ADAPT
Engaging Content
Engaging People

A World Leading SFI Research Centre

Science Foundation Ireland
For what's next

Leveraging Pre-trained Language Models for Gender Debiasing

Language Resources and Evaluation Conference (LREC) - 2022

Authors: Nishtha Jain[1], Maja Popovic[2], Declan Groves[3], Lucia Specia[4]

HOST INSTITUTION
Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

HOST INSTITUTION
DCU

PARTNER INSTITUTIONS
TU DUBLIN
University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath
Ireland's Global University
MTU
TUS
Maynooth University
OÉ Gaillimh
NUI Galway

[1]ADAPT Centre, Trinity College Dublin, [2]ADAPT Centre, Dublin City University, [3]Microsoft, Dublin, [4]Imperial College London, UK

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

Engaging Content
Engaging People

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

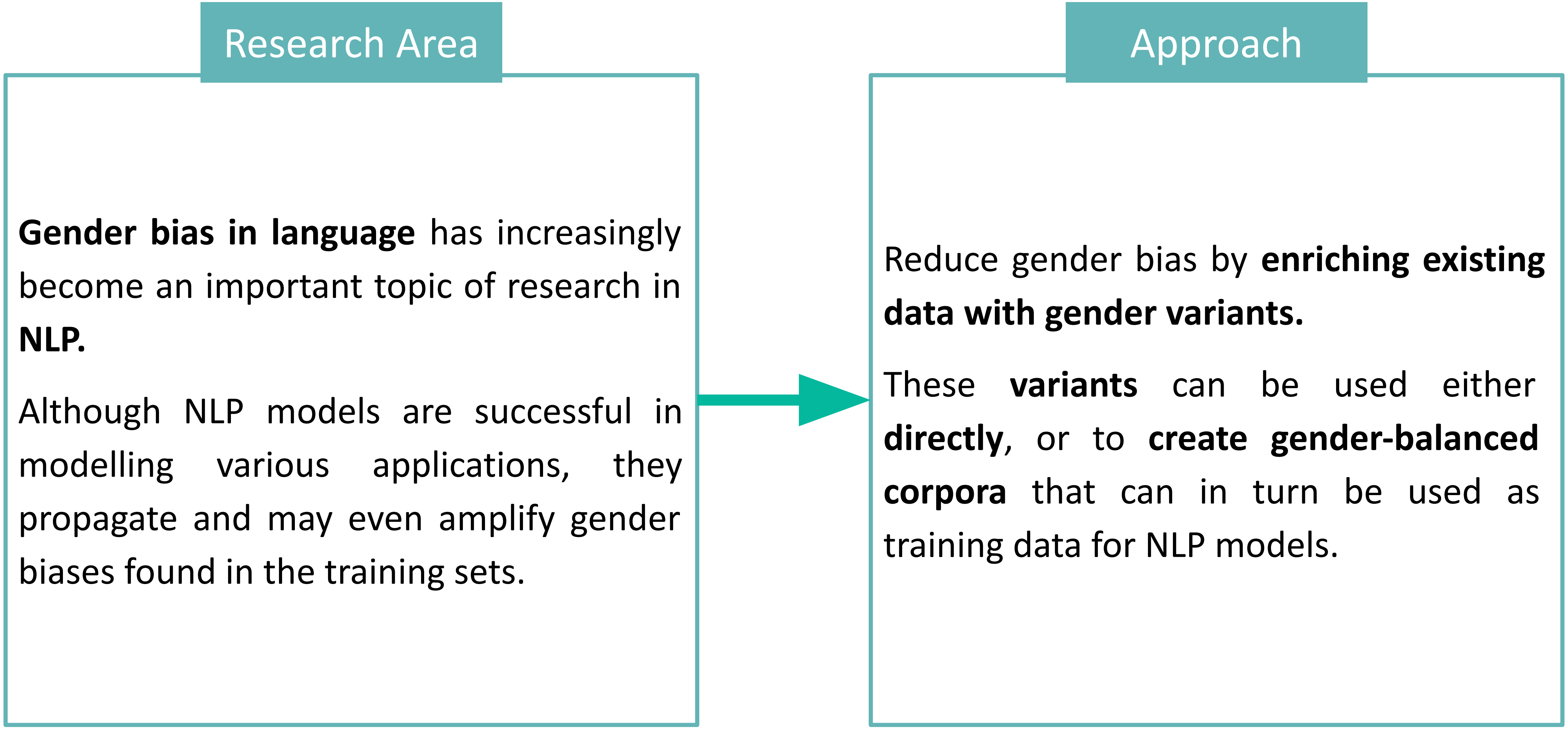
Research Area and Approach

Research Area

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.

Research Area and Approach



Research Area

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.

