Appendix to Generating Gender Augmented Data for NLP

Nishtha Jain[1], Maja Popovic[2], Declan Groves[2], Eva Vanmassenhove[4]

[1]ADAPT Centre, Trinity College Dublin,
[2]ADAPT Centre, Dublin City University,
[3]Microsoft, Dublin,
[4]Department of CSAI, Tilburg University

[1,2]firstname.lastname@adaptcentre.ie, [3] degroves@microsoft.com,[4]e.o.j.vanmassenhove@tilburguniversity.edu

1 Appendix

1.1 Training Data Split

This section provides a detailed split of the data used in this research work, including OpenSubs corpus and the data obtained from our industry partner. Table 1 mentions the split of neutral and re-genderable segments in each set.

set	type	# of segments	# of running words
training for NMT rewriter	all	2 193 657	15 540 108
	neutral	1 145 781	8 360 144
	re-genderable	1 047 876	7 179 964
development	all	1 018	4 350
	neutral	432	2 021
	re-genderable	506	2 329
test	all	3 066	12 996
	neutral	1 289	6 045
	re-genderable	1 777	6 951
structured test1	all	5 648	25 605
	neutral	2 304	11 607
	re-genderable	3 344	13 998
unstructured test1	all	15 892	88 537
	neutral	14 497	81 075
	re-genderable	2 752	14 680
training for classifier	all	7 692	36 250
	neutral	3 440	17 438
	re-genderable	4 252	18 812

Table 1: Data statistics

1.2 POS Sequences and Rewriting Rules

This section mentions in detail the word categories analyzed in the industry sourced data. The corresponding POS sequences and rules formulated to rewrite gender variants for those particular word categories are given in respective tables.

Clitic Pronoun Candidates

Table 2 consists of the POS sequences of gendered utterances that contain past clitic pronouns.

POS Sequences including regenderable clitic pronouns (PPC)	Rewriting Rules for each PPC
PPC-Vfin-FS	
FS-PPC-Vfin-FS	
PPC-Vfin-ADV-FS	
Vfin-CQUE-PPC-Vfin-FS	
NEG-PPC-Vfin-FS	
ADV-PPC-Vfin-FS	"lo" => "la"
ADV-CM-PPC-Vfin-FS	"la" => "lo"
PPC-Vfin-CM-NC-FS	"los" => "las"
ADV-NEG-PPC-Vfin-FS	"las" => "los"
NEG-PPC-Vfin-ADV-FS	
Vfin-CQUE-NEG-PPC-Vfin-FS	
NEG-Vfin-CQUE-PPC-Vfin-FS	

Table 2: POS sequences and rewriting rules for clitic pronouns

Demonstrative Pronoun Candidates

Table 3 consists of the POS sequences of gendered utterances that contain demonstrative pronouns.

POS Sequences including regenderable demonstrative pronouns (DM)	Rewriting Rules for each DM
Vfin-DM-FS	"este" => "esta"
FS-Vfin-DM-FS	"esta" => este"
FS-INT-Vfin-DM-FS	"estas" => "estos"
NEG-Vfin-DM-FS	"ese", => "esa"
FS-NEG-Vfin-DM-FS	"esa" => "ese"
DM-FS	"esos" => "esas"
DM-SE-Vfin-FS	"esas" => "esos"
ADV-Vfin-DM-FS	"aquel", => "aquella"
DM-NEG-FS	"aquella" => "aquel"
DM-PPX-Vfin-FS	"aquellos" => "aquellas"
FS-CC-DM-FS	"aquellas" => "aquellos"
DM-NEG-Vfin-FS	"estos" => "estas"

Table 3: POS sequences and rewriting rules for demonstrative pronouns

Past Participles Candidates

Table 4 consists of the POS sequences of gendered utterances that contain past participles.

