Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2188–2195
Marseille, 20-25 June 2022

Leveraging Pre-trained Language Models for Gender Debiasing

Nishtha Jain[1], Maja Popovic[1], Declan Groves[2], Lucia Specia[3]
[1]ADAPT Centre, [2]Microsoft, [3]Imperial College London
[1]Trinity College Dublin, Ireland, [1]Dublin City University, Ireland, [2]Dublin, Ireland, [3]London, UK
[1]{firstname.lastname}@adaptcentre.ie, [2]declan.groves@microsoft.com, [3]l.specia@imperial.ac.uk

Abstract

Studying and mitigating gender and other biases in natural language have become important areas of research from both algorithmic and data perspectives. This paper explores the idea of reducing gender bias in a language generation context by generating gender variants of sentences. Previous work in this field has either been rule-based or required large amounts of gender balanced training data. These approaches are however not scalable across multiple languages, as creating data or rules for each language is costly and time-consuming. This work explores a light-weight method to generate gender variants for a given text using pre-trained language models as the resource, without any task-specific labelled data. The approach is designed to work on multiple languages with minimal changes in the form of heuristics. To showcase that, we have tested it on a high-resourced language, namely Spanish, and a low-resourced language from a different family, namely Serbian. The approach proved to work very well on Spanish, and while the results were less positive for Serbian, it showed potential even for languages where pre-trained models are less effective.

Keywords: gender debiasing, language generation, pre-trained language models

1. Introduction

Gender bias in language has increasingly become an important topic of research in natural language processing (NLP). Although NLP models are successful in modelling various applications, they propagate and may even amplify gender biases found in the training sets. While the study of bias in artificial intelligence is not new, techniques to mitigate gender bias in NLP are relatively nascent (Sun et al., 2019). This paper proposes a new method to reduce gender bias by enriching existing data with gender variants. These variants can be used either directly, for example, in the context of language generation applications to provide multiple gender-marked outputs, or to create gender-balanced corpora that can in turn be used as training data for NLP models.

More specifically, the approach is devised to process either original or automatically generated sentences in gender-marked languages such as Spanish, where grammatical gender is expressed by morphology (see Section 3) to provide its gender variants. The method does not address only gender of persons but all instances of grammatical gender (persons, animals, objects). For example, given the sentence in Spanish "¡Gracias, querid**a**!" ("Thank you, darling!"), which uses the feminine noun "querid**a**", the goal is to generate the masculine counterpart: "¡Gracias, querid**o**!". Conversely, given the sentence (also in Spanish) "Enviad**o**." ("Sent."), which uses the masculine version of the past-participle verb "send", the goal is to generate its feminine counterpart "Enviad**a**.".

Our approach is inspired by work in the area of text infilling (Zhu et al., 2019), the process of finding suitable fill-in-the-blank words for a text where some words are missing, given their sentential context. The idea is that the context will be telling which words could fill the blanks. This method has applications in areas such as historical document restoration, article writing, and text editing. We propose to use such technique for paraphrasing gender-marked words in a sentence, i.e. we remove such words from a sentence and find replacements for them, which are then further filtered to ensure they are paraphrases that only vary in gender. The main challenges in this approach are to (1) select words whose grammatical gender can be changed, (2) find appropriate variants in context, and (3) ensure sentence cohesion when multiple words can be changed. We describe how we address these challenges in Section 4. We test this approach on a high-resource language (Spanish) as well as a low-resource language (Serbian) using two corpora of sentences with different levels of linguistic complexity (Section 5) and discuss the performance we obtain on both in Section 6.

2. Related Work

There has been significant work in the field of mitigating gender bias, however most of the approaches address this problem from an algorithmic perspective. There has been less research in the direction of this work, namely generating gender variants as the final output and/or using them as a data augmentation strategy for debiasing models.

(Sun et al., 2019) describe various techniques that can be used to mitigate gender bias in NLP, mainly grouped in two strategies: 1) debiasing using data manipulation and 2) debiasing by adjusting algorithms. Manual data augmentation has been widely used as a technique to add data to a gender-unbalanced corpus, however it is expensive, especially for large datasets.

