with a rule to classify the sections into dialogues and scenes. As Figure 1 shows, a common format of movie scripts is to align the shot headings, transitions and scene descriptions vertically, and uses a larger indentat for dialogues. So, the indent size can be used to identify dialogues. Since the indentat size may vary across different scripts. Our rule assumes the sections as dialogues if they have larger indent compared to `FADE IN` in the same script and the others as scenes.

Silver Parses Construction The rule-based preprocessing introduces many noises. We then designed a statistical method to automatically determine the threshold indent of dialogues. First, we compute the averaged ratio μ of dialogues in a script and its standard variation σ. Second, we keep adding sections with the largest indent sizes to the set of dialogues, until the ratio of added sections becomes larger than $\mu+\sigma$. Finally, we keep the left sections as scenes. If none of the indentation size can reach the ratio of dialogues in the range of $\mu\pm\sigma$, the movie script was seen as a failure case. We designated the successfully processed scripts with the dialogues/scene labels as the "silver" set which consists of 29% of the scripts.

Section Classifier For the failure scripts from the previous step and the scripts without `FADE IN` markers, we trained a BERT-based section classifier using 137,042 labeled sections from the silver set to label them. The classifier achieved 99.31% accuracy on a held out validation set. The outputs are our final parses.

3.3 Personality Collection and Mapping

We collect human rated MBTI types from *PDB*. Movie scripts are the blueprint for the actor's performance. An actor's body language, dialogue, and contexts are all described in the scripts (Jhala, 2008). Human rater's perception of a character's personality from the movies would be consistent with the script's description. In total, we collected MBTI types of 28,653 characters. Each character has an id, name, vote count, and voters' agreement on each MBTI dimension. For example, the MBTI profile in Figure 1 has 300 voters, with different agreement rate along each dimension. To ensure the quality of personality voting, we removed character profiles with <3 voters and <60% agreement rate so some characters do not have all the 4 dimensions. When the user starts rating, the rating interface hides the previous rater's choices. Thus, the rater would not have prior bias. We then matched the characters' personality profiles to the scripts, if the name can be softly matched to the dialogue title or the recognized named entities in the scenes. Table 1 shows the core statistics of our dataset.

	Dimension	Train(%)	Dev(%)	Test(%)
(a)	E/I	45.9/51.8	49.6/49.0	52.6/44.2
	N/S	36.6/60.4	41.8/54.0	41.4/55.0
	T/F	54.7/43.2	45.8/50.8	46.0/52.8
	J/P	46.4/51.3	47.2/51.2	45.6/53.0

		Mean	Min	Max
(b)	# dialogues/character	76.90	0	776
	# words/dialogue	917.74	1	12, 536
	# scenes/character	41.08	0	495
	# words/scene	1,381.47	1	25,457

Table 1: Distribution of two personality types per dimension (a) and core statistics (b) in `Story2Personality`.

3.4 Statistics of Human Agreement

Table 2 lists the human agreement score on each MBTI dimension, on which we compute the human accuracy and approximate human macro-F1 scores.

The raters are most divided in annotation of N, with an average agreement is 91.06% and the standard deviation 0.11. One reason is that the perceptual style dimension N/S measures how the individual obtain information. Comparing with dimensions related to attitudes (E/I) or decision making (T/F, J/P) (Jung, 2016) perceptual style is more implicit. Specifically, S is seen as attending to sensory stimuli, while N describes a more detached, insightful analysis of events and stimuli (Boyle, 1995). They are more difficult to determine from the explicit story narratives.

	Mean	Min	Max	STD	#Character
I	94.43%	60%	100%	0.10	1,783
E	94.22%	60%	100%	0.10	1,679
N	91.06%	60%	100%	0.11	1,347
S	93.32%	60%	100%	0.11	2,082
T	94.22%	60%	100%	0.10	1,851
F	93.68%	60%	100%	0.10	1,617
P	93.68%	60%	100%	0.10	1,825
J	93.72%	60%	100%	0.10	1,644

Table 2: Descriptive statistics of voters' agreement

	Correct scene	Correct dialogue
Ramakrishna et al. (2017)	85%	93%
Our parser	**97%**	**100%**

Table 3: Comparison of correct parsing results.

