Rewriting Fictional Texts Using Pivot Paraphrase Generation and Character Modification

Dou Liu[1](✉), Tingting Zhu[1], Jörg Schlötterer[2], Christin Seifert[1,2], and Shenghui Wang[1]

[1] University of Twente, Enschede, The Netherlands
{d.liu-2,t.zhu}@student.utwente.nl, shenghui.wang@utwente.nl
[2] University of Duisburg-Essen, Essen, Germany
{christin.seifert,joerg.schloetterer}@uni-due.de

Abstract. Gender bias in natural language is pervasive, but easily overlooked. Current research mostly focuses on using statistical methods to uncover patterns of gender bias in textual corpora. In order to study gender bias in a more controlled manner, we propose to build a parallel corpus in which gender and other characteristics of the characters in the same story switch between their opposite alternatives. In this paper, we present a two-step fiction rewriting model to automatically construct such a parallel corpus at scale. In the first step, we paraphrase the original text, i.e., the same storyline is expressed differently, in order to ensure linguistic diversity in the corpus. In the second step, we replace the gender of the characters with their opposites and modify their characteristics by either using synonyms or antonyms. We evaluate our fiction rewriting model by checking the readability of the rewritten texts and measuring readers' acceptance in a user study. Results show that rewriting with antonyms and synonyms barely changes the original readability level; and human readers perceive synonymously rewritten texts mostly reasonable. Antonymously rewritten texts were perceived less reasonable in the user study and a post-hoc evaluation indicates that this might be mostly due to grammar and spelling issues introduced by the rewriting. Hence, our proposed approach allows the automated generation of a synonymous parallel corpus to study bias in a controlled way, but needs improvement for antonymous rewritten texts.

Keywords: Text rewriting · Gender parallel corpus · Paraphrase generation · Character modification

1 Introduction

A fictional text contains a compelling plot and a series of characters. The description and narration of these characters play a crucial role in the readers' acceptance of the story. If we rewrite a fictional text in which the characters change

C. Seifert and S. Wang—Contributed equally.

K. Ekštein et al. (Eds.): TSD 2021, LNAI 12848, pp. 73–85, 2021.
https://doi.org/10.1007/978-3-030-83527-9_6

their gender[1] or other characteristics, for example, if the Little Red Riding Hood is not a girl but a boy, or Peter Pan is actually disciplined instead of being mischievous, do readers still think that the story is acceptable and fun to read?

Measuring readers' reactions to a gender-parallel story with the same storyline as the original but whose characters have a different gender or opposite characteristics brings an interesting perspective and a controlled way to study gender bias. In order to conduct this type of studies at scale, an automated process that produces such a parallel corpus at a large scale is desirable.

In this paper, we propose a two-step approach for rewriting fictional texts and constructing a parallel corpus. First, a fictional text is paraphrased by a pivot-based paraphrase generation model, to keep the storyline but vary its expression. We apply this step in order to increase the textual diversity in the corpus. In a second step, we modify the gender and other characteristics of the characters using a combination of Named Entity Recognition, Part-of-Speech Tagging, and Character Recognition. We compare four different methods to identify the synonyms or antonyms to describe person characteristics, and perform a technical evaluation as well as a user study to assess the readability of the rewritten texts and readers' acceptance. A web demo, source code and evaluation examples are made available[2].

2 Related Work

Gender bias for characters in literary works is pervasive, but easily neglected or overlooked [6]. Current research on gender bias focuses on statistical semantics of gender-sensitive words or patterns in large scale corpora [3,10,27]. For gender stereotype detection, Cryan et al. [5] demonstrated end-to-end classification approaches to outperform lexicon-based approaches in terms of robustness and accuracy, even on moderately sized corpora. Field and Tsvetkov [9] use propensity matching and adversarial learning to reduce the influence of confounding factors in identifying gender bias in comments. Rudinger et al. [24] rely on a high-quality corpus containing contrast text to detect gender bias in co-reference resolution models. Habash et al. [11] manually created a gender parallel corpus and proposed an approach to de-bias the output of a gender-blind machine translation with gender-specific alternative re-inflections. While the aforementioned methods study or address gender-bias, they typically rely on large-scale (parallel) corpora. Our goal is not to study gender bias itself, but to develop an automatic process for the creation of a gender parallel corpus, which can then be used to study gender bias in a controlled way. That is, we seek a method to rewrite the original text, switching gender and other characteristics.

