Leveraging Pre-trained Language Models for Gender Debiasing

Language Resources and Evaluation Conference (LREC) - 2022

Authors: Nishtha Jain¹, Maja Popovic², Declan Groves³, Lucia Specia⁴
Outline

Introduction
  ○ Research Area and Approach
  ○ Inspiration and Adaptation

Methodology
  ○ An example in Spanish
  ○ Filtering techniques

Test Sets for Evaluation

Evaluation and Comparison
  ○ Spanish
  ○ Serbian

Conclusions and Future Work
Outline

Introduction
  ○ Research Area and Approach
  ○ Inspiration and Adaptation

Methodology
  ○ An example in Spanish
  ○ Filtering techniques

Test Sets for Evaluation

Evaluation and Comparison
  ○ Spanish
  ○ Serbian

Conclusions and Future Work
Gender bias in language has increasingly become an important topic of research in NLP.

Although NLP models are successful in modelling various applications, they propagate and may even amplify gender biases found in the training sets.
**Gender bias in language** has increasingly become an important topic of research in NLP.

Although NLP models are successful in modelling various applications, they propagate and may even amplify gender biases found in the training sets.

Reduce gender bias by *enriching existing data with gender variants*. These *variants* can be used either *directly*, or to *create gender-balanced corpora* that can in turn be used as training data for NLP models.
Outline

Introduction
  ○ Research Area and Approach
  ○ Inspiration and Adaptation

Methodology
  ○ An example in Spanish
  ○ Filtering techniques

Test Sets for Evaluation

Evaluation and Comparison
  ○ Spanish
  ○ Serbian

Conclusions and Future Work
Inspiration and Adaptation

INSPIRATION:
Inspired by work in the area of *text infilling* (Zhu et al., 2019)
Inspiration and Adaptation

**INSPIRATION:**
Inspired by work in the area of text infilling (Zhu et al., 2019)

**ADAPTATION:**
Use the technique for paraphrasing gender-marked words in a sentence

The main challenges in this approach are to:
- select *words* whose *grammatical gender* can be *changed*
- find *appropriate variants* in context
- ensure *sentence cohesion* when multiple words can be changed.

We test this approach on a high-resource language (Spanish) as well as a low-resource language (Serbian)
Outline

Introduction
  ○ Research Area and Approach
  ○ Inspiration and Adaptation

Methodology
  ○ An example in Spanish
  ○ Filtering techniques

Test Sets for Evaluation

Evaluation and Comparison
  ○ Spanish
  ○ Serbian

Conclusions and Future Work
What is the approach?

Candidate sentence S to debias → POS tag sentence S → Can word \( w_i \) have alternative gender

- Yes: Query BERT for replacements of the word \( w_i \)
  → Apply filtering techniques to return best alternative \( w_i \)
  → Replace \( w_i \) in the candidate sentence with \( W_i \)

- No: Replace \( w_i \) as is in its original form

For each word \( w_i \) in sentence S

Is new sentence grammatically correct?

- Yes: New (debiased) sentence is returned
- No: \( w_i \) is kept as in its original form
Outline

Introduction
  ○ Research Area and Approach
  ○ Inspiration and Adaptation

Methodology
  ○ An example in Spanish
  ○ Filtering techniques

Test Sets for Evaluation

Evaluation and Comparison
  ○ Spanish
  ○ Serbian

Conclusions and Future Work
Candidate sentence: *actúa muy tranquila*.
Candidate sentence: *actúa muy tranquila*.

How it works through the pipeline to generate a gender variant?
Outline

Introduction
  ○ Research Area and Approach
  ○ Inspiration and Adaptation

Methodology
  ○ An example in Spanish
  ○ Filtering techniques

Test Sets for Evaluation

Evaluation and Comparison
  ○ Spanish
  ○ Serbian

Conclusions and Future Work
Filtering Techniques

Filtering techniques are as follows:

1. Baseline
2. POS-tag based filtering - only this one is used for Serbian
3. Normalised character-level edit distance ranking (ccer)
4. Length and prefix penalty (ccer⁺)
5. Lo/La interchanging (only for Spanish)
6. Language tool API
Outline

Introduction
  ○ Research Area and Approach
  ○ Inspiration and Adaptation

Methodology
  ○ An example in Spanish
  ○ Filtering techniques

Test Sets for Evaluation

Evaluation and Comparison
  ○ Spanish
  ○ Serbian

Conclusions and Future Work
1) Sentences have a specific structure using the rules from (Jain et al., 2021) eg. 
   VERB ADVERB ADJECTIVE
2) Sentences with a shorter length
3) At most one word which has a possible gender variant
4) # regenderable sentences > # neutral sentences
Test Sets for Evaluation - Spanish and Serbian

