Automatic Gender Identification and Reinflection in Arabic

Nizar Habash, Houda Bouamor,[†] Christine Chung
Computational Approaches to Modeling Language Lab
New York University Abu Dhabi, UAE
[†]Carnegie Mellon University in Qatar, Qatar
{nizar.habash,cic266}@nyu.edu, hbouamor@qatar.cmu.edu

Abstract

The impressive progress in many Natural Language Processing (NLP) applications has increased the awareness of some of the biases these NLP systems have with regards to gender identities. In this paper, we propose an approach to extend biased single-output gender-blind NLP systems with gender-specific alternative reinflections. We focus on Arabic, a gender-marking morphologically rich language, in the context of machine translation (MT) from English, and for first-person-singular constructions only. Our contributions are the development of a system-independent gender-awareness wrapper, and the building of a corpus for training and evaluating first-person-singular gender identification and reinflection in Arabic. Our results successfully demonstrate the viability of this approach with 8% relative increase in BLEU score for first-person-singular feminine, and 5.3% comparable increase for first-person-singular masculine on top of a state-of-the-art gender-blind MT system on a held-out test set.

1 Introduction

The impressive progress in the last decade in many Natural Language Processing (NLP) applications, from machine translation (MT) to dialogue system, has increased awareness of some of the biases these systems have with regards to gender identities. A case in point is the *I-am-a-doctor/ I-am-a-nurse* MT problem in many morphologically rich languages. While English uses gender-neutral terms that hide the ambiguity of the first-person gender reference, many morphologically rich languages need to use different grammatically gender-specific terms for these two expressions. In Arabic, as in other languages with grammatical gender, gender-blind single-output MT from English often results in أنا طبيب *ÂnA Tbyb*[1] 'I am a [male] doctor'/أنا ممرضة *ÂnA mmrDħ* 'I am a [female] nurse', which is inappropriate for female doctors and male nurses, respectively.

Part of this problem comes from human-generated data that mirrors the social biases and inequalities of the world we live in, and that results in biased models and representations. Many research efforts responded to this problem by debiasing and balancing the models created from the data through model modification or data augmentation (Font and Costa-jussà, 2019; Zmigrod et al., 2019). However, ultimately, even the most balanced and unbiased of models can be useless in gender-blind systems that are designed to generate a single text output. Such systems are doomed to unsurprisingly pass on the biases of the models they use, as demonstrated in the doctor/nurse example above. In contrast, gender-aware systems should be designed to produce outputs that are as gender-specific as the input information they have access to. The input gender information may be contextual (e.g., the input 'she is a doctor'), or extra linguistics (e.g., the gender feature provided in the user profile in social media). But, there may be contexts where the gender information is unavailable to the system (e.g., 'the student is a nurse'). In such cases, generating both gender-specific forms or a gender-neutral (gender-ambiguous) form is more appropriate.

In this paper, we propose an approach that extends the possibly biased output of gender-blind NLP systems with gender-specific reinflections. This is a monolingual postprocessing rephrasing task that wraps around a gender-blind system to make it gender-aware, through identifying if there are gender-specific phrases in its output and of-

[1] Arabic transliteration is in the HSB scheme (Habash et al., 2007).

fering alternative reinflections instead. The selection of the gender-specific form is then left to the user or another automatic component has access to extra-linguistic information, such as profile gender. For example, the Arabic gender-blind MT output translating English ‘I am a nurse’ as أنا ممرضة *ÂnA mmrDħ* ‘I am a [female] nurse’ is turned into two gender-marked output options: (a) أنا ممرضة *ÂnA mmrDħ* ‘[First Person Singular Feminine]’, and (b) أنا ممرض *ÂnA mmrD* ‘[First Person Singular Masculine]’. Since the output of the gender-blind NLP system is not necessarily always masculine or feminine, our approach requires two components: **gender identification** and **gender reinflection**, which can be modeled jointly or in cascade. The approach is system-independent and can be used with MT, dialogue systems, etc., as well as, to balance corpora through augmentation by adding reinflected copies of gender-specific constructions.

We focus on Arabic, a gender-marking morphologically rich language, in the context of MT from English, and for first-person-singular constructions only. We only work on first-person constructions because they tend to be gender-neutral in English. Furthermore, as sentences may involve multiple gendered references, we wanted to control for the number of combinations. We plan to extend to multiple references in future work.