Most previous work of **text rewriting** (e.g. [12,15,25,28]) has focused on style transfer, e.g., transferring modern English to Shakespearean English [12]. Santos et al. [25] presented an encoder-decoder network to transfer offensive

[1] In this paper, we simplify the notion of gender to the binary concept of male and female.

[2] https://doi.org/10.5281/zenodo.4904849.

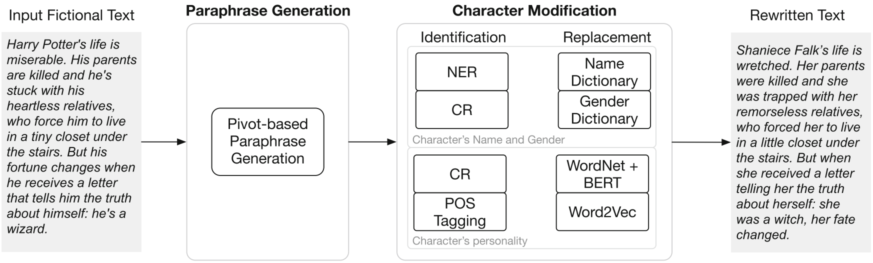

Input Fictional Text
Harry Potter's life is miserable. His parents are killed and he's stuck with his heartless relatives, who force him to live in a tiny closet under the stairs. But his fortune changes when he receives a letter that tells him the truth about himself: he's a wizard.
Paraphrase Generation
Pivot-based Paraphrase Generation
Character Modification
Identification
Replacement
NER
CR
Name Dictionary
Gender Dictionary
Character's Name and Gender
CR
POS Tagging
WordNet + BERT
Word2Vec
Character's personality
Rewritten Text
Shaniece Falk's life is wretched. Her parents were killed and she was trapped with her remorseless relatives, who forced her to live in a little closet under the stairs. But when she received a letter telling her the truth about herself: she was a witch, her fate changed.

Fig. 1. Overview of the fictional texts rewriting model. First, the input text [23] is paraphrased using a pivot language. Second, we detect characters using NER and CR and modify their gender and other characteristics using dictionaries and language models.

language. Woodsend et al. [28] used automatically extracted rewriting rules to increase the amount of labeled data for semantic role labeling. Changing style in texts can be seen as a **Paraphrase Generation** (**PG**) problem, which in turn can be considered a monolingual machine translation problem. Bannard et al. [1] introduced a pivot-based method for PG based on the assumption that two English phrases that translate to the same phrase in a pivot language are potential paraphrases. Moreover, Zhao et al. [29] leveraged multiple MT engines for paraphrase generation and showed that this approach is useful for obtaining valuable candidate paraphrases.

Approaches for **Named Entity Recognition** (**NER**) have transformed from rule-based methods (e.g., [19]) to machine-learning based methods (e.g., [14]) to deep learning methods with automatic feature generation (e.g., [16]). Qi et al. [22] introduced a toolkit for NLP with the state-of-art performance in NER tasks. Similarly, the CoreNLP toolkit [17] encompasses methods for **Coreference Resolution (CR)** with the state-of-the-art performance. The replacement process in our rewriting model requires a notion of **semantic similarity**. Mikolov et al. [20] proposed Word2Vec to obtain a feature space capturing the semantic similarity between words. Faruqui et al. [8] introduced retrofitting word vectors to incorporate knowledge from semantic lexicons and showed that these word vectors perform better on semantic tasks than those word vectors trained without it. As a sub-task of BERT [7], mask-filling can also be used to select the most similar word when several alternative words are given for a sentence.

