Can Language Model Learn Ethics? Predicting Characters' Morality by Learning Character Embedding

Su-Young Bae[1], Eun-Chong Kim[2] and Yun-Gyung Cheong[2]

[1] Department of AI, The Social Innovation Convergence Program, Sungkyunkwan University, Suwon, Korea, sybae01@skku.edu
[2] Department of AI, Sungkyunkwan University, Suwon, Korea, {prokkec, aimecca}@ eskku.edu

Abstract

Story generation and analysis have been research for decades. However, while ethics is becoming essential consideration for developing AI applications, few research deals with morality in narrative. To bridge the gap, we propose an morality judgment task using story books in this paper. We present the framework to build embeddings for representing characters to predict the character's morality. We preprocessed existing data for building datasets for evaluation. We carry out a number of experiments and the results suggest that word embedding models can learn character's morality. This paper reports the results and discuss our findings.

Keywords— *Ethics, Morals, Story Character Embedding, Masked Language Modeling*

I. INTRODUCTION

With the advent of large-scale transformer-based models [1][4][7], word embeddings can learn various contexts. Some studies on narrative suggest that these embeddings can learn the story context, exhibiting various characteristics associated with story, character, and scenes. For instance, [6] builds character embeddings using BiLSTM models for story generation, and [5] used BERT and Elmo to predict the quality and popularity of movie scripts. However, there is a lack of research investigating characters' morality, while morality is becoming essential issue as artificial intelligence machines are increasingly intervening human social activities such as counseling, health care, and education [8].

To bridge the gap, we aim to learn characters' morality using character-centric embeddings. In this paper, we test whether existing language models can predict the morality of a character in stories. For evaluation, we fine-tune a number of BERT based masked language models using the datasets we built to extract character embeddings. Then, we calculate the characters' morality score by applying the character embeddings to moral datasets. We carried out evaluation and the results suggest that the character embeddings using the BERT model can capture characters' morality. Furthermore, the performance can be improved when using our preprocessed datasets and a masking scheme.

In this study, we make the following contributions. 1. We propose an morality judgment task using story books corpora and moral stories datasets. 2. We present the framework to build embeddings for representing characters to predict the character's morality. 4. We propose data preprocessing for building character-focused embedding. 4. We implement the framework and carry out experiments to evaluate the efficacy of our framework.

II. METHOD

We focus on character embeddings in stories and test whether the embeddings built using masked language models can represent the characters' morality or not. Fig.1 and Fig.2 illustrate the overall process of building character embeddings.

A. Dataset

Story Dataset for Character Embeddings: We chose the Harry Potter series authored by J. K. Rowling and obtained the data from the Harry Potter Books Corpora at the Kaggle site[1]. There is also contains a charater list file containing character names and their bios in the series. We use all of the harry potter book corpora and we believe that the data are large enough to obtain character embeddings to represent character's morality.

In addition, the characters in this book series have conflicting ethics, conforming our purpose to compare the characters in terms of morality.

Morality Dataset: For evaluating whether fine-tuned masked language models can learn characters' morality,

[1] https://www.kaggle.com/datasets/balabaskar/harry-potter-books-corpora-part-1-7