Our contributions are the development of a system-independent gender-awareness wrapper, and the building of a corpus for training and evaluating first-person-singular gender identification and reinflection in Arabic. For gender identification, we compare rule-based and machine learning methods using our annotated corpus. For gender reinflection, we use a character-level neural MT (NMT) model in a single step (identify and reinflect, jointly), and as the second part of a two-step (identify then reinflect) system. Our results successfully demonstrate the viability of this approach with 8% relative increase in BLEU score for first-person-singular feminine, and 5.3% comparable increase for first-person-singular masculine on top of a state-of-the-art gender-blind MT system on a held-out test set.

Next, we discuss some related work (Section 2) and Arabic linguistic facts (Section 3). We present our Arabic parallel gender corpus in Section 4, gender identification in Section 5, and gender reinflection and MT results in Section 6.

2 Related Work

Gender bias has been detected, studied, and partially addressed for standard and contextualized word embeddings in a number of studies (Bolukbasi et al., 2016; Caliskan et al., 2017; Sutton et al., 2018; Basta et al., 2019; Garg et al., 2018; Zhao et al., 2018, 2019). These studies showed that training word embeddings on large human produced corpora such as news text leads to encoding societal biases including gender and race. Some of these studies focused on quantifying the bias, and proposed approaches for mitigating it within word embeddings.

In the context of data augmentation solutions, Lu et al. (2018) introduced *counterfactual data augmentation* (CDA), a generic methodology to mitigate bias in neural NLP tasks, where for each training instance, a copy with an intervention on its targeted words is added, replacing each with its partner, while maintaining the same ground truth. The goal here is to encourage learning algorithms to not pick up on biased distinctions. Building on CDA, (Zmigrod et al., 2019) presented a generative model that allows conversion between masculine inflected and feminine inflected sentences in four morphologically rich languages (Hebrew, Spanish, French and Italian) with a focus on animate nouns.

Specifically for MT, Rabinovich et al. (2016) presented work on the preservation of author gender. Some researchers suggested improvement through co-reference resolution (Gonzales and Tuggener, 2017; Luong and Popescu-Belis, 2016). Vanmassenhove et al. (2018) conducted a series of experiments to improve morphological agreement and improve translation quality in NMT systems for 20 language pairs (*none of which were Arabic*). They compiled large datasets from Europarl (Koehn, 2005), including speaker gender and age, and trained NMT systems with the tagged language pair. They showed that providing tags that indicate the speaker’s gender to the system leads to significant improvements. Similarly, Elaraby et al. (2018) marked speaker and listener gender as meta-data input on the source sentence in an English-to-Arabic NMT system. The training data came from OpenSubtitle (Lison and Tiedemann, 2016). The authors used rules to identify the gender in the Arabic text. Prates et al. (2018) used Google Translate to translate a set consisting of a list of jobs and gender-specific sentences from a variety of gender-neutral languages into English. They showed that occupations related to science, engineering and mathematics present a strong stereotype towards the male gender. More recently, Font and Costa-jussà (2019) studied the impact of gender debiasing on NMT between English and Spanish using debiased and gender-neutral word embeddings.

Google Translate publicly announced an effort to address gender bias for a few languages in different degrees and contexts (Help, 2019). As of the time of writing this paper, the system shows both feminine and masculine translations for some single words in certain languages; and provides gender-specific pronominal translations for some gender ambiguous cases (i.e., Turkish-English MT). In our work, we also evaluate on the output of Google Translate.

This paper sits in the intersection of efforts like data augmentation for morphologically rich languages (Zmigrod et al., 2019) and gender-aware MT (Vanmassenhove et al., 2018; Elaraby et al., 2018). Similarly to Zmigrod et al. (2019), we are interested in reinflection, but we implement it as character-based NMT. While Vanmassenhove et al. (2018) and Elaraby et al. (2018) expect gender meta-information as input, we propose a gender-aware post-processing approach, that applies gender identification and reinflection.

3 Arabic Linguistic Facts

We present three specific challenges for Modern Standard Arabic (MSA) NLP with attention to gender expression and MT.