In the area of machine translation, (Vanmassenhove et al., 2018) propose an approach for mitigating gender

bias by gender tagging, where the original sentence in the source language is tagged for speaker/author gender. Such tagging usually helps the machine translation model to select the intended gender of a speaker in the translation. However, when translating from gender-unmarked languages, it requires meta information such as gender of author of the text, which is not always available.

The work in (Zmigrod et al., 2019) is based on counterfactual data augmentation (CDA), which considers gender of animate occupational nouns from a prepared list. For each such noun, the gender is swapped. The morpho-syntactic tags for the rest of the tokens in the sentence are calculated using Markov random fields to generate meaningful alternatives for the words related to this noun (articles, adjectives). The approach is quite effective for sentences that contain this type of nouns, however it cannot handle more generic sentences that contain adjectives or other gendered words as it depends on lists with gender variants and in order to cover more gender phenomena by this method, the lists would become too large.

(Habash et al., 2019) introduces gender variant generation for Arabic sentences containing only first person personal pronouns. The approach consists of a machine learning-based classifier to classify a given sentence as masculine (requiring feminine variant), feminine (requiring masculine variant) or neutral (not requiring any variant). Subsequently, a tag is added to the sentence to inform the type of sentence, as in (Vanmassenhove et al., 2018). For each sentence, a gender variant is generated using a neural machine translation (NMT) rewriter trained on a manually created gender parallel corpus consisting of original sentences and their gender variants. This approach highlights the utility of NMTs in this task, however it requires a large amount of training data for both the gender classifier and the NMT rewriter. Also, this approach has only been tested on first-person pronouns.

(Jain et al., 2021) follows a similar line of work for Spanish, also using an NMT rewriter for generating gender variants. However, they avoid the cost-effective parallel corpus creation process by proposing an automated way to create such corpus. For that, they use language-specific rules designed to extract a set of sentences with a pre-defined syntactic structure (for example "VERB ADVERB ADJECTIVE") and apply gender transformation to the gendered words in those sentences. This method is effective, however not scalable because the approach is unable to handle any sentences which do not follow those specific syntactic structures covered by the set of rules.

Our work aims to design an automatic approach that (1) can cover gender-marked words with any POS tag, not only for a limited set (Zmigrod et al., 2019; Habash et al., 2019), (2) does not depend on a fixed set of rules as in (Jain et al., 2021) and is therefore scalable, and (3) does not require parallel corpora as in the automatic NMT rewriters developed in (Habash et al., 2019; Jain et al., 2021).

3. Gender-related Language Properties

In both languages explored in this work, grammatical gender is expressed by morphology. However, there are several differences between the two languages.

In Spanish, apart from animate nouns, the following POS classes have grammatical gender: adjectives, pronouns, as well as past participles when acting as adjectives, however not when they are forming past tense. In addition, pronouns are sometimes attached to verbs as suffix (e.g. "a ver" = let me see, "a verlo" = let me see it). Gendered words in Spanish can be easily distinguished by suffixes, for example "cansado/cansada" (tired), and the only additional inflection which can interfere is number, namely singular or plural ("cansados/cansadas"). Both types of suffixes are relatively straightforward: genders are distinguished by "o/a", "e/a" or adding "a", while plural is denoted by adding suffix "s".

This is, however, different in Serbian. While the gender is also expressed via suffix, there are several additional factors which have to be taken into account. First, in addition to gender and number, case is also expressed through inflection so that many different forms can interfere with the gender form. Also, the definition of plural is more complex, involving several possible suffixes without any specific rules so that it can also interfere with gender. For example, the word "druga" (feminine "second" or "other") can be singular feminine nominative but also plural masculine genitive, while its opposite gender variant "drugi", besides singular masculine nominative, can be plural masculine nominative. Also, past participles are gendered even when they are forming past tense.