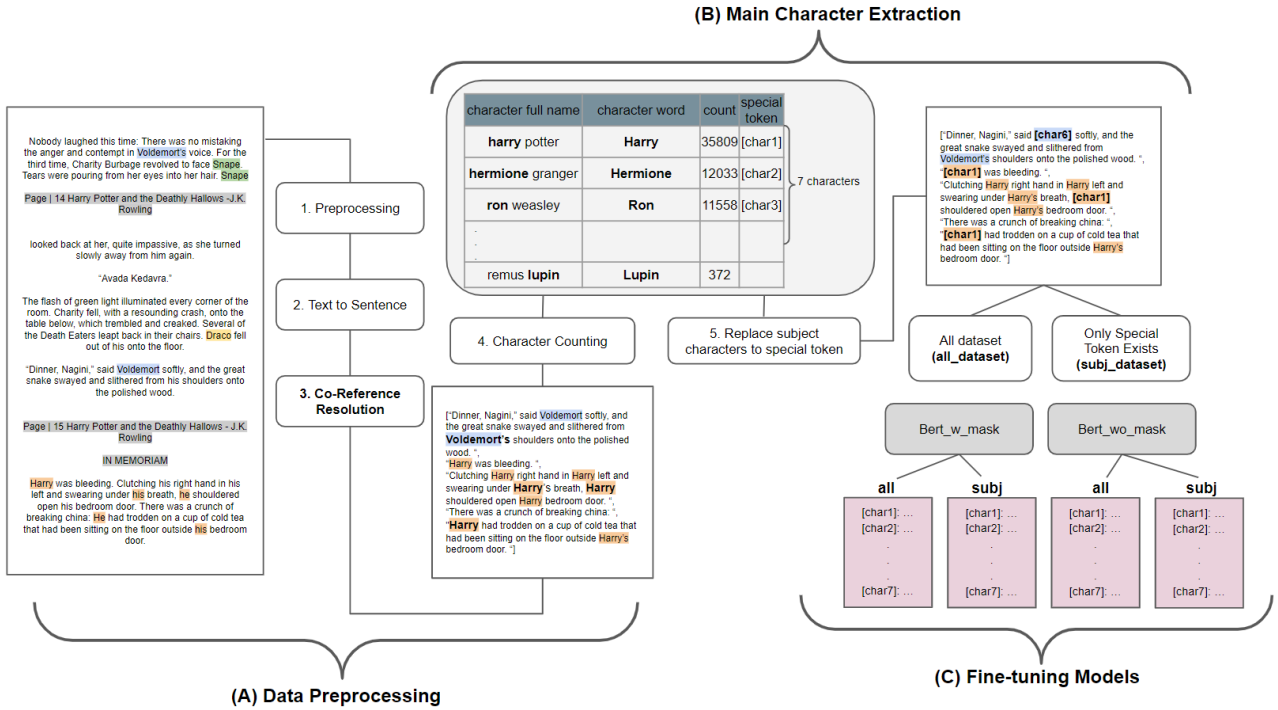

Fig. 1. In the process of (A), we preprocess a story dataset. In (B), we extract seven main characters in a story by counting the number of character names. And the last part, (C), we fine-tune three masked language models each to get seven character embeddings in different settings.

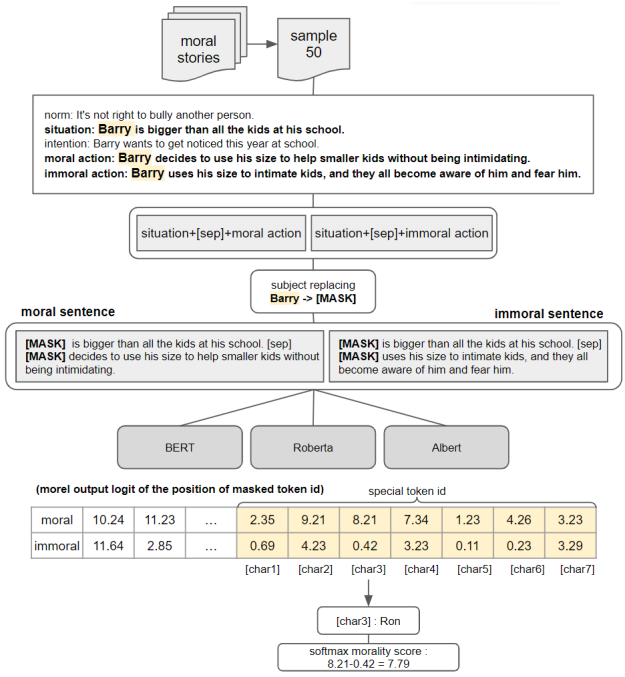

Fig. 2. (D) Process of evaluating character embeddings' morality.

we use the moral stories dataset at [2]. This is a crowd-sourced dataset of structured, branching narratives to study grounded, goal-oriented social reasoning [2]. It consists of seven categories: Norm, Situation, Intention, moral action, moral consequence, immoral action, and immoral consequence. We use the situation, moral action, and immoral action categories for our evaluation. The situation describes the story's social setting that introduces story participants. The moral action is an action performed by the actor that fulfills the intention while observing the norm. On the other hand, the immoral action is an action performed by the actor that meets the intent while violating the norm. We randomly selected 50 situations and action sets for evaluation.

