The Arabic Parallel Gender Corpus 2.0: Extensions and Analyses

Bashar Alhafni, Nizar Habash, Houda Bouamor†

Computational Approaches to Modeling Language Lab, New York University Abu Dhabi

†Carnegie Mellon University in Qatar

{alhafni, nizar.habash}@nyu.edu, hbouamor@qatar.cmu.edu

Introduction

- New corpus for Arabic gender identification and rewriting in contexts involving one or two target users (I and/or You) with different grammatical gender preferences.
- Expands on Habash et al. (2019)'s Arabic Parallel Gender Corpus v1.0 by including 1st and 2nd grammatical gender references covering singular, dual, plural constructions; and adding 6 times more sentences, reaching over 80K sentences (590K words).

Annotation and Selection

- 63K English-Arabic OpenSubtitles sentences.
- Annotated by four professional linguists.
- **Gender Identification**
  - Identify the genders of the 1st and 2nd person references in each sentence: F (feminine), M (masculine), B (ambiguous), or N (non-existent).
  - Identify the dual and plural gendered references ('!').
  - Avoid any heterocentric assumptions (e.g., 'you are my husband' is labeled as BM and not FM). All proper names are treated as a gender ambiguous (B).
- **Gender Rewriting**
  - In case gendered references exist in a sentence, introduce all of its possible opposite gender forms.
  - Rewriting is limited to morphological reinflections and word substitutions (i.e., 1-to-1 alignment at the word-level).

English	Arabic	Label	Reinflection Label	Reinflection	
I wanna thank you	أريد أن أشكرك	BB			(a)
I have something to say	لدي شيء لأقوله	BN			(b)
I'm so happy for you	أنا سعيدة من أجلك	FB	MB	أنا سعيد من أجلك	(c)
We were coming to see you	نحن قادمات لرؤيتك	F!B	M!B	نحن قادمون لرؤيتك	(d)
Because I'm your big brother	لأنني أخوك الكبير	MB	FB	لأنني أختك الكبيرة	(e)
We're ready	نحن مستعدون	M!B	F!B	نحن مستعدات	(f)
I know, babe	أعلم ذلك يا عزيزتي	BF	BM	أعلم ذلك يا عزيزي	(g)
I respect you [*plural*]	أنا أحترمكن	BF!	BM!	أنا أحترمكم	(h)
I'm right here dad	أنا هنا يا أبي	BM	BF	أنا هنا يا أمي	(i)
I love you [*plural*] so much	أحبكم كثيرا	BM!	BF!	أحبكن كثيرا	(j)
I'm sorry, you're going to have to leave	آسف يجب أن ترحل	MM	FM	آسفة يجب أن ترحل	(k)
			MF	آسف يجب أن ترحلي	(l)
			FF	آسفة يجب أن ترحلي	(m)
Baby, I'm so scared right now	أنا خائفة للغاية يا عزيزي	FM	MM	أنا خائف للغاية يا عزيزي	(n)
			FF	أنا خائفة للغاية يا عزيزتي	(o)
			MF	أنا خائف للغاية يا عزيزتي	(p)
I'm glad you made it home, mom	أنا سعيد بعودتك يا أماه	MF	FF	أنا سعيدة بعودتك يا أماه	(q)
			MM	أنا سعيد بعودتك يا أبتاه	(r)
			FM	أنا سعيدة بعودتك يا أبتاه	(s)
Don't call me a fool	لا تناديني بالغبية	FF	MF	لا تناديني بالغبي	(t)
			FM	لا تنادني بالغبية	(u)
			MM	لا تنادني بالغبي	(v)

(a)

Original Corpus						
Sentences		Label	Reinflection Label			
36,980	63.7%	BB				
1,123	1.9%	FB	MB			
1,940	3.3%	MB	FB			
5,210	9%	BF	BM			
12,164	21%	BM	BF			
68	0.1%	FF	MF	FM	MM	
135	0.2%	FM	MM	FF	MF	
117	0.2%	MF	FF	MM	FM	
298	0.5%	MM	FM	MF	FF	
58,035						

(b)

Balanced Corpus						
Input	$Target_{MM}$	$Target_{FM}$	$Target_{MF}$	$Target_{FF}$	Sentences	
BB	BB	BB	BB	BB	36,980	46%
FB	MB	FB	MB	FB	3,063	3.8%
MB	MB	FB	MB	FB	3,063	3.8%
BF	BM	BM	BF	BF	17,374	21.6%
BM	BM	BM	BF	BF	17,374	21.6%
FF	MM	FM	MF	FF	618	0.8%
FM	MM	FM	MF	FF	618	0.8%
MF	MM	FM	MF	FF	618	0.8%
MM	MM	FM	MF	FF	618	0.8%
					80,326	

Corpus Statistics

- **Original Corpus** 8.2% of the original selected sentences were dropped resulting in 58,035 sentences
- **Balanced Corpus** Gender balance the corpus by integrating the rewritten sentences.

Quantifying Bias in Gender-Unaware Machine Translation

- Translated the English side of the Input balanced corpus to Arabic using the Google Translate API.
- Studied the bias in the Arabic translations of gender specific Arabic and English sentences.

Selected $Sentences_{ar}$	Count	$Target_{MM}$	$Target_{FM}$	$Target_{MF}$	$Target_{FF}$	Multi-Reference
ALL	80,326	13.5	13.1	11.4	11.0	13.6
BB	36,980	14.0	14.0	14.0	14.0	14.0
ALL - BB	43,346	13.1	12.4	9.3	8.6	13.4
BM BF	34,748	13.1	13.1	8.6	8.6	13.3
MB FB	6,126	12.9	9.6	12.9	9.6	13.6
MM FM MF FF	2,472	12.9	9.5	9.5	6.7	13.5