3 Approach

Figure 1 provides an overview of our approach. The input is a fictional text, e.g., the summary of a published book or a short story written by a human author. Rewriting is then performed in two steps: i) paraphrase generation to express the original text differently and ensure variance in the corpus and ii) character modification, to detect characters and modify their characteristics. Applied to a corpus of fictional texts with controlled modifications, the original texts and

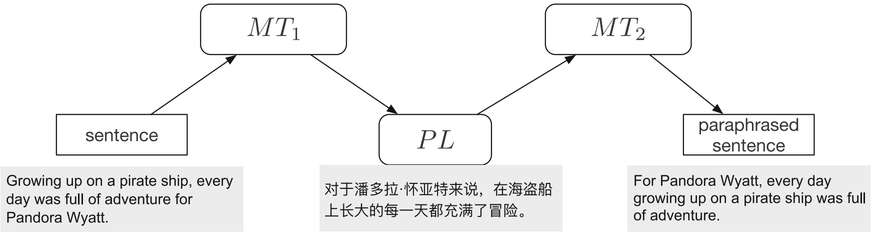

MT_1
MT_2
sentence
PL
paraphrased sentence
Growing up on a pirate ship, every day was full of adventure for Pandora Wyatt.
对于潘多拉·怀亚特来说，在海盗船上长大的每一天都充满了冒险。
For Pandora Wyatt, every day growing up on a pirate ship was full of adventure.

Fig. 2. A single-pivot paraphrase generation system first translates the input sentence [18] to the pivot language (PL) and then back to the original language.

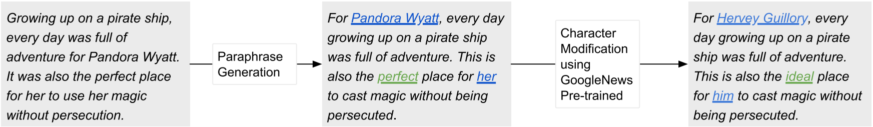
the rewritten ones form a parallel corpus to which readers' preferences can be measured.

3.1 Paraphrase Generation

The paraphrase generation step is applied to increase the textual diversity in the corpus. Conducting a gender bias user study with only changing the fictional characters' gender may easily reveal the study purpose to users and influence their behavior. Further, the original author's writing style could be a confounding factor when studying gender bias. If the parallel corpus is composed of only paraphrased texts, i.e., in a text pair one text that is paraphrased and one that is paraphrased and modified, instead of the original text and a paraphrased and modified version, this factor can be eliminated.

We use a single-pivot paraphrasing system as proposed by Bannard [1]. The basic assumption for pivot-based paraphrase generation (PG) is that two English phrases translated to the same phrase in a pivot language, are potential paraphrases. A single-pivot PG system (cf. Fig. 2), is defined as the triple (MT_1, PL, MT_2), with MT_1 being a machine translation engine which translates a source sentence into the pivot language, PL is the pivot language, and MT_2 a machine translation engine which translates the intermediate sentence in the pivot language back to the source language. The final translation result is the paraphrase of the original sentence. We choose Chinese as the pivot language because Chinese is the mother tongue of the two first authors and therefore the intermediate translations can be manually checked, and Chinese is a gender-neutral language, which prevents introducing additional gender bias during the paraphrasing step. Chinese has gendered pronouns just like English, but does not have grammatical gender in the sense of noun class distinctions.

We use the LingoCloud API [2] for translating the texts from English to Chinese and back, i.e., as MT_1 and MT_2. Figure 2 shows an example for the orginal English sentence, it's paraphrased version and the intermediate Chinese translation.



Growing up on a pirate ship, every day was full of adventure for Pandora Wyatt. It was also the perfect place for her to use her magic without persecution.

For Pandora Wyatt, every day growing up on a pirate ship was full of adventure. This is also the perfect place for her to cast magic without being persecuted.

For Hervey Guillory, every day growing up on a pirate ship was full of adventure. This is also the ideal place for him to cast magic without being persecuted.

Fig. 3. A rewriting example. The original text [18] is paraphrased before the characteristics of the main character are detected and modified.

3.2 Character Modification

After the original text is paraphrased, we modify the characters in the text, by first detecting the characters and subsequently modifying their name, gender and characteristics (cf. Fig. 1). Figure 3 presents the rewriting steps using an example.

To **detect characters** and their mentions, we use Named Entity Recognition (NER) and Coreference Resolution (CR) from CoreNLP [17] and Stanza [22]. An initial comparison on CoreNLP and Stanza's NER performance shows that Stanza provides more accurate results[3], and therefore, we use Stanza to detect the PERSON entities. The CR from CoreNLP is used to detect all mentions of the characters.