POS Sequences including regenderable past participles (Vadj)	Rewriting Rules for each Vadj
Vadj-FS	if word suffix is "ado", "ido", "cho" => last letter to "a" if the word suffix is "ada", "ida", "cha" => last letter to "o" if word suffix is "ados", "idos", "chos" => last two letters "as" if the word suffix is "adas", "idas", "chas" => last two letters "os"
Vfin-Vadj-FS	
Vadj-CC-Vadj-FS	
Vfin-ADV-Vadj-FS	
FS-Vfin-Vadj-FS	
FS-Vadj-FS	
ADV-Vadj-FS	
ADV-Vfin-DM-FS	
Vadj-ADV-FS	
FS-Vfin-ADV-Vadj-FS	
ADV-CM-Vadj-FS	
ADV-Vfin-Vadj-FS	
NEG-Vadj-FS	

Table 4: POS sequences and rewriting rules for past participles

Adjective Candidates

Table 5 consists of the POS sequences of gendered utterances that contain adjectives.

POS Sequences including regenderable adjectives (ADJ)	Rewriting Rules for each ADJ
ADJ-FS	if suffix "o" => "a" if suffix "dor" => "dora" if suffix "os" or "dores" => last two letters to "as" if suffix "dora" => "dor" if suffix "doras" => "dores" if suffix "a" => "o" if suffix "as" => "os"
Vfin-ADJ-FS	
FS-Vfin-ADJ-FS	
ADV-ADJ-FS	
Vfin-ADV-ADJ-FS	
FS-ADJ-FS	
FS-Vfin-ADV-ADJ-FS	
ADV-Vfin-ADJ-FS	
NEG-Vfin-ADJ-FS	
FS-INT-ADJ-FS	
VMfin-Vinf-ADJ-FS	
SE-Vfin-ADJ-FS	
ADJ-CC-ADJ-FS	

Table 5: POS sequences and rewriting rules for adjectives

Clitic Pronouns Attached to Verbs

Table 6 consists of the rewriting rules applied to gendered utterances that contain clitic pronouns attached to verbs.

For clitic pronouns attached to verbs, if a VCL tag is present in the POS sequence of the sentence then it represents a VCL candidate[1]. Table 6 represents the rules to tackle such structures.

Neutral Past Participle Structures

Table 7 consists of the POS sequences which contain past participles which should not be regendered.

[1]POS tags for this category are not very clean, many of verbs with clitic pronouns are tagged as a simple verb infinitive, therefore this rule was included (infinitives without clitic pronouns cannot end with "lo/la/los/las").

Rewriting Rules for each clitic pronoun attached to a verb
if suffix "lo" => "la"
if suffix "la" => "lo"
if suffix "los" => "las"
if suffix "las" => "los"

Table 6: POS sequences and rewriting rules for clitic pronouns attached to verbs

POS Sequences including past participles (Vadj) which should not be regendered
NC-Vadj-FS
FS-NC-Vadj-FS
Vadj-CC-Vadj-FS
VHfin-Vadj-ART-NC-FS
FS-NC-Vadj-FS
ART-NC-SE-VHfin-Vadj-FS
ART-NC-Vfin-Vadj-FS
ADV-Vadj-NC-FS
FS-ADV-Vadj-NC-FS
Vfin-ADV-Vadj-NC-FS
FS-Vfin-ADV-Vadj-NC-FS

Table 7: POS sequences containing past participles which should not be regendered

Neutral Adjective Structures

Table 8 consists of the POS sequences containing adjectives which should not be regendered.

POS Sequences including adjectives (ADJ) which should not be regendered
FS-ADJ-NC-FS
Vfin-ART-NC-ADJ-FS
FS-Vfin-ART-ADJ-NC-FS
FS-INT-ADJ-NC-FS
NC-ADJ-FS
ART-NC-Vfin-ADJ-FS
ADV-ADJ-NC-FS
FS-ADV-ADJ-NC-FS
Vfin-ADV-ADJ-NC-FS
FS-Vfin-ADV-ADJ-NC-FS

Table 8: POS sequences containing adjectives which should not be regendered