Approach

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

Engaging Content
Engaging People

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

Engaging Content
Engaging People

Introduction

- ○ Research Area and Approach
- ○ Inspiration and Adaptation

Methodology

- ○ An example in Spanish
- ○ Filtering techniques

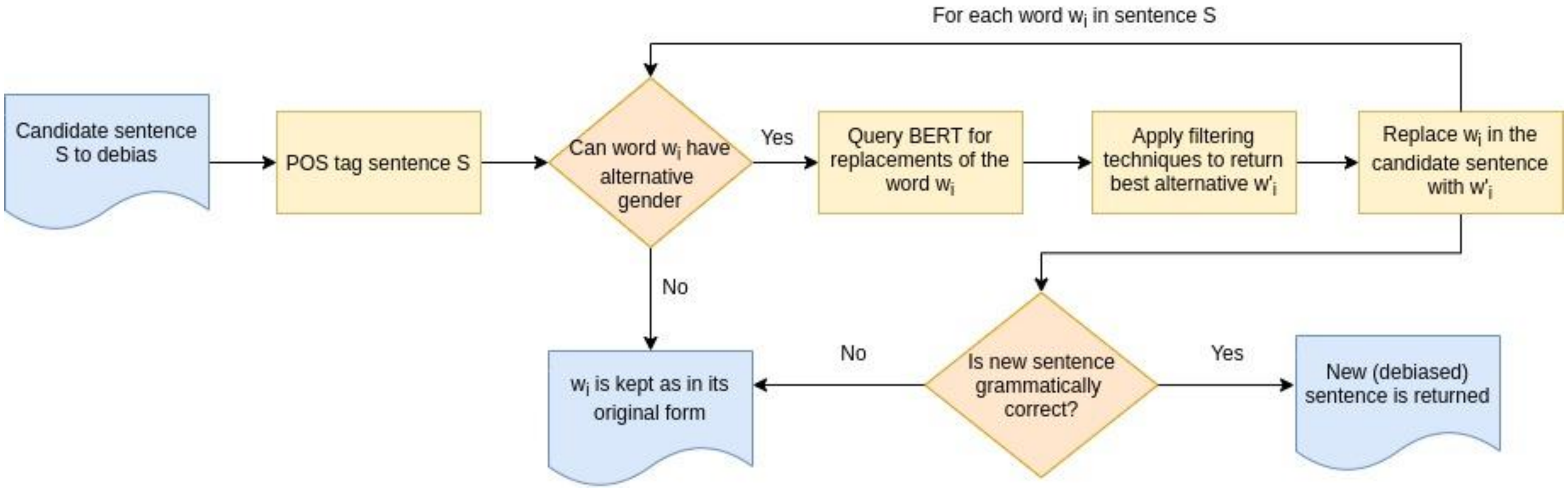
Test Sets for Evaluation

Evaluation and Comparison

- ○ Spanish
- ○ Serbian

Conclusions and Future Work

What is the approach?



For each word w_i in sentence S
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
w_i is kept as in its original form
No
Is new sentence grammatically correct?
Yes
New (debiased) sentence is returned