4 Dataset Analysis

Script Parsing Results We compared our parsing results with the results of the state-of-the-art open-sourced script parser (Ramakrishna et al., 2017), which employs many human written rules, with a human study. We randomly selected five scene descriptions and five dialogue sections in 10 common movies, giving 100 snippets for evaluation (40 from the silver set). Then we manually compared the parsing results with the original movie scripts. Table 3 shows the results. Our parser outperforms Ramakrishna et al. (2017) with a large margin. Most mistakes of (Ramakrishna et al., 2017) is to recognize scenes as dialogues. There are other parsers but did not publish the code or data, so we cannot conduct human study for comparison. A state-of-the-art learning model (Agarwal et al., 2014) reports 91% accuracy on line-level classification. In a preliminary study, we achieve 99% on this task, but finally choose do conduct more accurate section-level classification as in Section 3.2.

Human performance We take the majority vote of each character's MBTI types as the groundtruth. This gives an averaged 93.54% human accuracy across the four personality dimensions on our test data. Computing humans' macro-F1 score lacks an analytical form from the agreement scores. Therefore we make an approximation by sampling three voters (the minimum number of voters in our dataset) for each character and treating them like the predictions of three different models. This gives overall >95% scores which is much higher than model performance (in Table 6). The statistics of human agreement on MBTI dimensions is shown in Table 2. Table 4 lists the distribution of all the

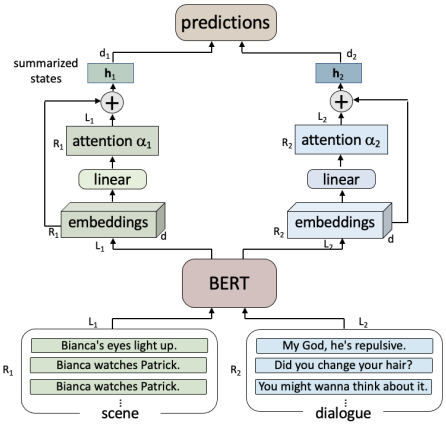

Figure 2: Multi-row multi-view BERT model architecture.

16 MBTI types in our dataset, together with a representative movie character for each type.

Personality	%	Example
ISTJ	8.41%	*Darth Vader* ("Star Wars")
ISTP	8.07%	*Shrek* ("Shrek")
ESTP	8.21%	*Han Solo* ("Star Wars")
ESTJ	6.52%	*Boromir* ("The Lord of the Rings")
ISFJ	6.41%	*Forrest Gump* ("Forrest Gump")
ISFP	6.49%	*Harry Potter* ("Harry Potter")
ESFJ	4.88%	*Cher Horowitz* ("Clueless")
ESFP	7.06%	*Jack Dawson* ("Titanic")
INFJ	4.80%	*Edward Cullen* ("Twilight")
INFP	5.42%	*Amélie Poulain* ("Amélie")
ENFP	3.90%	*Anna* ("Frozen")
ENFJ	3.75%	*Judy Hopps* ("Zootopia")
INTJ	4.26%	*Michael Corleone* ("The God Father")
INTP	3.75%	*Neo* ("The Matrix")
ENTP	4.94%	*Tyler Durden* ("Fight Club")
ENTJ	4.88%	*Patrick Bateman* ("American Psycho")

Table 4: Distribution of the 16 MBTI personality types in `Story2Personality`

5 Experiments

Baselines We build two baseline models.

•**SVM**, the LinearSVC from sklearn.svm. We extracted top 20K word unigram, bigram, and trigram features according to term frequency after removing stop words. We set C=0.1.

• **BERT**, fine-tuning the out-of-box BERT, with a linear head on the '[CLS]' token's final layer embedding for classification.

Our Method We propose the multi-view multi-row BERT (**MV-MR BERT**) classifier (Fig. 2) which is an extension of BERT to deal with the long inputs and handle the verbal and non-verbal information differently. First, to handle the long input per character, we borrow the idea from fusion-

	1K	1.5K	2K	∼2.5K
SVM	50.33	52.19	54.56	55.41
BERT	54.32	55.42	55.58	55.59

Table 5: Learning curve on N/S.