Finally, different from Spanish, there is a third gender in Serbian, namely the neuter gender, which might interfere with other forms as well. This gender category[1], also existing in other languages such as German, is predominantly used for objects and general concepts, and it is neither masculine nor feminine. There is, however, no major reason to include neuter gender into providing gender variants because it is not used for people (except for kids) and it is often used for general concepts where gender variants would not make sense. For example, in the sentence "it is nice to see you", the adjective "nice" has neuter gender and there is no possible alternative variant.

4. Methodology

This section explains the core approach initially developed for Spanish and subsequently minimally adapted and tested for Serbian (in Section 4.1).

[1]This is not to be confused with neutral sentences mentioned later in the paper. Neutral sentences in the context of providing gender variants, are sentences that do not require a gender variant and should be left unchanged.

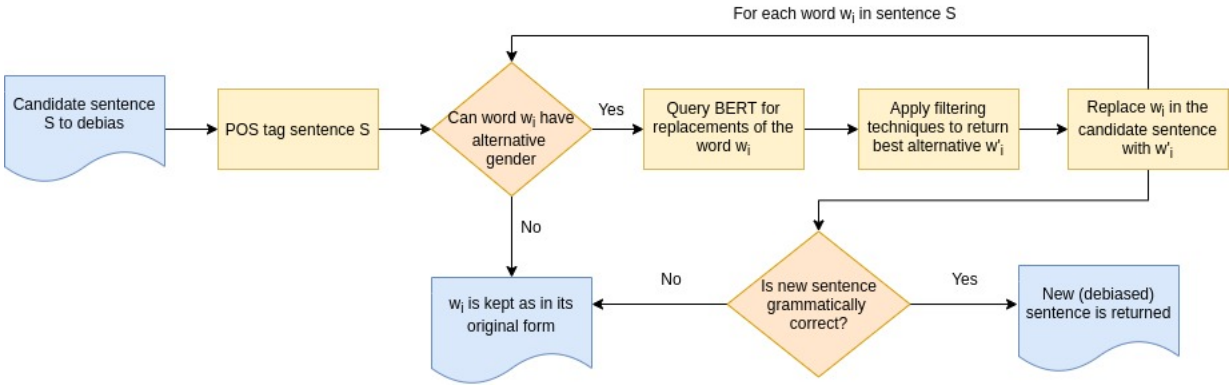

Figure 1: Methodology: Given a candidate sentence, it is first POS tagged. Words that have gender marking based on the morphological information returned by the POS tagger are then masked and BERT is queried to return replacements for that word in the context of the original sentence. Applying filtering techniques, the best alternative suggested by BERT is used to create the gendered variant. This process is repeated for each word that has gender marking. The sentence is finally checked for grammaticality. If the resulting sentence is ungrammatical, the original sentence is kept.

Figure 1 shows our methodology developed to generate gender variants. Its core is inspired by the idea of text infilling (Zhu et al., 2019), but with a specific goal and a few additional steps. More specifically, it works in the following way: a text is first POS-tagged using a state-of-the-art POS tagger, namely Stanza (Qi et al., 2020), which also produces morphological information. This step marks words with gender information (any POS tag) if they exist in the sentence. Once the word(s) and its position(s) within the sentence are identified, each particular word is masked (i.e. replaced by a placeholder symbol) to create the input for a pre-trained language model, namely BERT (Devlin et al., 2019)[2]. We then query BERT as a masked language model having as input the full sentence as context, with one masked word at a time. The top 100 words returned are considered as candidates to replace the masked token, much like a fill-in-the-blank task. In case a sentence consists of multiple gendered words whose variants need to be generated, this is done recursively by replacing the words in their original order.

As a **baseline**, the masked token is replaced with the first of the 100 alternatives returned by BERT. For our final approach, we add two main steps: a **filtering** step to select the best amongst the 100 alternatives, and an overall sentence **grammar check** step to decide if the best replacements as selected by the previous steps make up for a grammatical sentence.