Morphological Richness Arabic is a morphologically rich language that inflects for gender, number, person, case, state, aspect, mood and voice, in addition to allowing a number of attachable clitics (prepositions, particles, pronouns) (Habash, 2010). Wright (1955) classifies nouns according to their gender into three classes: masculine (**M**), feminine (**F**), and those that can be either masculine or feminine (**B**). Examples include طبيب *Tabiyb* 'male doctor' [M], طبيبة *Tabiybaħ* 'female doctor' [F], and words like طريق *Tariyq* 'road' [B]. Arabic adjectives have gender-specific forms (M or F). But some pronouns and some verb conjugations can be used for either masculine or feminine (B). For example, the pronoun أنا *ÂanA* 'I', and the first-person-singular perfect and imperfect verbal conjugations (e.g., كتبت *katabtu* 'I wrote' and أكتب *Âaktub* 'I write') are all gender-ambiguous (B).

The Arabic agreement system between verbs and their subjects, and between nouns and their adjectives does not just involve gender, number, case and state, but also a lexical feature called *rationality* – a quality typically associated with human actors (Alkuhlani and Habash, 2011). For instance, while adjectives modifying rational nouns agree with them in gender and number; adjectives modifying irrational plural nouns are always feminine and singular.

Orthographic Ambiguity Arabic is also orthographically ambiguous due to the optionality of diacritic specification in the written form. This optionality can lead to gender ambiguous orthographic forms as some gender-specific forms only differ in diacritics (short vowel specification).[2] For example, the word كتبت can be diacritized as *katabta* 'you [masc.sing] wrote' or *katabti* 'you [fem.sing] wrote', and it is ambiguous with yet two other forms: *katabtu* 'I [fem/masc] wrote' and *katabat* 'she wrote'. In this regard, orthographic ambiguity reduces gender bias. But it is still an issue for speech synthesis systems (Halabi, 2016).

In general, for first person expressions, we expect the verbal sentences to be gender-ambiguous (B), and the copular/equational sentences involving adjectives and rational nouns to be gender-specific ([M] or [F]). We will present an analysis of our data in the next section that confirms this.

Orthographic Noise MSA *unedited* text tends to have a large percentage (∼23%) of spelling errors (Zaghouani et al., 2014). Most common errors involve Alif-Hamza (Glottal Stop) spelling (أ، إ، آ، ا *A, Ā, Ǎ, Â*), Ya spelling (ي، ى *y, ý*), and the feminine suffix Ta-Marbuta spelling (ة، ه *h, ħ*). These errors are so common, that in Arabic NLP, Alif/Ya normalization is standard preprocessing (Habash, 2010), and Alif/Ya specification is done as postprocessing (El Kholy and Habash, 2010). Since the Arabic text we use from the OpenSubtitles Corpus (Lison and Tiedemann, 2016), a collection of translated movie subtitles, has many spelling errors of the above mentioned kinds, we evaluate MT within an orthographically normalized space (more details in Section 6).

[2] We will use the label B to refer to inherent gender ambiguity, as well as gender ambiguity resulting from undiacritized spelling.

			Original Corpus		Balanced Corpus					
			(a)	(b)	(c)	(d)	(e)			
Sentences	**Words**	**Words^{MF}**	**Input**	**Reinflected**	**Input**	**Target^{M}**	**Target^{F}**	**Sentences**	**Words**	**Words^{MF}**
10,242	74,702	0	B		B	B	B	10,242	74,702	0
362	2,720	422	F	M^r	F	M^r	F	362	2,720	422
636	4,710	743	M	F^r	M	M	F^r	636	4,710	743
					M^r	M^r	F	362	2,720	422
					F^r	M	F^r	636	4,710	743
11,240	82,132	1,165						12,238	89,562	2,330

Table 1: Statistics of the original corpus we annotated and the balanced version we report on in the paper experiments. Words^{MF} refers to the count of gender-marking words, specifically. M^r and F^r are the reinflected versions of the F and M labelled sentences, respectively, in the same rows they appear in.

English	Original Arabic	Gender	Reinflection
I have no interest in that.	ليس لدي مصلحة في هذا	B	
He shot at me!	لقد أطلق النار علي!	B	
I'm leaving.	أنا راحلة	F	أنا راحل
I'm rich! I'm rich!	أنا غنية أنا غنية	F	أنا غني أنا غني
I am a Muslim and a Hindu and a Christian and a Jew.	أنا مسلم و هندوسي و نصراني و يهودي	M	أنا مسلمة و هندوسية و نصرانية و يهودية
I'm the new attending.	أنا الأخصائي الجديد .	M	أنا الأخصائية الجديدة .

Table 2: Examples from the Arabic Parallel Gender Corpus including original sentence, its gender and its reinflection to the opposite gender where appropriate.