To **modify characters**, we use a POS tagger to detect adjectives in the near proximity (at most two tokens away) to the detected characters, assuming that those adjectives describe their characteristics. Once the characteristics are identified, we rewrite the text by switching the character to the opposite gender and rewrite attributes by either using synonyms or antonyms. To change the *gender* of a fictional character, we first determine its gender using dictionaries. For our model, we simplify the notion of gender to a binary attribute[4]. We use a gender-name dictionary[5] and a dictionary with pairs of gender-specific words (e.g. *prince* vs *princess*)[6]. When the gender of the character is determined, we randomly select a common first name with the opposite gender and a common surname from the Surname list[7] or choose the opposite gender-specific to replace the original name: For example, *Princess Diana* is replaced by *Prince Philippe*, or *Harry Potter* with *Shaniece Falk* (cf. Fig. 1). One rewriting option is to replace the identified characteristics (i.e., the adjectives in the near proximity) with their *synonyms*. In other words, similar characteristics are now associated with the opposite gender. We compare the following four methods to replace adjectives:

[3] CoreNLP treats the first name and the last name as two different name entities while Stanza combines them into one.

[4] We will address this limitation in future work.

[5] https://data.world/alexandra/baby-names.

[6] www.fit.vutbr.cz/imikolov/rnnlm/word-test.v1.txt.

[7] https://data.world/uscensusbureau/frequently-occurring-surnames-from-the-census-2000.

- **WordNet:** We select a synonym from WordNet [21].
- **Self-trained Word2Vec (W2V_own):** We train a Word2Vec model using our own corpus of book summaries that is additionally retrofitted with a semantic lexicon as described in [8]. We select the most similar adjective based on semantic similarity.
- **Google News Word2Vec (W2V_google):** We use the Word2Vec model pre-trained on Google News corpus[8] to select the most similar adjective.
- **BERT Mask Filling (BERT):** We consider the adjective as a masked token, and use the prediction of a pre-trained BERT model [7] as replacement.

Another rewriting option is to replace the characteristics with their antonyms, in order to support possible study in gender bias or other fields which need to change the characteristics to the opposite. To replace characteristics with their *antonyms* we follow nearly the same methodology as above except for the following adaptions:

- **W2V_own and W2V_google:** In order to find the opposite adjectives in the vector space, we subtract the adjectives' sentimental polarity (i.e., positive or negative) and add the opposite polarity. For instance, for the positive word *optimistic*, we select the word whose embedding is closest to the vector representing *optimistic - positive + negative*. We use SenticNet [4] to obtain the sentimental polarity of the adjective.
- **BERT:** We combine WordNet with the pre-trained BERT model to select appropriate alternatives to replace the adjectives. First, a set of antonyms of the target adjective is selected from WordNet. With the target adjective masked out, each candidate antonym is ranked by the pre-trained BERT model. The antonym with the highest score, i.e., the one that fits the pre-trained language model best, is chosen as the replacement.

4 Evaluation

We evaluate our approach on a corpus of English fictional texts. To assess the quality of our rewriting model, we evaluate NER and CR separately, assess the readability of the rewritten texts, and conduct a user study to measure the performance of different rewriting methods and the overall acceptance of the rewritten texts.

4.1 Dataset

Our corpus consists of the summaries of 72,487 English books that are catalogued as fictional juvenile works in the WorldCat library catalogue[9]. The average length of the summaries is 216 words, and the vocabulary size is 26,768 words. This dataset is used to train our own Word2Vec embeddings for the character replacement, as mentioned in Sect. 3.2. This corpus also serves as the original fictional texts based on which the parallel corpus is built.

[8] https://github.com/mmihaltz/word2vec-GoogleNews-vectors.

[9] https://www.worldcat.org/.