We devised the following filtering techniques:

- **POS tag-based filtering**: This technique requires POS tagging all alternatives returned by BERT and selecting the ones with the same POS and all morphological features (number, tense, case, etc.) as the original word, except for gender, where we flip masculine to feminine and vice-versa.

  Since BERT generates around 100 variants for each sentence, and each of those have to be POS tagged, POS-basd filtering is computationally expensive and not optimal for real-world applications.

- **Normalised character-level edit distance ranking (ccer)**:

  Since POS-based filtering is computationally expensive when selecting the best variant, we investigated alternative approaches based on edit-distance. The idea is based on the fact that the gender in both languages is expressed through morphology – the assumption is that the string representing the replacement has to be very close to the original string in terms of its characters. We use several variants of normalised character-based edit distance in order to rank the possible replacements and choose the one with the shortest distance:

  – *Alternative character error rate (acer)* the number of changed characters normalised by the number of characters in the candidate alternative word:

  $$acer = \frac{editDistance}{len(altWord)}.$$

  – *Original character error rate (ocer)* the number of changed characters normalised by the number of characters in the original word:

  $$ocer = \frac{editDistance}{len(origWord)}.$$

[2]Specifically: `https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased`

– *Combined character error rate (ccer)* the number of changed characters normalised by the number of characters in both original and alternative words:

$$ccer = \frac{2 \times editDistance}{leng(altWord) + len(origWord)}.$$

- **Length and prefix penalty (ccer$^+$)** : This was introduced to penalise (a) any changes in length between original and alternative words, (b) alternative words whose first letter is different from the first letter in the original word (gender changes are only expressed via suffices in Spanish). In either case, we increment the edit distance by 2.

- **"Lo/la" interchanging**: for the masculine pronoun "lo" and its feminine alternative "la", BERT often proposes the neutral pronoun "le" which cannot be filtered out by previously described methods. Therefore, we applied additional "brute-force" filtering to always interchange between those two words instead of exploring other alternatives.

- **LanguageTool API**[3]: To avoid potentially ungrammatical replacements generated despite the filtering techniques, we use LanguageTool API to check a sentence for various grammatical errors: if there is any error in the regendered sentence, we revert it to the original sentence. Any grammar checker could be used in this step.

4.1. Serbian

This section presents additional work needed and the challenges in porting the approach initially proposed for Spanish to a very different language, from a different language family - namely Serbian. Our claim is that the core of the methodology can be kept as is (POS tagger followed by masked language model), but the filtering heuristics need adaptation (requires some knowledge of the language). We note, however, that these are mostly very general heuristics, rather than rules as in (Jain et al., 2021). However, we also recall that our core text infilling approach assumes two resources: a POS tagger and a BERT-like model, the quality of which will have an impact on the results. Serbian brings challenges in both of these directions:

POS tagger We used Classla - a fork of Stanza adapted for better performance on Serbian and other languages (Slovenian, Croatian, Macedonian and Bulgarian). On manual inspection, the quality of the POS tagger seemed good, except for a few cases where it was difficult to disambiguate some words in different contexts, and thus incorrect variants were generated. For example, for the word 'druga', the tag is predicted as "singular, feminine, nominative", which is correct in some contexts, but in the given context the correct tag is "plural, masculine, genitive".

BERT model We had to resort to a multilingual version of BERT[4]. We tried one built specifically for Serbian[5], BERTić (Ljubešić and Lauc, 2021), but the training objective for this model was different: instead of training a language model to predict masked words, it was trained to predict whether a word was the same as in the original text or replaced. Therefore, it was not possible to use it for providing different gender variants.

The multilingual BERT was built using data from Wikipedia for 104 languages, and we assume Serbian to have had relatively small coverage. That means that a lot of words we attempt to find replacements for either (i) do not exist in the model at all, (ii) are segmented into sub-word units (which brings issues to the approach as discussed in previous sections), or (iii) come from corpora of other, related languages, like Czech or Polish, which results in incorrect replacements.

The described challenges, together with the differences in language properties described in Section 3, allowed us to test the potential of the approach in a much different setting than Spanish which is considered a high-resourced language.