4 The Arabic Parallel Gender Corpus

For the kind of experiments we conduct in this paper, we need a corpus of first-person-singular Arabic sentences that are gender-annotated and gender-translated. That is, for every sentence in such corpus, we would like the gender of the sentence's speaker to be identified as *B* (gender-ambiguous), *F* (feminine) or *M* (masculine); and for the F and M cases, we would like the equivalent opposite gender form. Such a corpus needs to also be paired with English translations to support possible MT experiments. To the best of our knowledge, no such corpus exists for Arabic, nor for any other language. We plan to make this resource publicly available.[3] We describe next the approach we followed to build this corpus.

Corpus Selection We decided to use a subset of the sentences from the OpenSubtitles 2018 corpus (Lison and Tiedemann, 2016). We selected this corpus because it has parallel English and Arabic sentences, and because it contains a lot of first-person-singular sentences. We first extracted all the English-Arabic sentence pairs that include first-person-singular pronouns in the English side: *I, me, my, myself, mine*. We used English because it is not a pro-drop language like Arabic. There were 8.5 million sentences of this kind, 5.7 million of which do not include a second person pronoun (*you, your, yourself, yours*). In this work, we decided to focus on the first-person-singular exclusively and excluded all second person cases. Out of this rich set, we selected 12,000 sentences to be annotated. All the Arabic sentences were white-space-and-punctuation tokenized, as well as morphologically analyzed and lemmatized using the MADAMIRA toolkit for Arabic NLP (Pasha et al., 2014).

Corpus Annotation Four Arabic native speakers (three female and one male) annotated the corpus. The instructions were simple. First, they are to identify the grammatical gender of the singular speaker in each sentence and then label it as F (feminine), M (masculine), or B (ambiguous). Second, for the F and M cases, the annotators are to copy the sentence and minimally modify it so that it expresses the opposite gender and remains fully grammatical; they are only allowed to use word substitutions, i.e., no additions or deletions so that the total number of words is maintained. For most words, the gender reinflection maintained the same lemma, e.g., الطبيب *AlTbyb* 'the doctor' [M] is reinflected as الطبيبة *AlTbybħ* 'the doctor' [F]. However, gender-specific nouns that cannot reinflect in the same lemma are mapped

[3] http://resources.camel-lab.com

appropriately to a related lemma expressing the opposite gender. For example, the word أم *Âm* 'mother' is mapped to أب *Âb* 'father'.[4] Proper names are all treated as gender-neutral (B), even when they have strong gender-specific associations, and as such are not reinflected. The annotators were made aware of hetreo-centrist interpretations and were instructed to suspend any preconceived assumptions, e.g., the sentence تلك زوجتي 'That's my wife' is given the label B, not M. Finally, the annotators were also instructed to flag bad translations or malformed sentences. Examples from our corpus are illustrated in Table 2. The average pairwise inter-annotator agreement on a 60-sentence set that was annotated by all annotators is quite high (97.2%), suggesting the task is reasonable. The points of disagreement were plausible different interpretations. For example, the word متأخرا *mtÂxrA* 'late' in the sentence أستيقظ متأخرا *ÂstyqĎ mtÂxrA* 'I wake up late' was interpreted as an adverb (which would not gender-inflect) or as an adjective (which would).

The Original Corpus After the annotation was completed, we excluded all sentences with malformed input, sentences with Latin characters, and sentences with Arabic-Arabic gender misalignment due to annotation errors. This resulted in a set of 11,240 sentences (82,132 words), which constitute our Original Corpus Input (Table 1, column (a)). In this corpus, about 91% of all the sentences are gender-ambiguous (B). Interestingly, the M sentences are almost twice as many as the F sentences. All of the gender-specific sentences were reinflected (M → F^r and F → M^r), resulting in an additional 989 sentences (7,430 words) (Table 1, column (b)). Among the words of the first-person-singular gender-specific sentences, 1,165 are gender-specific (15.7%). The percentage of these words in the whole corpus is 1.4%.