Table 1. Example errors for NER and CR

Method	Sentence	Error
NER	*...who were the three people she spoke of when* ***Death*** *carried her away' Casey must...*	The person name 'Death' is not identified correctly
CR	*...As* ***she*** *explores the wreckage of* ***her*** *own marriage,* ***Plump*** *offers a beautifully told ..*	Plump, the pronouns 'she' and 'her' refer to the same person 'Plump'. The pronouns are annotated, but 'Plump' is not annotated

4.2 Performance of NER and CR

To evaluate NER and CR, the first two authors independently annotated 20 randomly chosen texts and agreed on a final annotation in case of differences. We obtained F_1 scores for NER on Stanza of 0.953 and CR with CoreNLP of 0.868. The results show that Stanza and CoreNLP perform well on our dataset. Stanza is able to identify persons with high accuracy, while CoreNLP can identify referential relationships between the fictional characters and personal pronouns with reasonable accuracy. However, we observe some errors, as shown in Table 1. Such errors may prevent readers from accepting the rewritten texts, while this is not necessarily related to any implicit biases.

4.3 Text Readability

Ideally, apart from changing the gender and characteristics of the character, the original texts should remain consistent in its story. This also applies to readability; that is, the rewritten text should be at the same or similar readability level as the original one. Therefore, we assess the impact of our model on the readability of the rewritten texts. As readability measures, we use the Flesch-Kincaid Grade Level [13] and the Automated Readability Index [26]. The former score is a linear combination of average sentence length and the average number of syllables of a word, while the latter calculates readability as a linear combination of characters per word and sentence length.

We calculate the readability in terms of school grade level for both, the original and the rewritten texts. Both measures of readability are then averaged for each text. The readability difference is then calculated by subtracting the original text's grade level from that of the rewritten one. We assess how different stages in our model, i.e., the paraphrase generation and different character modification methods, affect the readability. Text readability evaluation is performed on 100 randomly selected texts. The results are shown in Table 2. The metrics are defined as follows:

Diff$_{\textbf{absolute}}$. Averaging each method's absolute differences shows the absolute difference on the readability without considering if the rewritten text is easier or harder for the reader to read.

Table 2. Result of readability evaluation. The *Original–Rewritten* results are between original and rewritten text (either directly after paraphrasing or the final rewritten text generated by different replacement methods). The *PG-CM* results are between the paraphrased text and the final rewritten one.

Indicator	Method									
	Paraphrase generation (PG)	Character modification (CM)								
		Synonymous rewriting				Antonymous rewriting				Mean
		WordNet	W2V_own	W2V_google	BERT	WordNet	W2V_own	W2V_google	BERT	
$\text{Diff}_{\text{absolute}}$ (Original-Rewritten)	1.94	2.01	2.07	1.96	2.27	2.00	2.10	1.98	2.27	2.08
$\text{Diff}_{\text{simplify}}$ (Original-Rewritten)	1.61*	1.08*	1.48*	1.20*	1.84*	1.18*	1.53*	1.33*	1.85*	1.43
$\text{Diff}_{\text{complicate}}$ (Original-Rewritten)	0.98	2.77	1.78	2.27	1.27	2.48	1.72	1.94	1.27	1.94
$\text{Diff}_{\text{absolute}}$ (PG-CM)	-	1.22	0.72	0.96	0.74	1.06	0.69	0.75	0.75	0.86
$\text{Diff}_{\text{simplify}}$ (PG-CM)	-	0.05	0.16	0.06	0.44*	0.06	0.18	0.07	0.45*	0.18
$\text{Diff}_{\text{complicate}}$ (PG-CM)	-	3.50*	1.70*	2.70*	0.90	2.98*	1.52*	2.04*	0.92	2.03

* This indicates that the majority (more than half) of rewritten texts are easier (or more difficult) to read.

$\text{Diff}_{\text{simplify}}$. This only takes into account the rewritten texts that have lower grade level on readability and show the average difference.

$\text{Diff}_{\text{complicate}}$. This indicator calculates the average difference only if the rewriting text has a higher readability level.

