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).

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.