B. Data Preprocessing

Since our aim is to verify the pre-trained masked language models' ability to capture characters' morality, we did not train language models using other story datasets. We combined all of the seven books into one corpus. The first preprocessing step eliminates irrelevant tokens in the datasets, such as page numbers, book titles, chapter titles (grey highlighted in (A) section in Fig.1), and stopwords. The second stage splits the corpus into sentences using the sentence tokenize functionality provided by the nltk package[3]. Short sentences tend not to help learn character embeddings. Therefore, we remove the sentences with less than five words from the dataset.

[2] https://huggingface.co/datasets/demelin/moral_stories

[3] https://www.nltk.org/api/nltk.tokenize.html

C. Main Character Extraction

After preprocessing datasets, we extract the main characters in stories and replace them with special tokens. In order to extract main characters, we first apply coreference resolution using the SpanBERT model [3]. We define characters that appear more than 3,000 times in the text as the major characters and the others are minor characters. We believe that meaningful embeddings would be created when the corpus is sufficient, and set the threshold value of appearance as 3,000. To compute the frequencies, we look up the character list file included in the dataset. A characters' name is made up of the first and the last name, e.g., 'Hermione Granger.' When computing its frequency, we need to count the appearances of 'Hermione', and 'Granger' as well.

For building character embeddings, we convert character names into special tokens. In this study we use special tokens, such as '[char1]', '[char2]', '[char3]', ... '[charN]' and add them to the models' vocabulary. We replace the main characters with special tokens when they serve as the subjects of the sentences. We disregard the sentences when a main character serves as the object because we aim to embed the character's morality. In the case of the sentence 'Malfoy hits Harry', for example, the hit action is associated with Malfoy but not with Harry since Harry is the object of this sentence. Then, Malfoy in the sentence is replaces with '[char2].' We create two datasets for comparison: one that contains all the sentences and the other that contains only the sentences that has special tokens (sentences with their subjects as main characters).

D. Fine-tuning Models

We use three masked language models to compare their performances to estimate the characters' morality. We fine-tune each model using the two datasets we built (all_dataset, subj_dataset) and two types of masking: masking 15 percent of entire datasets(w_random) and masking only special token ids(wo_random).

E. Morality Estimation

We use the morality dataset to compute the morality score of each character embedding. We first randomly select 50 samples in moral stories and make a sentence by concatenating situation, moral action, and situation, immoral action using [sep] token. Then, we mask the words corresponding to the subject of the sentence. Taking the two sentences (moral sentence and immoral sentence) as the input text, the three fine-tuned masked language models output logits in the position of mask token ids((logits[mask_token_id]). After applying the softmax and sum functions, we compare logit scores at the position of each special token ids(logits[mask_token_id][character_special_token_id]) between moral text and immoral text.

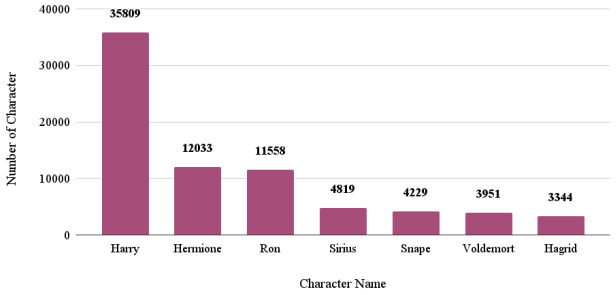

Fig. 3. Seven main characters in the Harry Potter Books corpora by descending order.