5. Test sets

The approach proposed in this paper does not require any task-specific labelled training set. In our experiments, we use different kinds of test sets to evaluate its performance containing a mix of "neutral" sentences, for which no gender variants are possible, and "regenderable" sentences, which contain one or more words for which gender variants are possible. For each sentence in our test set, a parallel sentence was manually generated as the gold-standard, consisting of either of the regendered variant of a regenderable sentence or a copy of the original sentence for neutral sentences. The statistics of our test sets can be seen in Table 1.

For **Spanish**, we used three test sets: SPANISH 1, SPANISH 2 and SPANISH 3. The first two sets, SPANISH 1 and SPANISH 2, were provided to us by our industry partner Microsoft, initially as one single set

SPANISH 1: this test set was extracted from the initial set using the rules from (Jain et al., 2021). It contains sentences with a shorter length and at most one word which has a possible gender variant. In this test set, the number of regenderable sentences is slightly higher than the number of neutral sentences.

SPANISH 2: This test set consists of the rest of the sentences from the initial set that do not conform with the rules. Therefore, it cannot be processed by the rule-based approach because the rules would fail. The set contains slightly longer sentences and a much larger

[3] https://dev.languagetool.org/public-http-api

[4] https://huggingface.co/bert-base-multilingual-cased

[5] https://huggingface.co/classla/bcms-bertic

# Sentences	Spanish			Serbian
	1	2	3	
All	5,648	17,249	3,066	1,921
Regenderable	3,344	2,752	1,777	1,586
1 word	3,344	2,602	-	1,295
2 words	0	134	-	265
3 words	0	16	-	24
4 words	0	0	-	2
Neutral	2,304	14,497	1,289	335
# Words	25,605	95,755	12,995	12,334
words/sent	4.53	5.55	4.23	6.42

Table 1: Data set statistics: number of sentences & words, and average number of words per sentence.

proportion of neutral sentences. Also, in a number of regenderable sentences, more than one word can have a gender variant. Therefore, it is a more challenging test set than SPANISH 1.

The third test set, SPANISH 3, has been extracted from the publicly available monolingual OpenSubtitles (Jörg Tiedemann, 2012) corpus again using the rules from (Jain et al., 2021). It majorly consists of shorter sentences with a slightly higher number of regenderable sentences as compared to neutral sentences. This test set will be released publicly to make the research carried out in this paper reproducible.

For **Serbian**, no test sets were readily available, so we created one from the monolingual OpenSubtitles (Jörg Tiedemann, 2012) corpus. It consists of longer sentences than Spanish, with a significantly higher number of regenderable sentences than neutral ones, and those sentences contain up to four regenderable words. No rules were used to create this test set, it was only manually inspected to check and remove possible errors. This test set is a much more challenging test set than the Spanish ones, both because of its structure as well as because of the differences between language characteristics. This test set will also be publicly available.

In all test sets, the majority of regenderable sentences contains only one regenderable word.

6. Results

In this section we present the results for Spanish and Serbian in the form of word-level accuracy, namely the proportion of all words which match the words in the gold-standard of the gender variant: for neutral sentences, it is identical to the original sentence, and for regenderable sentences, it contains correct gender variants for all regenderable words. The word-level accuracies are presented in Table 2 for all sentences, as well as separately for neutral and regenderable sentences. We discuss these for each language in what follows, along with a qualitative manual analysis of outputs in order to identify the main problems and challenges for each of the languages.

6.1. Spanish

In Table 2, it can be noted that the baseline model, i.e. replacing the original word with the first alternative returned by BERT performs as the worst on both the test sets overall and on regenderable sentences, while for neutral sentences the accuracy is quite high. This could be expected, as without any filtering, the method often chooses the alternative word identical to the original one, which is beneficial for neutral sentences.

POS-based filtering improves the accuracy of the regenderable sentences, however deteriorates neutral sentences in the first set. Nevertheless, since POS-based filtering is computationally too expensive for real-time applications, as mentioned in Section 4, we attempted to replace it by edit-distance based filtering.