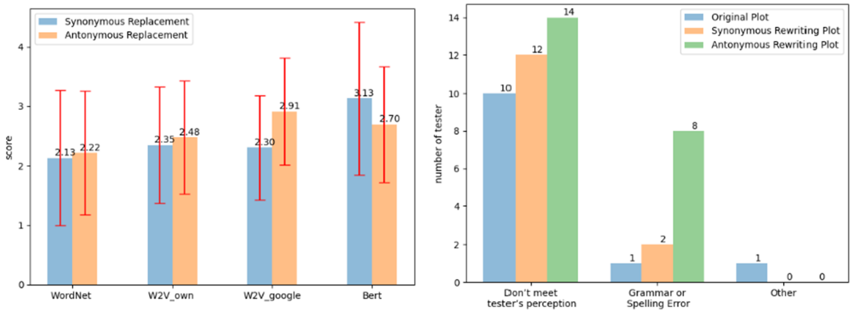
Overall, the readability level of the rewritten texts differs from that of the original by two grades, as indicated by the mean $\text{Diff}_{\text{absolute}}$ (Original–Method). Most rewritten texts are easier to read, as the symbol * is mostly attached to the indicator $\text{Diff}_{\text{simplify}}$ (Original–Method). It also shows that the paraphrase generation step has a more significant influence on readability than character modification. The difference between paraphrased text and character modified text (*PG-CM*) shows that, except for the BERT Mask Filling method, other three replacement methods tend to make the texts more complicated to read. A closer inspection also indicates that the first three methods tend to provide more obscure adjectives as the replacement.

In conclusion, the paraphrasing in our model impacts readability by mostly simplifying the original texts. Synonymous Rewriting and Antonymous Rewriting have no significant difference in terms of the impact on readability. The BERT Mask filling method tends to simplify the texts, while the other three replacement methods do the opposite.

4.4 User Study

A two-stage user study was conducted to evaluate our rewriting methods and the rewritten texts' overall acceptance. We collected human readers' assessments on different rewriting methods applied on individual sentences in the first stage. In the second stage, we investigated human readers' overall attitude towards the synonymously or antonymously rewritten texts. In total, 23 students in the English speaking Master program of the University participated in the user study.

Rewriting Method Evaluation. As described in Sect. 3.2, we implemented four different methods to modify the character's characteristics by replacing corresponding adjectives with their synonyms or antonyms. In the first stage of the user study, we collect human readers' judgments to determine which method produces the most naturally rewritten sentences. We randomly selected two groups of sentences, each containing two individual sentences. For the first group, each sentence was synonymously rewritten while the sentences in the second were antonymously rewritten. During the user study, participants were asked to judge for each of the four rewritten sentences the appropriateness of replacements on a 5-point Likert scale (1: completely disagree, 5: completely agree) and the question is phrased as "Please rate the above sentences based on whether their substitutions(the bold words) are appropriate". In order to avoid bias, the original sentences were not shown to the participants.



(a) Text quality scores.

(b) User Feedback

Fig. 4. User study results. (a) shows the approval scores of the sentences rewritten by different methods (a higher score is better). (b) shows the frequencies of the reasons chosen by the participants when judging a plot is unreasonable.

The user study results are shown in Fig. 4a. For the synonymously rewritten sentences, the BERT method had the highest score. WordNet, W2V_own, W2V_google showed better results for rewriting with antonyms than with synonyms. W2V_google was perceived best for antonymously rewritten sentences. The inferior performance for antonyms vs. synonyms using BERT can be explained as follows: the method first gets candidate antonyms from WordNet before calculating their probability scores; however, some antonyms do not exist in the pre-trained BERT model. This may lead to a less appropriate candidate being chosen as the replacement. The result suggests to use BERT for rewriting using synonyms and W2V_google to rewrite the texts antonymously.

Overall Acceptance. In the second stage, participants were asked to evaluate six fictional texts (two original, two synonymously rewritten and two antonymously rewritten ones). These six texts depict six different storylines. Based on our preliminary manual inspection and the readability evaluation results, we chose the BERT method to modify characters' characteristics for Synonymous Rewriting and Antonymous Rewriting. The six fictional texts were presented to the participants in a randomized order. In detail, the participants were asked "How do you feel after reading?" first and needed to select one level from *reasonable*, *almost reasonable* or *unreasonable* for each text. Additionally, the participants were asked "Why do you think the plot is reasonable or unreasonable?" and needed to select the major reason for their choices or provide their own reasons.