For example, Fig.2-(D), there are two mask tokens in a moral sentence and an immoral sentence. So we apply the softmax function of each position of mask token. And add them all. If you want to look at the probability of 'Ron' entering the mask token of the sentence, you can look at the value of the '[char3]' special token index position.

We define the *morality score* of a character as:

$$morality_{character} = score_{moral} - score_{immoral} \quad (1)$$

In the equation, $score_{moral}$ denotes the summation of softmax output logit scores at character special token id in moral text, and $logit_{immoral}$ is a sum of softmax output logit scores at character special token id in moral text. We fine-tune each model with different settings for evaluation.

III. Experiments

A. Models

We used the following three masked language models to compare character embeddings' morality scores.

BERT-base-uncased: the bert-base-uncased model has 110M parameters and supports masked language modeling. This model can learn input embeddings by masking input texts and predicting the masked words. The vocab size of the bert-base model is 30,522, and the mask token id is 103.

RoBERTa-base: the roberta-base model uses dynamic masking and is trained on large-scale text corpora dataset. The vocab size of the Roberta model is 50,265, and the mask token id is 50,264.

ALBERT-base-v2: the albert-base model is a light version of the Bert model. It has a reduced model size and improved model performance. The vocab size of the ALBERT model is 30,000 and the mask token id is 4.

Both the BERT and ALBERT models encode a mask token as $[MASK]$, while RoBERTa uses a mask token as $\langle mask \rangle$. We add seven special tokens denoting the main characters to the vocabulary, changing its size to 30,532 for BERT, 30,010 for ALBERT, and 50,275 for RoBERTa. We set hyperparameters such that epoch is 5, the batch size is 8, adam optimizer, and the learning rate is 5e-5 in the NVIDIA GeForce RTX 3,090 environment.

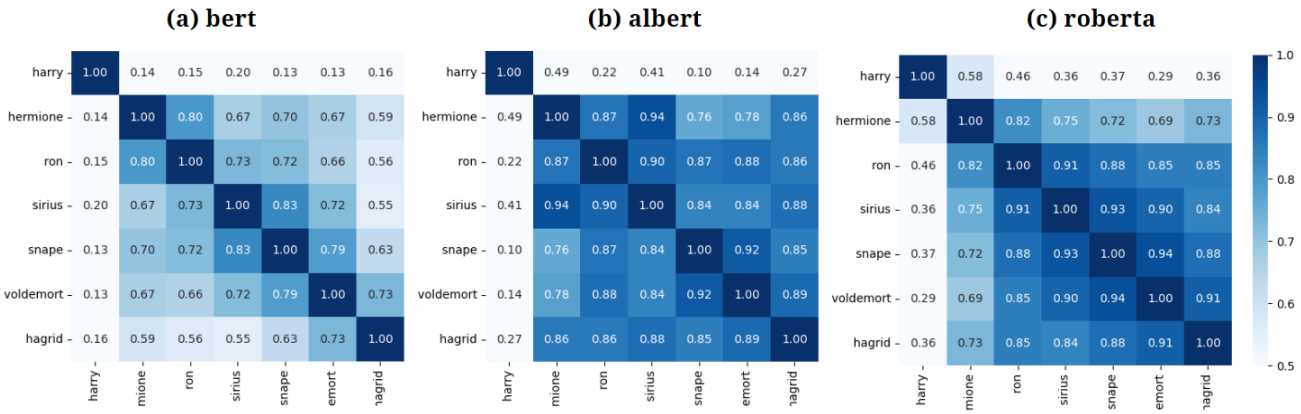

Fig. 4. Similarity heatmaps between the main characters. The three heatmaps are obtained using the three models and the additional masking method(w_random) with all datasets(all_dataset)

B. *Character Embedding*

The total number of a sentence in the Harry Potter Books dataset is 56,132. Each book text consists of the title, pages, chapter titles, and the contents. We eliminate the content-unrelated information to obtain 49,906 sentences (all_dataset). We then extract seven characters from the dataset. Fig.3 shows the histograms of the number of characters' appearances in the entire dataset. In addition, we built another dataset that contains only the sentences with special tokens to obtain 25,318 sentences (subj_dataset). Finally, we train the three models (BERT, ALBERT, and RoBERTa) with the two datasets (all_dataset and subj_dataset) with different masking methods to obtain the character embeddings.