The three normalisation variants for edit-distance yielded very similar accuracies for SPANISH 1, although the *ccer* variant (normalising over both original and alternative word length) seems to be the best. Therefore, we used only this variant SPANISH 2 and 3 as well as for Serbian.

Compared to POS-based filtering, edit-distance filtering improves accuracy for regenderable sentences while deteriorating neutral sentences. This can be expected, since without comparing POS tags and performing morphological analysis of the original and alternative words, a number of words in neutral sentences are replaced by some similar words, such as sentence 1) in Table 3: the noun "casa" (house) is replaced by another noun "cosa" (thing) only because the edit distance is small.[6]

Another observed problem is replacing singular by plural and other way round instead of the gender change (example 2 in Table 3), because edit distance is the same both for number and for gender variant. Therefore, penalising edit distance for different word lengths as well as for different prefixes $ccer^{+}$ was introduced, which further improved the accuracy on both test sets and both sentence types (Table 2).

The problem with pronouns "lo" and "la" can be seen in sentence 3) in Table 3: the masculine variant "lo" is replaced by the neutral pronoun "le" instead of its feminine variant "la". Since it cannot be distinguished even by improved edit distance $ccer^{+}$, "brute-force" pronoun interchanging was applied, which notably improved the performance on regenderable sentences without much impact on neutral ones. The main reason for this large improvement is the very high number of sentences containing these pronouns in both test sets.

Sentences 4) and 5) in Table 3 illustrate two more frequent problems, namely generating non-existing words (4) and removing accents from words (5). Both these problems were improved by adding the language tool on the top, obtaining higher accuracy both on regenderable and especially on neutral sentences.

[6] "casa" and "cosa" are both female gender marked nouns and thus not variants of each other.

Word level accuracy [%]	All	Regenderable	Neutral
Spanish 1			
rule-based (Jain et al., 2021)	**99.3**	**99.3**	**100**
baseline	84.0	74.3	96.0
POS tagger	88.9	84.0	94.8
acer (normalising over alternative word length)	88.1	85.6	91.0
ocer (normalising over original word length)	88.2	85.7	91.2
ccer (normalising over both)	88.3	85.8	91.3
ccer$^+$ (with penalising length and prefix)	90.2	88.6	92.3
ccer$^+$ + "lo/la" pronoun interchanging	92.7	93.2	92.1
ccer$^+$ + "lo/la" pronoun interchanging + language tool	**94.8**	93.3	**96.5**
ccer$^+$ + "lo/la" pronoun interchanging + POS tagger	94.6	**94.2**	95.1
Spanish 2			
rule-based (Jain et al., 2021)	rules fail (not applicable)		
baseline	93.2	78.2	96.0
POS tagger	**95.9**	79.8	**98.9**
ccer	84.9	83.4	85.1
ccer$^+$	88.8	88.9	88.8
ccer$^+$ + "lo/la" pronoun interchanging	89.2	92.2	88.6
ccer$^+$ + "lo/la" pronoun interchanging + language tool	94.7	92.1	95.1
ccer$^+$ + "lo/la" pronoun interchanging + POS tagger	92.9	**93.3**	92.8
Spanish 3			
rule-based (Jain et al., 2021)	99.6	99.3	100.0
baseline	82.1	72.1	**93.8**
ccer	76.1	75.8	76.4
ccer$^+$	84.9	84.4	85.4
ccer$^+$ + "lo/la" pronoun interchanging	87.3	**90.1**	84.1
ccer$^+$ + "lo/la" pronoun interchanging + language tool	**92.1**	89.1	**95.5**
Serbian			
baseline	**84.5**	81.5	**99.5**
ccer$^+$	80.7	78.6	91.5
ccer$^+$ + POS tags	83.2	80.0	**99.3**
ccer$^+$ + POS tags for pronouns only	84.2	**81.8**	96.3

Table 2: Performance as word-level accuracy, comparing against previous work for Spanish, the baseline (no filtering), and the various filtering techniques.