The Balanced Corpus Given the stark gender imbalance as well as the small ratio of gender-specific sentences, we opted to balance the corpus by introducing the reinflected sentences (M^r and F^r) as if they were original input and pair them with their original input as their reinflection. In Table 1 the added sentence statistics appear in the two additional rows under Balanced Corpus. In Table 1 columns (c), (d) and (e), we define three versions of the balanced corpus, which we will refer to and use in the rest of the paper. The first is the balanced Arabic input corpus (henceforth, *Balanced Input*), which matches the original input plus the added reinflected sentences. The second is a masculine target corpus (Target^M) containing only B, M and M^r sentences. And the last is a feminine target corpus (Target^F) containing only B, F, and F^r sentences. All three corpora naturally have the same number of sentences, words, and gender-specific words. Given the addition of the reinflected sentences, the percentage of all gender-specific sentences in the balanced corpus is 16.3% and the number of masculine and feminine sentences is the same. The balanced corpora were all divided randomly and in parallel into training (TRAIN: 70% or 8,566 sentences), development (DEV: 10% or 1,224 sentences) and blind test (TEST: 20% or 2,448 sentences). The balanced corpus DEV and TEST English side sentences were also machine translated through Google Translate's API to create the DEV^{GT} and TEST^{GT} sets[5] (See Section 6).

The Synthetic Corpus Given the very small number of gender-specific words in the corpus, we created a synthetic corpus consisting of short gender-inflected sentences using an Arabic morphological analyzer and generator (Taji et al., 2018). We covered 6,447 adjectives and 2,172 rational nouns (8,619 total) producing 25 different expressions for each in parallel, in masculine and feminine form. The 25 expressions consisted of simple nominal sentences, including constructions with كان وأخواتها 'Kan and her sisters', and إن وأخواتها 'Inna and her sisters'. For example, for the masculine adjective سعيد *sςyd* 'happy' we include the sentences أنا سعيد *ÂnA sςyd* 'I am happy', كنت سعيدا *knt sςydA* 'I was happy', لست سعيدا *lst sςydA* 'I was not happy', etc., and their feminine versions, respectively, أنا سعيدة *ÂnA sςydħ* 'I am happy', كنت سعيدة *knt sςydħ* 'I was happy', لست سعيدة *lst sςydħ* 'I was not happy', etc. The choice of expressions was influenced by a sample manual analysis, which we discuss in Section 5.1. In total, the synthetic corpus has 226,175 sentence pairs covering 479,518 words on each side. We use this corpus for training purposes only.

[4] We are aware that sentences with gender-specific English words, e.g., *widow and widower*, will be mismatched with the reinflected Arabic. We do not consider this to be a problem from the point of view of the spirit of the task as a whole.

[5] Google Translate's API – April 22-23, 2019.

Arabic	انا مسرورة لأني بالمدينة اليوم							
English	I'm just glad I was in town tonight							
Tokens	*AnA*	*msrwrħ*	*lÂny*			*bAlmdynħ*		*Alywm*
Gloss	I	happy	for	that	I	in	the city	today
Features	`pron+1s`	`adj+fs`	`li_prep`	`conj_sub+`	`1s_pron`	`bi_prep`	`noun+fsi`	`noun+msi`
	first person singular pronoun	*feminine singular adjective*	*preposition clitic*	*subordinating conjunction*	*first person singular pronoun*	*preposition clitic*	*feminine singular irrational noun*	*masculine singular irrational noun*

Table 3: Example of the morphological features used in automatic gender identification. In the third row, the Arabic words are presented in transliteration from left to right. The features are paired with the words they are generated from. The Gloss is the literal translation of the word. The English translation is from the OpenSubtitles corpus.

5 Automatic First-Person-Singular Gender Identification

We define the task of automatic first-person-singular gender identification as taking a sentence from our Balanced Input corpus (DEV and TEST) and predicting a label from the set {B, F, M} that indicates the gender-specificity of the first person speaker. We present four models for accomplishing this task. The first is a rule-based baseline, and the other three are machine-learning models trained on the TRAIN set of the Balanced Input corpus. Two of the machine learning models and the rule-based one make extensive use of automatically determined morphological features. All system development and parameter tuning was done using the DEV set. We report only the TEST results here. The DEV and TEST results were very similar. We discuss the morphological features next, followed by the four models, then we present our results and discussion.

5.1 The Morphological Features

We started this effort with an analysis of 100 samples from the training data: 50 from B cases, and 50 from M/F cases. For each case, we manually identified how the first person singular aspect of the task, and how the gender aspect of the task are realized linguistically. We identified three categories for the first person singular: as pro-dropped subject of a verb, as the pronoun أنا *ÂnA* 'I', and as the pronominal clitics نِي+ *+ny* and يِ+ *+y* 'me, my'. As for gender-specific forms, they were associated with adjectives and rational nouns. It was interesting to see that not a single case of the B sentences had an adjective or rational noun referring to the first person. Among the gender-specific cases, 96% of them (all but 2 cases out of 50) appeared as simple copular sentences with some variations involving كان وأخواتها 'the so-called Kan and her sisters', or إن وأخواتها 'the pseudo verbs so-called Inna and her sisters'.