in-decoder (Izacard and Grave, 2020). Since the complexity of Transformers is $O(RL^2)$ (with R the number of rows and L the length per row), when L is very large, we can split it into multiple segments to reduce the quadratic term. Next, we rely on the attention over all the segments to fuse the information. Specifically, we split the input content $\mathcal{D}$ of a character into multiple segments $\mathcal{D} = \{\mathcal{S}_i\}_{i=1}^{R}$, and encode all the segments in a minibatch as $\mathbf{H} = \text{BERT}(\mathcal{S}_i) \in \mathbb{R}^{R\times L\times d}$, where d is the hidden state size. Then a linear head is applied to get the attention score across tokens in all the rows as $\alpha = \text{softmax}(\mathbf{H}W + b) \in [0, 1]^{R\times L}$. The final summarized representation of the input $\mathcal{D}$ is thus the weighted summation $\mathbf{h}_\mathcal{D} = \sum_{i=1}^{R}\sum_{j=1}^{L}\alpha_{ij}\mathbf{H}_{ij}$. Second, to handle both the dialogue and behavioral description a character, our multi-view model receives an input pair $(\mathcal{D}^{\text{dial}}, \mathcal{D}^{\text{scene}})$, then uses a shared BERT and separated linear heads to compute the summarized states $\mathbf{h}_\mathcal{D}^{\text{dial}}$ and $\mathbf{h}_\mathcal{D}^{\text{scene}}$. The two vectors are fed into a fully-connected layer for prediction. For the scene descriptions, we prepend a special token "[ent]" to the target character's name to denote its position. The attention α^{scene} is only computed on these special tokens.

Results and Analysis Following Flekova and Gurevych (2015), we use macro-averaged F1 as evaluation metric. Table 6 shows the main results on the four MBTI dimensions. Peak performance was achieved by our MV-MR BERT. The result suggests using both dialog and action scene descriptions consistently improved model performance.

The results are generally low compared to human performance, showing the task is challenging to existing models. We analyzed the learning curve of BERT model by adding the training data from 1K to 2.5K characters (Table 5). The model performance did not change a lot in the development dataset. Figure 3 gives further evidence for the challenge of our task, which shows the dev and test results are not highly-correlated, meaning that by achieving near perfect accuracy on the training data, the models largely overfit the noises instead of capturing real clues.

Figure 3 gives further evidence for the challenge

Model	E/I	N/S	T/F	J/P
SVM	54.65	55.41	52.83	56.18
BERT	$56.06_{\pm 0.73}$	$55.59_{\pm 3.36}$	$57.13_{\pm 0.97}$	$57.59_{\pm 1.40}$
MV-MR BERT	$\mathbf{57.50}_{\pm \mathbf{2.04}}$	$\mathbf{57.42}_{\pm \mathbf{4.27}}$	$\mathbf{60.33}_{\pm \mathbf{0.93}}$	$\mathbf{59.83}_{\pm \mathbf{1.42}}$
- multiview	$57.30_{\pm 1.91}$	$57.05_{\pm 1.80}$	$57.04_{\pm 2.05}$	$57.39_{\pm 2.21}$
Human Perf.	$98.19_{\pm 0.60}$	$97.82_{\pm 0.10}$	$98.51_{\pm 0.67}$	$98.03_{\pm 0.19}$

Table 6: Macro F1 scores on the four dimensions

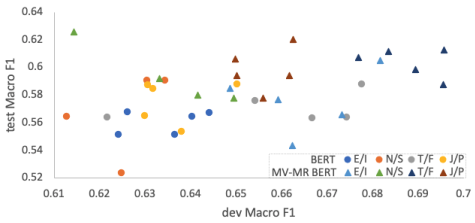

Figure 3: Dev vs. test F1 scores of BERT-based models.

	0.5K	1K	2K	4K
J/P	59.79	58.18	60.35	58.77
T/F	59.91	64.31	63.32	65.42
N/S	56.15	53.86	55.65	57.18
I/E	61.64	61.05	62.87	62.69

Table 7: Ablation experiment on input length.

of our task, which plots the dev versus test scores during our model selection. It shows the dev and test results are not highly-correlated, meaning that by achieving near perfect accuracy on the training data, the models largely overfit the noises instead of capturing real clues. Both length and multiview have an improvement on model performance, but length has a slightly smaller impact, as shown in Table 7, when increasing the number of input tokens, the performance is not greatly affected.

6 Conclusion

We develop a movie script parser and proposed a new narrative understanding benchmark, `Story2Personality`, which enables neural model training for understanding characters. We evaluate several classifiers on our task – while our multi-view multi-view BERT model achieves a substantial improvement over the SVM and BERT baselines, there is a huge gap compared to human performance. This indicates our dataset a valuable and challenging task for future research. In the future, we will continue to expand our dataset and build downstream applications.

Acknowledgements

This research was supported, in part, by the NSF (USA) under Grant Numbers CNS–1948457.