Note: We use bert-base-wwm-uncased for Spanish and bert-multilingual for Serbian

Outline

Engaging Content
Engaging People

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

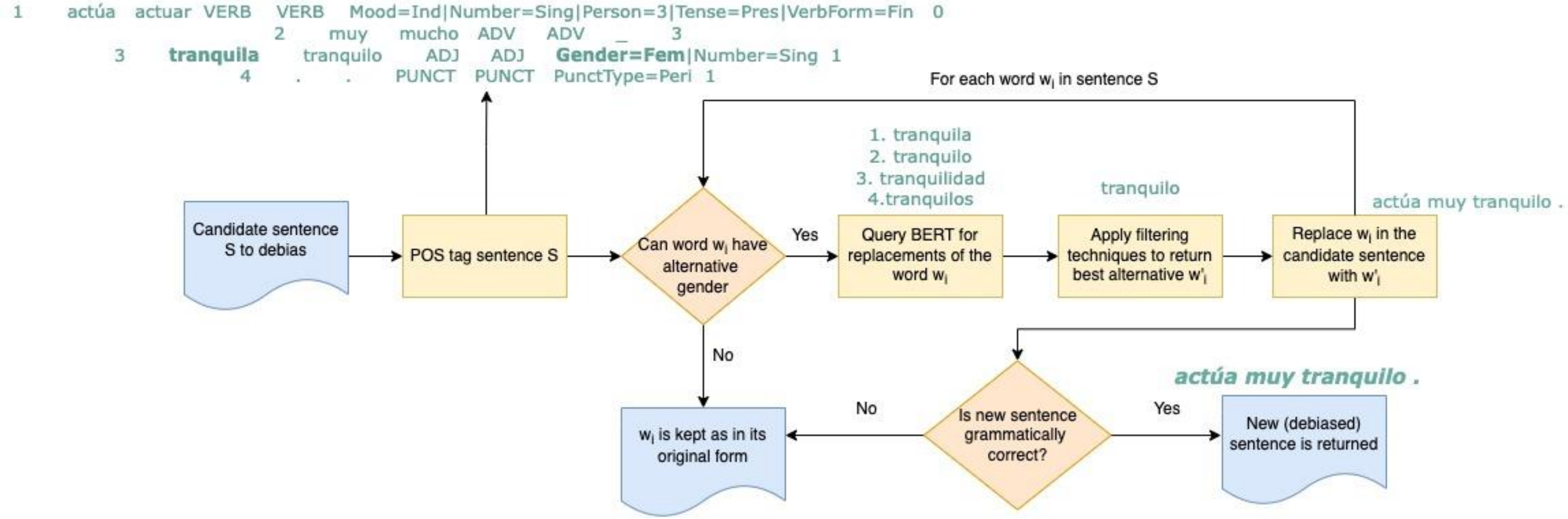
Example in Spanish

Candidate sentence: *actúa muy tranquila .*

Example in Spanish

Candidate sentence: *actúa muy tranquila .*

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



```
1      actúa   actuar  VERB    VERB    Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0
                                  2       muy     mucho   ADV     ADV     _       3
          3       tranquila       tranquilo       ADJ     ADJ     Gender=Fem|Number=Sing  1
                          4       .       .       PUNCT   PUNCT   PunctType=Peri  1
```

For each word w_i in sentence S
1. tranquila
2. tranquilo
3. tranquilidad
4.tranquilos
tranquilo
actúa muy tranquilo .
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
w_i is kept as in its original form
No
Is new sentence grammatically correct?
Yes
actúa muy tranquilo .
New (debiased) sentence is returned

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

Engaging Content
Engaging People

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

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

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

Engaging Content
Engaging People

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

Extracted from OpenSubtitles [1]

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

[1] 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. **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

Extracted from OpenSubtitles [1]

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

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/

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

Evaluation and Comparison - Spanish

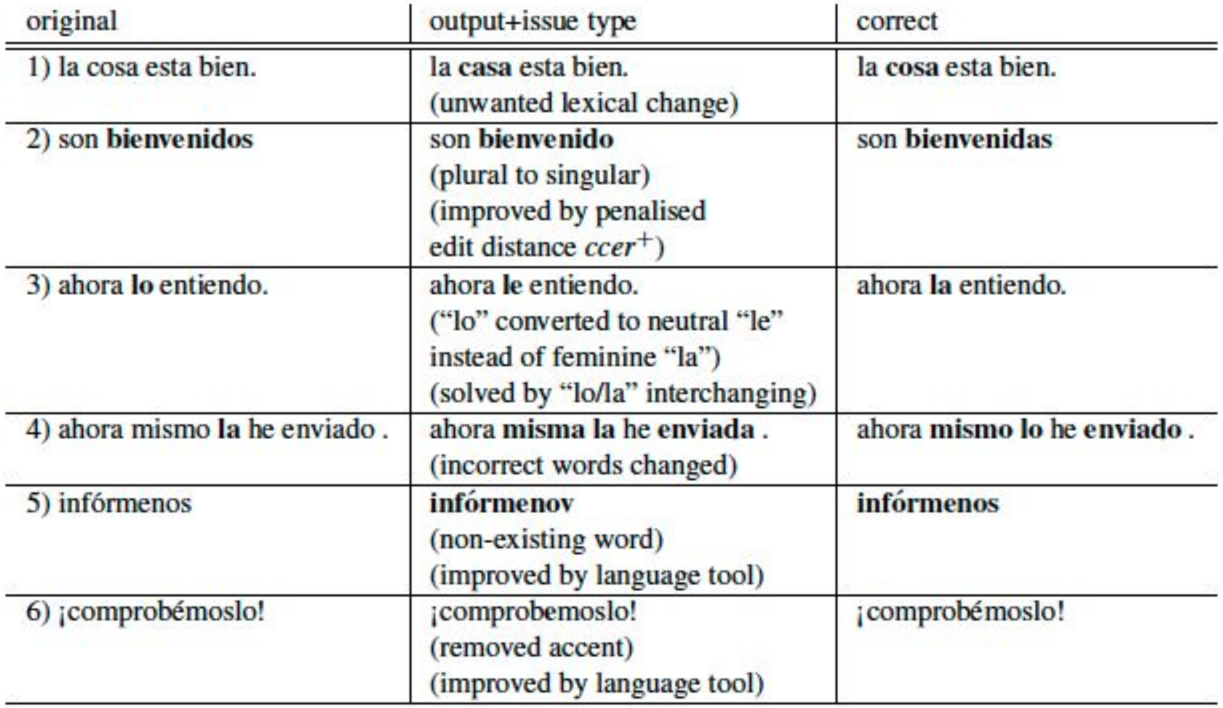
What is the evaluation measure?