For all the collected and created corpora (manually translated, synthetic and machine translated), we generated a parallel morphologically analyzed feature corpus using the MADAMIRA Arabic analysis toolkit (Pasha et al., 2014). Since MADAMIRA uses the SAMA analyzer (Graff et al., 2009) which does not provide functional features for gender and number, and rationality, we extended MADAMIRA's analyses using the work of Taji et al. (2018). We further extracted a set of specific morphological features that we determined to be relevant from the analysis we did. An example of the features associated with a sentence from our corpus is shown in Table 3.

5.2 The Rule-based Model

Given the insights developed from our initial analysis, we created a simple regular expression that operates on the morphological features discussed above. This was intended as a baseline system. The regular expression captured any context in which a first-person-singular indicator (e.g., the pronoun أنا *ÂnA* 'I', a copular verb or pseudo verb with first person subject, or a subordinating conjunction with a first person pronominal clitic) *followed by* a singular rational noun or singular adjective. The gender of the noun or the adjective determines the label for the sentence (M or F). If there is no match, the sentence receives the label B. The rule-based model does not include any lexical features and does not require any training data.

	Rule-based			Lexicalized			Delexicalized			Joint		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
B	92%	99%	95%	93%	98%	96%	93%	98%	95%	94%	98%	96%
F	96%	56%	71%	84%	65%	73%	91%	65%	75%	90%	72%	80%
M	81%	53%	64%	81%	57%	67%	80%	54%	64%	84%	62%	71%
Average	**89%**	69%	77%	86%	73%	79%	88%	72%	78%	**89%**	**78%**	**83%**

Table 4: First person singular gender identification results on TEST. P, R, and F1 refer to Precision, Recall and F1-score, respectively. Average is the *Macro Average* of values in its column.

5.3 The Machine Learning Models

As part of the development of the machine-learning models, we experimented with a very large number of learning algorithms, vectorization features, and hyper parameters. This included the use of sentence2vec embeddings trained on large collections of text, and neural models, which were not competitive due to the limited training data size. We only report below on the settings and models that were determined to be optimal during development.

We trained three models, all using logistic regression with a liblinear solver, and using features derived from the input sentences or their morphological features. For the input sentence, we normalized the Alif/Ya forms. We used *character* n-gram features from length 1 to length 7, *word* n-gram features (from length 1 to length 7), and morphological n-gram features (from length 1 to length 7). We imposed a limit of 20,000 features on each of the character, word, and morphological n-grams. All these models were implemented using the scikit-learn toolkit (Pedregosa et al., 2011).

The three machine learning models are as follows. The **Lexicalized Model** only used the input sentence character and word n-gram features as presented above. The **Delexicalized Model** only used the morphological n-gram features as presented above. And the **Joint Model** used both sets of features concatenated for each sentence.

5.4 Results and Discussion

Table 4 presents the results of the four models described above, on the blind TEST set. For each model, we report the precision (P), recall (R) and F1-score (F1) for the three labels (B, F, and M), and their macro averages. While F and M are balanced, B is about 84% of all cases.

With regards to the overall performance, the Joint model outperforms all models in terms of macro-average F1. Across all models, the precision, recall and F1 scores for B are the highest, which makes sense given the higher proportion of training data. We tried several techniques for balancing the corpus, but none improved the overall scores. Interestingly, the scores for F are always higher than M. This may be attributed to the fact that feminine is the *marked* feature in Arabic, where specific endings are easy to detect, e.g., the feminine singular suffix is ة+ +$\hbar$.

The Rule-based model is the least performing, although it is very competitive given that it was ‘human learned’ from 100 examples only (50 B, and 50 M/F). If we use a comparable training set (50 B with 50 M and F pairs), the Lexicalized, Delixicalized and Joint macro average F1 scores decrease to 56%,54%, and 60%, respectively, all below the Rule-based model. The Rule-based model also has very high precision, comparable to that of the Joint model; but it trades off with the lowest recall. This is expected and typical of rule-based models.

The Delexicalized and Lexicalized models have comparable scores and generally lower precision and higher recall than the Rule-based model. The Joint model seems to successfully increase both recall and precision (with a slight reduction of precision for F in the Delexicalized model). This suggests that the Joint model brings together complementary strengths from the Lexicalized and Delexicalized models.