IV. Results

A. *Similarities of Character Embeddings*

Before evaluating character embeddings' morality, we compute the cosine similarity scores between the main characters' embeddings using the three different fine-tuned models and the additional random masking(w_random) with all datasets (all_dataset). Fig.4 exhibits that the scores generated by the BERT model tend to be lower than the other models. We also found that the similarity scores associated with Harry are relatively lower than the similarity scores engaging the other characters. We reason that this is related with the high frequency of Harry in the text. To verify this, we need to check with other stories in the future. In addition, Hermione is most similar to Harry, although the similarity score is not high. This makes sense because Harry and Hermione are close to each other in the story and share similar view of life.

B. *Character Morality*

First, we compare the morality scores obtained from character embedding trained by the three embedding models, using all dataset(all_dataset) and the dataset(subj_dataset) containing main characters only. To check differences between the models using two different data types (e.g., all_dataset and subj_dataset), we combine the two outputs produced by each embedding model applying w_random and wo_random. Table 1 shows the morality scores using the different data types. We note the scores in bold when the scores correctly predict the character's morality. We believe that Harry, Hermione, and Ron are ethical and Sirius, Snape, Voldemort are unethical. Hagrid is a controversial character; he is good, but some of his actions are immoral. For example, Hagrid keeps illegal creatures in his house. However, we regard him as ethical in this study. Therefore, we expect positive morality scores for Harry, Hermione, Ron, and Hargid, and negative scores for Sirius, Snape, and Voldemort.

Table 1 demonstrates that the BERT model outperforms the other models in predicting the morality of a character correctly for both of the *all* and the *subj* datasets. It is noted that the morality score for Harry is not high. In particular, ALBERT and RoBERTa produce negative scores for him regardless of the data types. This makes sense because Harry is good but takes violent actions against unethical characters. In addition, Voldemort is regarded as ethical for most models, except the RoBERTa model using the *subj* dataset. Voldemort generally orders other characters to commit crimes, but he himself does not commit crimes.

The results also suggest that the *subj* dataset is more useful, as the scores are higher for ethical characters than those produced using the *all* dataset. This is consistent with our hypothesis that the morality of a sentence is associate with its subject but not with its object. In addition, using the *subj* dataset produces positive scores for all the three models for Ron. Second, we check whether the additional masking method has an impact on estimating the character's morality. We build the *w_random* dataset by additionally masking 15 percent of the entire data for fine-tuning masked language models. We obtain the morality scores of each character when using this dataset and when us-

Dataset	Model	Harry	Hermione	Ron	Sirius	Snape	Voldemort	Hagrid
all	BERT	**0.08**	**0.57**	**0.35**	**-0.53**	**-0.20**	0.15	**0.25**
	ALBERT	-0.65	**0.07**	-0.35	0.24	0.31	0.09	-0.17
	RoBERTa	-0.59	-0.32	-0.19	**-0.04**	0.18	0.25	**0.02**
subj	BERT	**0.67**	**0.19**	**0.72**	**-0.65**	**-0.47**	0.15	-0.34
	ALBERT	-0.13	**0.94**	**0.63**	-0.36	0.17	0.02	**0.08**
	RoBERTa	-0.29	-0.02	**0.02**	**-0.07**	**-0.05**	**-0.01**	**0.03**

Table 1. Morality scores obtained from character embedding trained the three language models, using the *all* dataset which contains all the sentences and the *subj* dataset which contains the sentences with main characters only. Positive scores denote ethical characters, while negative score denote unethical characters.