From the feedback collected during this study stage, 73.91% of the participants judged the original fictional texts reasonable or almost reasonable, and 69.57% judged the synonymously rewritten texts reasonable or almost reasonable. This is interesting because it suggests that switching gender barely makes the whole story less reasonable. At the same time, only 52.17% had the same judgments towards the antonymously rewritten texts. The reasons for judging a text not reasonable are provided in Fig. 4b. An almost equal amount of users did not think the texts meet their perception, no matter they were original or rewritten. We do not know the exact reasons for the mismatch between the texts and human reader's perception, but synonymous rewriting does not seem to have an influence. Slightly more users deemed the antonymously rewritten texts unreasonable due to not meeting their perception. Antonymous rewriting is clearly a more difficult task, as many participants found grammar or spelling mistakes in the antonymously rewritten texts. However, these grammar issues are a potential reason why less users deem antonymously rewritten texts reasonable.

4.5 Detecting Spelling and Grammar Issues

The user study reveals a high percentage of antonymously rewritten texts not deemed reasonable, due to grammar issues. Here, we investigate, whether grammar issues can be identified programmatically as a pre-requisite to potentially resolve them and improve antonymous rewriting in future work. We use the open source software LanguageTool[10] containing over 4,800 rules to detect potential spelling mistakes and grammar errors. For a description of the categories please refer to the LanguageTool community[11].

We randomly selected 100 fictional plots from our corpus on which we perform the paraphrase generation and character modification with all eight characteris-

[10] https://pypi.org/project/language-tool-python/, python wrapper for https://languagetool.org.

[11] https://community.languagetool.org/rule/list?lang=en.

Table 3. Spelling and grammar errors of 100 fictional texts at different stages. The first column indicates the stage of the rewriting model and the remaining columns represent categories of spelling or grammar errors defined by **LanguageTool**.

Stage	Misspelling	Typographical	Whitespace	Grammar	Style	Uncategorized	Duplication
Original	401	193	46	22	7	2	0
PG	421	53	44	20	9	0	3
Synonymous rewriting							
W2V_own	769	110	101	68	9	1	20
W2V_google	996	45	102	50	8	0	6
WordNet	933	133	102	48	11	1	1
BERT	650	109	93	10	11	1	2
Antonymous rewriting							
W2V_own	768	111	101	50	10	1	9
W2V_google	833	39	102	54	8	0	1
WordNet	767	130	101	39	9	1	1
BERT	833	107	91	24	9	1	1

tic replacement methods. The results of the spelling and grammar check for each stage (from the original text over paraphrasing to the final rewritten text produced by a particular method) are shown in Table 3. The rather high amount of misspellings (401) in the original text can be explained by the made-up names for the character and locations in the fictional plots. Since many rare names exist in our name dictionary, the number of misspelling errors increases after we replace the character's name during the character modification. The paraphrasing step barely changes the amount of errors, except for the typographical issues, which decrease from 193 to 53. We found many incorrectly opening and closing quotation marks in the original text, an issue the paraphrasing step mitigates.

The character modification part is also found to introduce additional whitespace errors. These are due to tokenization, i.e., treating a single word wrongly as two tokens when splitting a sentence into individual tokens. The increase of duplication errors for W2V_own may be explained by the limited vocabulary of our corpus.

While the text quality evaluation (cf. Sect. 4.4) suggests the use of W2V_google for antonymous rewriting, the spelling and grammar check results confirm our choice of BERT. BERT outperforms the other methods in particular in terms of grammar issues, both for synonymous and antonymous rewriting. Still, the amount of grammar issues in antonymous rewriting (24) is more than twice as high as in synonymous rewriting (10), confirming the results from the user study. Yet, these results suggest that grammar issues can be identified programmatically, providing an opportunity to also resolve them programmatically in the future.

5 Conclusion

In this paper, we proposed a fiction rewriting model to build a parallel corpus for future studies of gender bias. Our model combines paraphrase generation with character detection and characteristic replacement. The evaluation shows that, compared to the original fiction, the user acceptance of synonymously rewritten texts is roughly on par with the original, while antonymously rewritten texts perform worse. An analysis shows that rewriting with antonyms tends to generate unnatural texts or introduce grammatical mistakes. In conclusion, synonymously rewritten fictional texts produced by our approach can be deemed suitable for building a gender parallel corpus, while antonymous rewriting needs some future improvements, e.g., an automatic grammar correction. Future work could investigate the alteration of further aspects of the stories (e.g., location, events) and investigate their influence, as well as improve the naturalness of the rewriting method, especially for the antonymous rewriting.