Compared to the set-up containing both POS-based and edit distance filtering, as it can be seen in the last row of Table 2, the performance is similar to that of the language tool. It should however be noted that using a POS tagger is more computationally expensive than using the language tool API.

Compared to previous work (Jain et al., 2021), we obtain lower performance on SPANISH 1 and SPANISH 3. However, as explained in Section 5, these test sets were explicitly designed to conform with these rules, therefore they perform very well. On the other hand, the second test set SPANISH 2 cannot be processed at all by this approach since the syntactic structures of the sentences do not conform to the structures that the rules can be applied to. This illustrates the scalability of the approach described in this work.

All in all, the main problems for Spanish regenderable sentences are the words which remained unchanged. Apart from this, there are undesired conversions of the verb tense/person/mood, number (incorrect word changed) replacing letters with an accent by same letters without an accent, and also creating non-existing words.

As for neutral sentences, the main problem are unnecessary gender-related changes contributing to more than a half of all errors. Also, some words are converted into other tense/number/etc, or even into a non-existing word.

6.2. Serbian

In Table 2, it can be seen that the baseline system for Serbian performs fairly well, especially on neutral sentences. The first step further was to use the best version of edit-distance filtering designed for Spanish, but without the language tool (because it is not available for this language) and the rule for "lo-la" (because it is Spanish-specific). This method, however, deteriorated the accuracy because it resulted in a number of non-existing words. One reason is a large number of original words which were probably not seen in the training of BERT and therefore became segmented into subword units, and then alternatives were proposed for incorrect parts of the word. Another reason is that sometimes alternatives were not only unrelated to gender but were not even part of the Serbian language (some of them even contained characters from other languages,

original	output+issue type	correct
1) la cosa esta bien.	la **casa** esta bien. (unwanted lexical change)	la **cosa** esta bien.
2) son **bienvenidos**	son **bienvenido** (plural to singular) (improved by penalised edit distance $ccer^{+}$)	son **bienvenidas**
3) ahora **lo** entiendo.	ahora **le** entiendo. ("lo" converted to neutral "le" instead of feminine "la") (solved by "lo/la" interchanging)	ahora **la** entiendo.
4) ahora mismo **la** he enviado .	ahora **misma la** he **enviada** . (incorrect words changed)	ahora **mismo lo** he **enviado** .
5) infórmenos	**infórmenov** (non-existing word) (improved by language tool)	**infórmenos**
6) ¡comprobémoslo!	¡comprob**e**moslo! (removed accent) (improved by language tool)	¡comprobémoslo!

Table 3: Spanish examples comparing the generated output with the correct output to highlight the difference

however those were successfully filtered out by the edit distance).

Another problem with edit distance are personal pronouns: in Serbian, they definitely cannot be handled solely by edit distance and similar length because the gender variants are often more distant than in Spanish ("on-ona" (he/she, nominative) "njega-nje" (genitive), "mu-joj" (dative), "ga-je", "njega-nju" (accusative), "njim-njom" (instrumental), "njemu-njoj" (locative)). Also, suffix conversions are more complex than in Spanish (for example "bio/bila", "rekao/rekla", (past participles) "potreban/potrebna" (adjective), etc.). In addition, edit distance to other options such as plural or neuter gender is often smaller so that this variant becomes selected instead of the correct one. For example, feminine past participle "bila" (been) is closer to the neuter singular variant "bilo" and three plural variants "bili, bile, bila" than to its opposite gender variant, masculine singular past participle "bio".

Similar challenges occur with the adjectives, e.g. feminine version of "needed", "potrebna", is closer to neuter singular and all plurals "potrebno, potrebni, potrebne, potrebna" than to the desired gender variant "potreban".