6 Automatic First Person Singular Gender Reinflection

We define the task of first-person-singular gender reinflection as taking a sentence with an unspecified first-person-singular gender as input and generating two gender-appropriate versions, one masculine (B or M) and one feminine (B or F). We model the task in two ways: (a) as a single reinflection system, and (b) as a two-step identify-then-reinflect system.

6.1 Gender Reinflection as Character-based NMT

We recast the gender reinflection task as a MT task that maps the text from one source gender to a target gender. We use character-based NMT, which views the input and output sentences as sequences of characters rather than words and learn to encode and decode at the character-level. The main reason for this setup is that character-level representations are reported to be good in capturing and learning morphological aspects (Ling et al., 2015; Kim et al., 2016), which is important for a morphologically rich language like Arabic. Furthermore, character-level NMT modeling requires less vocabulary and helps reduce out-of-vocabulary by translating unseen words.

Our character-based NMT system is an encoder-decoder model that uses the general global attention architecture introduced by Luong and Manning (2015). All the NMT models we use have been trained with the OpenNMT toolkit (Klein et al., 2017) with no restriction on the input vocabulary size. Specifically, we use long short-term memory units (LSTM), with hidden units of size 500 and 2 layers in both the encoder and decoder. The model is trained for 13 epochs, using Adam with a learning rate of 0.002 and mini-batches of 40 with no pre-trained embeddings. Our char-level embeddings are learned within the training of the model.

Using different combinations of the data sets presented in Section 4, we build four reinflection models.

- **in-to-M** is a model trained to map from the Balanced Input corpus (and Synthetic F) to the Target^M corpus (and Synthetic M).
- **in-to-F** is a model trained to map from the Balanced Input corpus (and Synthetic M) to the Target^F corpus (and Synthetic F).
- **M-to-F** is a model trained to map from the Target^M corpus (and Synthetic M) to the Target^F corpus (and Synthetic F).
- **F-to-M** is a model trained to map from the Target^F corpus (and Synthetic F) to the Target^M corpus (and Synthetic M).

Single Direct Reinflection System The first two models (in-to-M and in-to-F) are used for the single system reinflection approach, where no input gender identification is needed. The in-to-M model is used to generate the M target; and the in-to-F model is used to generate the F target.

Two-step Identify-then-Reinflect System The last two models (M-to-F and F-to-M) are used in the two-step reinflection approach. We use the output of the best sentence-level Arabic gender identification model (Joint model) described in Section 5 to identify the gender of the sentence. Then, we proceed as follows. For the M target, if the identified input sentence gender is B or M, we pass the input through as is; otherwise, we reinflect the F sentence to M using the F-to-M model. And vice versa for the F target: if the identified input sentence gender is B or F, we pass the input through as is; otherwise, we reinflect the M sentence to F using the M-to-F model.

6.2 Experimental Results and Analysis

The character-based NMT reinflection models are trained using the 8,566 TRAIN sentence pairs and the 226,175 synthetic corpus sentence pairs (as discussed above). The DEV and TEST sets comprise 1,224 and 2,448 sentences, respectively. We compare two input settings: (a) the Balanced Input DEV and TEST, and (b) the English-to-Arabic Google Translate output of the English sentences corresponding the Balanced Input DEV and TEST, DEV^{GT} and TEST^{GT} (Section 4). We evaluate sentence gender reinflection against the DEV and TEST portions of the Target^F and Target^M corpora as references (also, Section 4). In addition to the single and two-step system, we include a "do-nothing" baseline that simply passes the input to the output as is.

Reinflection Evaluation Reinflection results for each setup are reported in Table 5 in terms of the MT metric BLEU (Papineni et al., 2002). It is important to note that all the reported scores are on AYT-normalized texts.[6] This normalization helps reduce the number of cases in which Alif, Ya, and Ta Marbuta are inconsistently represented in the references. The table specifies columns for Target M, and Target F, which indicate which reference is used for evaluation.