References

1. Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 597–604 (2005)
2. CaiyunWiki: LingoCloud API in 5 minutes (2020). https://open.caiyunapp.com/LingoCloud_API_in_5_minutes. Accessed 01 June 2020
3. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
4. Cambria, E., Li, Y., Xing, F.Z., Poria, S., Kwok, K.: Senticnet 6: ensemble application of symbolic and subsymbolic AI for sentiment analysis. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management, pp. 105–114 (2020)
5. Cryan, J., Tang, S., Zhang, X., Metzger, M., Zheng, H., Zhao, B.Y.: Detecting gender stereotypes: lexicon vs. supervised learning methods. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–11 (2020)
6. Davis, E.: The physical traints that define men and women in literature. https://pudding.cool/2020/07/gendered-descriptions/
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E., Smith, N.A.: Retrofitting word vectors to semantic lexicons. In: Proceedings of NAACL (2015)
9. Field, A., Tsvetkov, Y.: Unsupervised discovery of implicit gender bias. arXiv preprint arXiv:2004.08361 (2020)
10. Garg, N., Schiebinger, L., Jurafsky, D., Zou, J.: Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc. Nat. Acad. Sci. **115**(16), E3635–E3644 (2018)
11. Habash, N., Bouamor, H., Chung, C.: Automatic gender identification and reinflection in Arabic. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pp. 155–165 (2019)
12. Jhamtani, H., Gangal, V., Hovy, E., Nyberg, E.: Shakespearizing modern language using copy-enriched sequence-to-sequence models. arXiv preprint arXiv:1707.01161 (2017)

13. Kincaid, J.P., Fishburne, R.P., Jr., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report, Naval Technical Training Command Millington TN Research Branch (1975)
14. Klein, D., Smarr, J., Nguyen, H., Manning, C.D.: Named entity recognition with character-level models. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4, pp. 180–183. Association for Computational Linguistics (2003)
15. Lample, G., Subramanian, S., Smith, E., Denoyer, L., Ranzato, M., Boureau, Y.L.: Multiple-attribute text rewriting. In: International Conference on Learning Representations (2019). https://openreview.net/forum?id=H1g2NhC5KQ
16. Li, J., Sun, A., Han, J., Li, C.: A survey on deep learning for named entity recognition. IEEE Trans. Knowl. Data Eng., 1 (2020). https://doi.org/10.1109/TKDE.2020.2981314
17. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford CORENLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60 (2014)
18. McCollum, H.: Tangled Hearts. Highland Hearts, Entangled Publishing, LLC (2014). https://books.google.nl/books?id=XcRRAQAAQBAJ
19. Mikheev, A.: Automatic rule induction for unknown-word guessing. Computat. Linguist. **23**(3), 405–423 (1997)
20. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
21. Miller, G.A.: Wordnet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
22. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082 (2020)
23. Rowling, J., GrandPré, M.: Harry Potter and the Sorcerer's Stone. Grades 7–9, A.A.L. Books Inc., Series. A.A. Levine Books, Hoboken (1998). https://books.google.nl/books?id=dmouxgEACAAJ
24. Rudinger, R., Naradowsky, J., Leonard, B., Durme, B.V.: Gender bias in coreference resolution (2018)
25. Santos, C.N.D., Melnyk, I., Padhi, I.: Fighting offensive language on social media with unsupervised text style transfer. arXiv preprint arXiv:1805.07685 (2018)
26. Senter, R., Smith, E.A.: Automated readability index. CINCINNATI UNIV OH, Technical Report (1967)
27. Sun, T., et al.: Mitigating gender bias in natural language processing: Literature review. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1630–1640. Association for Computational Linguistics, Florence, Italy, July 2019. https://doi.org/10.18653/v1/P19-1159, https://www.aclweb.org/anthology/P19-1159
28. Woodsend, K., Lapata, M.: Text rewriting improves semantic role labeling. J. Artif. Intell. Res. **51**, 133–164 (2014)
29. Zhao, S., Wang, H., Lan, X., Liu, T.: Leveraging multiple MT engines for paraphrase generation. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 1326–1334 (2010)