For these reasons, and also because no language tool or grammar checker is available for this language, we applied the combination of edit distance and POS filtering, which resulted in better performance compared to the edit distance alone, especially for neutral segments. Still, the accuracies are slightly below the baseline due to a number of non-existing words. Another problem is that some of the forms (cases) of personal pronouns do not have any character in common (such as "mu/joj", "ga/je") and therefore normalised edit distance is maximal possible, 100%, we tried the third option: removed edit distance for pronouns, and retain it for other word classes. This method improved the overall accuracy by improving regenderable segments but deteriorated the quality of neutral segments because again it allowed more incorrect alternatives.

As for problems in neutral segments, they are similar as in Spanish texts: non-existing words are frequent, especially in the first variant (edit distance without POS), and different undesired changes can be found (related to gender, case, or number).

Some examples of different types of remaining challenges can be seen in Table 4.

The lower quality of the Serbian regendering process is due to two main factors: language properties (as noted above) and resources. For resources, contrary to Spanish BERT, multilingual BERT was trained on multiple languages, hence the alternatives were suggested from a common vocabulary of 104 languages (therefore segmenting a large number of words and not proposing reasonable alternatives). Also, there is no readily available language tool for Serbian for additional filtering of ungrammatical segments.

7. Conclusions and Future Work

The proposed approach performs quite well on the Spanish datasets, both simple and complex, with some very specific errors as shown in Table 3. A low-resourced, morphologically more complex language, namely Serbian, proved to be more challenging, as discussed in Section 6, mainly due to the lower quality of the POS tagger and the BERT model.

The main advantage of the approach is that it does not require any task-specific supervision. Also, it requires minimal language-specific heuristics which is transferable but requires some knowledge of the language. It thus provides an automatic way for generating gender variants using good pre-trained language models like BERT, followed by simple filtering strategies and, ideally, a grammar checker. This approach can be used,

original	output+issue type	correct
a **drugi** ?	a **drugi** ? (unchanged)	a **druga** ?
a baš je tada **otišao** kući ?	a baš je tada **otišlo** kući ? (neuter gender)	a baš je tada **otišla** kući ?
a **druge dve** da ostavimo ?	a **drugi** dva da ostavimo? (gender variant but for singular instead of plural)	a **druga dva** da ostavimo?
a jesi li i ti **bio** ?	a jesi li i ti **bili** ? (gender unchanged, singular instead of plural)	a jesi li i ti **bila** ?
a onda je ona **sišla** dole	a onda je on **sišila** dole (non-existing word)	a onda je on **sišao** dole
baš su **lepe** i **slatke** .	baš su **leps** i **slatni** . (non-existing words)	baš su **lepi** i **slatki** .

Table 4: Serbian examples comparing the generated output with the correct output to highlight the difference

among other things, to create a gender-balanced corpora which can subsequently be used to train different types of NLP models to reduced gender bias in their predictions.

Core directions for future work include using better pre-trained models such as XLM-R and more research into LM-based filtering, including purposely built LMs.

Future work could also include more research into how the approach generalises across different languages within the same family, e.g. Romance languages, versus languages in different families, such as Slavic languages, especially when it comes to the linguistic heuristics. It can also be the case that this approach would not apply to certain languages where gender is not given by POS tags and morphological features.

8. Bibliographical References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Habash, N., Bouamor, H., and Chung, C. (2019). Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy, August. Association for Computational Linguistics.

Jain, N., Popović, M., Groves, D., and Vanmassenhove, E. (2021). Generating gender augmented data for NLP. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 93–102, Online, August. Association for Computational Linguistics.

Ljubešić, N. and Lauc, D. (2021). BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine, April. Association for Computational Linguistics.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July. Association for Computational Linguistics.

Vanmassenhove, E., Hardmeier, C., and Way, A. (2018). Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November. Association for Computational Linguistics.

Zhu, W., Hu, Z., and Xing, E. P. (2019). Text infilling. *CoRR*, abs/1901.00158.

Zmigrod, R., Mielke, S. J., Wallach, H., and Cotterell, R. (2019). Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy, July. Association for Computational Linguistics.

9. Language Resource References

Jörg Tiedemann. (2012). *Parallel Data, Tools and Interfaces in OPUS*. European Language Resources Association (ELRA).