References

Apoorv Agarwal, Sriramkumar Balasubramanian, Jiehan Zheng, and Sarthak Dash. 2014. Parsing screenplays for extracting social networks from movies. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 50–58.

David Bamman, Brendan O'Connor, and Noah A Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. 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)*, pages 2138–2144.

Gordon H Bower and Daniel G Morrow. 1990. Mental models in narrative comprehension. *Science*, 247(4938):44–48.

Gregory J Boyle. 1995. Myers-briggs type indicator (mbti): some psychometric limitations. *Australian Psychologist*, 30(1):71–74.

Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. " let your characters tell their story": A dataset for character-centric narrative understanding. *arXiv preprint arXiv:2109.05438*.

Snigdha Chaturvedi, Mohit Iyyer, and Hal Daume III. 2017. Unsupervised learning of evolving relationships between literary characters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Snigdha Chaturvedi, Shashank Srivastava, Hal Daume III, and Chris Dyer. 2016. Modeling evolving relationships between characters in literary novels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Yu-Hsin Chen and Jinho D Choi. 2016. Character identification on multiparty conversation: Identifying mentions of characters in tv shows. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 90–100.

Elizabeth Clark, Yangfeng Ji, and Noah A Smith. 2018. Neural text generation in stories using entity representations as context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2250–2260.

David Cohen, Marilye Cohen, and Herbert Cross. 1981. A construct validity study of the myers-briggs type indicator. *Educational and Psychological Measurement*, 41(3):883–891.

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

Micha Elsner. 2012. Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 634–644.

David K Elson, Kathleen McKeown, and Nicholas J Dames. 2010. Extracting social networks from literary fiction.

Lucie Flekova and Iryna Gurevych. 2015. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1805–1816.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

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

Carl Jung. 2016. *Psychological types*. Routledge.

Evgeny Kim and Roman Klinger. 2019. Frowning frodo, wincing leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters. In *Proceedings of NAACL-HLT*, pages 647–653.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Philip Massey, Patrick Xia, David Bamman, and Noah A Smith. 2015. Annotating character relationships in literary texts. *arXiv preprint arXiv:1512.00728*.

Gerald Matthews, Ian J Deary, and Martha C Whiteman. 2003. *Personality traits*. Cambridge University Press.

James C McCroskey and Virginia P Richmond. 1996. *Fundamentals of human communication: An interpersonal perspective*. Waveland PressInc.

Isabel Briggs Myers. 1962. The myers-briggs type indicator: Manual (1962).

Alison H Paris and Scott G Paris. 2003. Assessing narrative comprehension in young children. *Reading Research Quarterly*, 38(1):36–76.

Anil Ramakrishna, Victor R Martínez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. 2017. Linguistic analysis of differences in portrayal of movie characters. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1669–1678.

Virginia P Richmond, James C McCroskey, and Mark Hickson. 2008. *Nonverbal behavior in interpersonal relations*. Allyn & Bacon.

Christopher Riley. 2009. *The Hollywood standard: the complete and authoritative guide to script format and style*. Michael Wiese Productions.

Yisi Sang, Xiangyang Mou, Jing Li, Jeffrey Stanton, and Mo Yu. 2022a. A survey of machine narrative reading comprehension assessments. *arXiv preprint arXiv:2205.00299*.

Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li, and Jeffrey Stanton. 2022b. Tvshowguess: Character comprehension in stories as speaker guessing. *arXiv preprint arXiv:2204.07721*.

Gary J Sipps and Ralph A Alexander. 1987. The multifactorial nature of extraversion-introversion in the myers-briggs type indicator and eysenck personality inventory. *Educational and Psychological Measurement*, 47(3):543–552.

Shashank Srivastava, Snigdha Chaturvedi, and Tom Mitchell. 2016. Inferring interpersonal relations in narrative summaries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Charles R Thomas. 1983. Field independence and myers-briggs thinking individuals. *Perceptual and motor skills*.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. *arXiv preprint arXiv:1903.03094*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291.

Ying Xu, Dakuo Wang, Penelope Collins, Hyelim Lee, and Mark Warschauer. 2021. Same benefits, different communication patterns: Comparing children's reading with a conversational agent vs. a human partner. *Computers & Education*, 161:104059.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, et al. 2022. Fantastic questions and where to find them: Fairytaleqa–an authentic dataset for narrative comprehension. *ACL'22*.