$$\text{Word-level accuracy} = \frac{\textit{\# words present both in gold-standard and in generated gender variant}}{\textit{total \# words}}$$

RESULTS:

Test Set	**Type**	**Rules** (Jain et al., 2021)	**Baseline**	**ccer$^+$ +“lo/la” pronoun interchanging + language tool**
Spanish 1	all	**99.3**	84.0	**94.8**
	neutral	**100**	96.0	**96.5**
	re-genderable	**99.3**	74.3	93.3
Spanish 2	all	NA	93.2	**94.7**
	neutral	NA	96.0	**95.1**
	re-genderable	NA	78.2	**92.1**
Spanish 3	all	99.6	82.1	**92.1**
	neutral	100	**93.8**	**95.5**
	re-genderable	99.3	72.1	89.1

Example outputs with potential errors - Spanish



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!	¡comprobemoslo! (removed accent) (improved by language tool)	¡comprobémoslo!

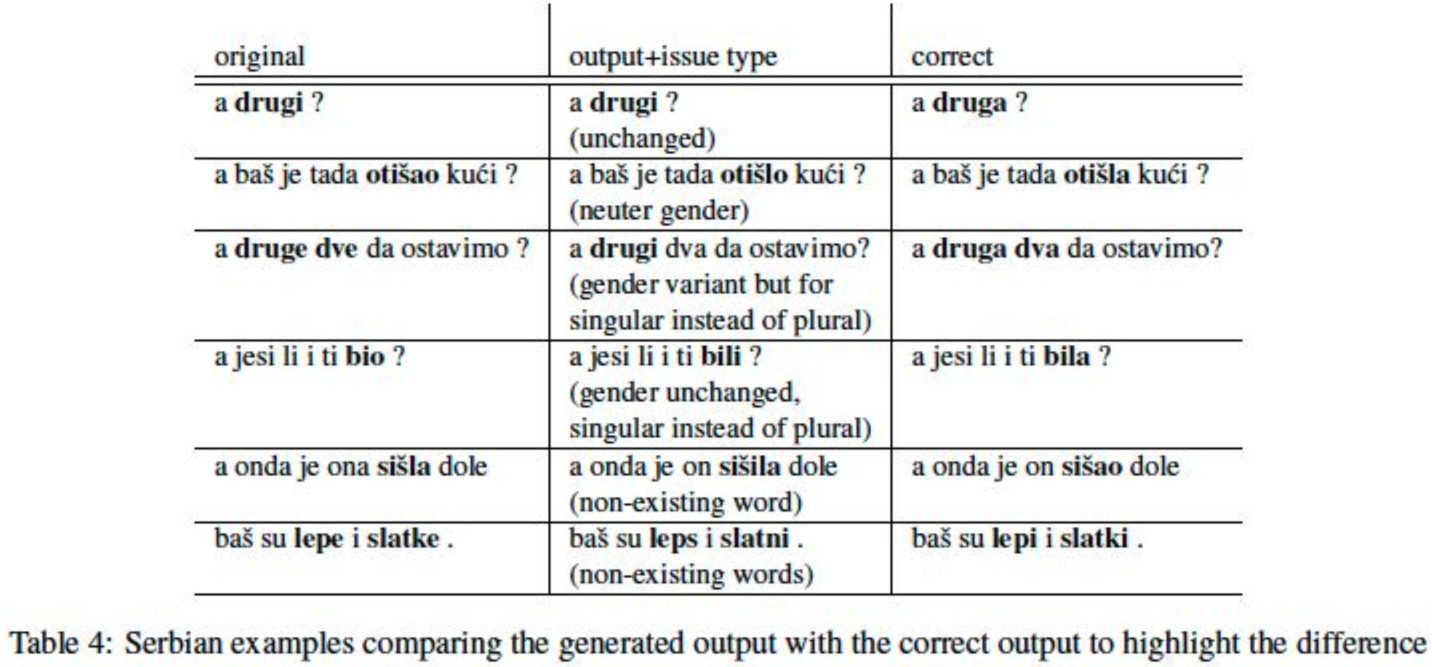
Table 3: Spanish examples comparing the generated output with the correct output to highlight the difference

Evaluation - Serbian

RESULTS:

Test Set	Type	Baseline	ccer$^+$	ccer$^+$ + POS tags	ccer$^+$ +POS tags for pronouns only
Serbian	all neutral re-genderable	**84.5** **99.5** 81.5	80.7 91.5 78.6	83.2 **99.3** 80.0	84.2 96.3 **81.8**

Example output with potential errors - Serbian



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

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

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

THANK YOU

Engaging Content
Engaging People