Random mask	Model	Harry	Hermione	Ron	Sirius	Snape	Voldemort	Hagrid
with mask	BERT	**0.94**	**1.18**	**0.60**	**-0.74**	**-0.79**	0.19	-0.04
	ALBERT	-0.60	**0.91**	**0.41**	**-0.27**	0.41	0.09	-0.02
	RoBERTa	0.00	-0.11	-0.30	**-0.10**	**-0.13**	**-0.04**	-0.04
without mask	BERT	-0.19	-0.42	**0.47**	**-0.45**	0.11	0.10	-0.05
	ALBERT	-0.19	**0.09**	-0.13	0.15	0.07	0.02	-0.06
	RoBERTa	-0.87	-0.23	**0.12**	**-0.02**	0.27	0.29	**0.09**

Table 2. Morality scores obtained from character embedding trained three models, each applying additional random masking and without random masking. Positive scores denote ethical characters, while negative score denote unethical characters.

ing the *wo_random* dataset, which is built by masking the special token ids. We then add the output produced using all_dataset and the output produced using the subj_dataset for each model. Table 2 lists the morality scores by using different masking methods.

Table 2 demonstrates that all the models applying additional masking predict characters' morality better than without using it. Again, the BERT model outperforms the other models, predicting correct morality 5 out of 7. Interestingly, RoBERTa tends to produce negative scores.

V. Conclusion and Future Work

In this paper, we make an attempt to learn characters' morality using character-centric embeddings. We present a framework to build character embeddings and propose a preprocessing to improve the embeddings. We evaluate characters' morality by applying our trained character embeddings to moral datasets. The results demonstrate that among existing masked language models, BERT model applying additional random masking and fine-tuning using the sentences including main characters only can capture characters' morality well. However, this study has several limitations.

First, we use only one particular book series to evaluate the proposed embeddings' ability to learn morality. And, the Harry Potter story is not reality-based, so the model has limitations in learning morality in actual society. Second, we do not consider the social networks in the story. In general, violent actions are considered immoral; however, such actions against villains can be regarded as ethical. It may be more reasonable to compare the number of ethical decisions among situations than calculating morality score to decide moral characters. However, when we calculate in this way, it is observed that all characters are generally immoral, which is thought that character embedding learns more about the mood of the book. Therefore, we will extend this study by using many datasets and additional methods about the social network to improve the character-centric embedding to represent general characters' morality for our future work.

VI. Acknowledgement

This research was partly supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Korea government (MEST) (No. 2019R1A2C1006316) and Institute of Information communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2019-0-00421, AI Graduate School Support Program), and the MSIT(Ministry of Science and ICT), Korea.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[3] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

[4] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942, 2019.

[5] Jung-Hoon Lee, You-Jin Kim, and Yun-Gyung Cheong. Predicting quality and popularity of a movie from plot summary and character description using contextualized word embeddings. In *2020 IEEE Conference on Games (CoG)*, pages 214–220, 2020.

[6] Danyang Liu, Juntao Li, Meng-Hsuan Yu, Ziming Huang, Gongshen Liu, Dongyan Zhao, and Rui Yan. A character-centric neural model for automated story generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1725–1732, 2020.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[8] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359, 2021.

SUMMARY OF THIS PAPER

A. *Problem Setup*

With the advent of large-scale transformer-based models, word embeddings can learn various contexts. However, there is a lack of research investigating characters' morality, while morality is becoming essential issue as artificial intelligence machines are increasingly intervening human social activities such as counseling, health care, and education.

B. *Novelty*

We propose an morality judgment task using story books in this paper. We present the framework to build embeddings for representing characters to predict the character's morality.

C. *Algorithms*

We preprocess datasets applying co-reference resolution algorithms and extract main characters using occurence. We get character embedding using 3 masked language models and evaluate characters' morality using moral stories datasets.

D. *Experiments*

We use the morality dataset to compute the morality score of each character embedding. We compare the morality scores obtained from character embedding trained by the three embedding models between dataset types and masking methods.