**Extracted from Microsoft**

### Spanish 1
1) Sentences have a specific structure using the rules from (Jain et al., 2021) eg. VERB ADVERB ADJECTIVE
2) Sentences with a shorter length
3) At most one word which has a possible gender variant
4) \# regenderable sentences > \# neutral sentences

### Spanish 2
1) Sentences do not have a specific structure using the rules from (Jain et al., 2021)
2) Sentences with longer length
3) More than one word which has a possible gender variant
4) \# neutral sentences >> \# regenderable sentences
**Test Sets for Evaluation - Spanish and Serbian**

<table>
<thead>
<tr>
<th>Spanish 1</th>
<th>Spanish 2</th>
<th>Spanish 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Sentences have a specific structure using the rules from (Jain et al., 2021) eg. <strong>VERB ADVERB ADJECTIVE</strong>&lt;br&gt;2) Sentences with a <strong>shorter</strong> length&lt;br&gt;3) <strong>At most one word</strong> which has a possible gender variant&lt;br&gt;4) # <strong>regenderable</strong> sentences &gt; # <strong>neutral</strong> sentences</td>
<td>1) Sentences <strong>do not</strong> have a specific structure using the rules from (Jain et al., 2021)&lt;br&gt;2) Sentences with <strong>longer</strong> length&lt;br&gt;3) <strong>More than one word</strong> which has a possible gender variant&lt;br&gt;4) # <strong>neutral</strong> sentences &gt;&gt; # <strong>regenderable</strong> sentences</td>
<td>1) Sentences have a specific structure using the rules from (Jain et al., 2021)&lt;br&gt;2) Sentences with a <strong>shorter</strong> length&lt;br&gt;3) <strong>More than one word</strong> which has a possible gender variant&lt;br&gt;4) # <strong>regenderable</strong> sentences &gt; # <strong>neutral</strong> sentences</td>
</tr>
</tbody>
</table>

---

1. [https://opus.nlpl.eu/](https://opus.nlpl.eu/)
Test Sets for Evaluation - Spanish and Serbian

**Extracted from Microsoft**

- **Spanish 1**
  1) Sentences have a specific structure using the rules from (Jain et al., 2021) eg.
  2) Sentences with a shorter length
  3) At most one word which has a possible gender variant
  4) # regenderable sentences > # neutral sentences

- **Spanish 2**
  1) Sentences do not have a specific structure using the rules from (Jain et al., 2021)
  2) Sentences with longer length
  3) More than one word which has a possible gender variant
  4) # neutral sentences >> # regenderable sentences

- **Spanish 3**
  1) Sentences have a specific structure using the rules from (Jain et al., 2021)
  2) Sentences with a shorter length
  3) More than one word which has a possible gender variant
  4) # regenderable sentences > # neutral sentences

**Extracted from OpenSubtitles[^1]**

- **Serbian**
  1) No rules
  2) Sentences with longer length
  3) Contain up to 4 regenderable words
  4) # regenderable sentences > # neutral sentences

---

[^1]: [https://opus.nlpl.eu/](https://opus.nlpl.eu/)
Outline

Introduction
  ○ Research Area and Approach
  ○ Inspiration and Adaptation

Methodology
  ○ An example in Spanish
  ○ Filtering techniques

Test Sets for Evaluation

Evaluation and Comparison
  ○ Spanish
  ○ Serbian

Conclusions and Future Work
What is the evaluation measure?

Word-level accuracy = # words present both in gold-standard and in generated gender variant

\[
\text{total # words}
\]

RESULTS:

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Type</th>
<th>Rules (Jain et al., 2021)</th>
<th>Baseline</th>
<th>ccer&quot;+&quot;lo/la&quot; pronoun interchanging + language tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish 1</td>
<td>all neutral</td>
<td>99.3</td>
<td>84.0</td>
<td>94.8</td>
</tr>
<tr>
<td></td>
<td>re-genderable</td>
<td>100</td>
<td>96.0</td>
<td>96.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>99.3</td>
<td>74.3</td>
<td>93.3</td>
</tr>
<tr>
<td>Spanish 2</td>
<td>all neutral</td>
<td>NA</td>
<td>93.2</td>
<td>94.7</td>
</tr>
<tr>
<td></td>
<td>re-genderable</td>
<td>NA</td>
<td>96.0</td>
<td>95.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NA</td>
<td>78.2</td>
<td>92.1</td>
</tr>
<tr>
<td>Spanish 3</td>
<td>all neutral</td>
<td>99.6</td>
<td>82.1</td>
<td>92.1</td>
</tr>
<tr>
<td></td>
<td>re-genderable</td>
<td>100</td>
<td>93.8</td>
<td>95.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>99.3</td>
<td>72.1</td>
<td>89.1</td>
</tr>
<tr>
<td>original</td>
<td>output+issue type</td>
<td>correct</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------------------------------------------------------</td>
<td>-----------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1) la cosa esta bien.</td>
<td>la casa esta bien.</td>
<td>la cosa esta bien.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(unwanted lexical change)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2) son bienvenidos</td>
<td>son <strong>bienvido</strong>  (plural to singular)</td>
<td>son <strong>bienvidas</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(improved by penalised edit distance <strong>eecer</strong>')</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3) ahora lo entiendo.</td>
<td>ahora <strong>le entiendo.</strong></td>
<td>ahora la entiendo.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(&quot;lo&quot; converted to neutral &quot;le&quot; instead of feminine &quot;la&quot;)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(solved by &quot;lo/la&quot; interchanging)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4) ahora mismo la he enviado .</td>
<td>ahora <strong>misma la he enviado</strong> .</td>
<td>ahora mismo <strong>lo he enviado</strong> .</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(incorrect words changed)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5) informemos</td>
<td>informenov (non-existing word)</td>
<td>informemos</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(improved by language tool)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6) ¡comprobemoslo!</td>
<td>¡comprobemoslo!</td>
<td>¡comprobemoslo!</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(removed accent)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(improved by language tool)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Type</th>
<th>Baseline</th>
<th>ccer⁺</th>
<th>ccer⁺ + POS tags</th>
<th>ccer⁺ + POS tags for pronouns only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serbian</td>
<td>all</td>
<td>84.5</td>
<td>80.7</td>
<td>83.2</td>
<td>84.2</td>
</tr>
<tr>
<td></td>
<td>neutral</td>
<td>99.5</td>
<td>91.5</td>
<td>99.3</td>
<td>96.3</td>
</tr>
<tr>
<td></td>
<td>re-genderable</td>
<td>81.5</td>
<td>78.6</td>
<td>80.0</td>
<td>81.8</td>
</tr>
<tr>
<td>original</td>
<td>output-issue type</td>
<td>correct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>--------------------------------------------</td>
<td>--------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a drugi ?</td>
<td>a drugi ? (unchanged)</td>
<td>a druga ?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a baš je tada otišao kuć i ?</td>
<td>a baš je tada otišlo kuć i ? (neuter gender)</td>
<td>a baš je tada otišla kuć i ?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a druge dve da ostavimo ?</td>
<td>a drugi dva da ostavimo? (gender variant but for singular instead of plural)</td>
<td>a druga dva da ostavimo?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a jesli li i ti bili ?</td>
<td>a jesli li i ti bili ? (gender unchanged, singular instead of plural)</td>
<td>a jesli li i ti bila ?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a onda je ona sišla dole</td>
<td>a onda je on sišla dole (non-existing word)</td>
<td>a onda je on sišao dole</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baš su lepe i slatke .</td>
<td>baš su leps i slatni . (non-existing words)</td>
<td>baš su lepi i slatki .</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

Introduction
○ Research Area and Approach
○ Inspiration and Adaptation

Methodology
○ An example in Spanish
○ Filtering techniques

Test Sets for Evaluation

Evaluation and Comparison
○ Spanish
○ Serbian

Conclusions and Future Work
Conclusions and Future Work

- Performs quite well on the Spanish datasets, both simple and complex, with some very specific errors
  - Serbian proved to be more challenging mainly due to the lower quality of the POS tagger and the BERT model

ADVANTAGES:
- No task-specific supervision required
- Requires minimal language-specific heuristics with some knowledge of the language
- Automatic way for generating gender variants using good pre-trained language models like BERT

FUTURE WORK:
- Using better pre-trained models such as XLMR and more research into LM-based filtering, including purposely built LMs
- 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
- Performs quite well on the Spanish datasets, both simple and complex, with some very specific errors
  - Serbian proved to be more challenging mainly due to the lower quality of the POS tagger and the BERT model

**ADVANTAGES:**
- No task-specific supervision required
- Requires minimal language-specific heuristics with some knowledge of the language
- Automatic way for generating gender variants using good pre-trained language models like BERT

**FUTURE WORK:**
- Using better pre-trained models such as XLMR and more research into LM-based filtering, including purposely built LMs
- 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
THANK YOU