For the Balanced Input, the best performance was achieved using the two-step system. The BLEU scores are very high because most of the

[6] AYT refers to the orthographic normalization of Alif-Hamza forms, Ya/Alif-Maqsura forms, and Ta-Marbuta/Ha forms (Habash, 2010)

	Balanced Input				**Google Translate Output**			
	DEV		TEST		DEV^{GT}		TEST^{GT}	
Target	M	F	M	F	M	F	M	F
Baseline	97.12	97.12	97.05	97.05	12.23	11.52	11.91	11.18
Single	95.43	95.64	96.12	95.93	**12.92**	**12.70**	**12.54**	**12.08**
Two Step	**98.00**	**97.92**	**98.22**	**98.31**	12.27	11.83	11.96	11.42

Table 5: BLEU results (all AYT normalized) for the Baseline, Single and Select systems on the DEV and TEST sets of the Balanced Corpus Input (Input^{ar}) and English-Arabic Google Translate output (Input^{GT}) for both F and M targets.

words are not changed between input and reference. The single system in fact introduced errors that made it worse than the do-nothing baseline. While in the baseline, 91.75% of DEV sentences are fully accurate; the two-step system sentence accuracy is 95.42% (M) and 94.68% (F), a ∼40% error reduction on average.

For the Google Translate results, the single system outperforms the two-step system and the baseline. On the TEST^{GT} set, the single system has an 8% relative increase in BLEU score for Target F, and 5.3% relative increase for Target M. The BLEU scores are much lower than the Balanced Input case since the actual input to the Google MT was English and many gender and non-gender related translation errors occur. Also, we only have a single MT reference to compare against. We suspect that the reason the two-step system did not do as well is that the gender identification component was not trained with the kind of input (and noise) generated by MT systems. One possible solution in the future is to train the gender identification component with MT/NLP output specifically.

Finally, an interesting side observation from this experiment is that automatic gender identification for the Google Translate Arabic output showed a 10-to-1 bias of M versus F, compared to the 50-50 distribution in the Balanced Corpus and the 2-to-1 bias in the Original Corpus. This further confirms the bias towards masculine forms in single-output MT systems.

Error Examples in MT output We conducted a limited analysis to understand the behavior of the NMT reinflection systems. While there were many cases that were handled properly, and cases of under-correction where the input is passed to the output as is; there were also cases of over-correction where words that should maintain their form are treated as gender-specific and modified. One example is the input word للدغدغه *lldγdγh* ‘for tickling’, which is erroneously turned into the nonsense word للدغد *lldγd*. There were also a few cases of very long repetitions in the output; as well as reduced output – simply leading to sentence length mismatch. All of these phenomena are unsurprising side effects of using character-based NMT models. In our experiments, they happened infrequently, but we plan to address them in future work.

7 Conclusions and Future Work

We presented an approach to gender identification and reinflection that can be used together with any NLP application that generates text interfacing with users. We also presented the first parallel gender corpus for Arabic. We plan on making this data set publicly available for research purposes. We demonstrated the use of the corpus in benchmarking the quality of different systems for automatic gender identification and reinflection in the context of producing gender-specific machine translation. Our results are very promising, but there is still a lot to improve.

In the future, we plan to extend our work beyond first-person sentences, annotate additional data sets, and explore other techniques for gender identification and reinflection. Among the techniques to plan to explore is word-level gender identification as a sequence labeling task. For gender reinflection, we plan to consider the approaches introduced by Cotterell et al. (2017) and Zmigrod et al. (2019). We are also planning to explore opportunities of hybrid approaches that exploit existing Arabic analysis and generation systems together with more advanced machine learning models. Finally, we are interested in expanding this work to include Arabic dialects.

Acknowledgments

We thank the New York University Abu Dhabi students who helped with the annotation and translation of the Arabic parallel gender corpus: Fatima Alaydaroos, Aya Bouhelal, Fatema AlNuaimi, Muhannad AlRamlawi, and Ella Noll. We also thank Ryan Cotterell, for helpful discussions. Finally, we thank the reviewers for their extremely helpful remarks and feedback.

References

Sarah Alkuhlani and Nizar Habash. 2011. A corpus for modeling morpho-syntactic agreement in Arabic: Gender, number and rationality. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, Portland, Oregon, USA.

Christine Basta, Marta Ruiz Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *CoRR*, abs/1904.08783.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356.6334.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. *CoRR*, abs/1706.09031.

Ahmed El Kholy and Nizar Habash. 2010. Orthographic and morphological processing for English-Arabic statistical machine translation. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*, Montréal, Canada. Montréal, Canada.

Mostafa Elaraby, Ahmed Y Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. Gender aware spoken language translation applied to English-Arabic. In *Natural Language and Speech Processing (ICNLSP), 2018 2nd International Conference on*, pages 1–6. IEEE.

Joel Escudé Font and Marta R Costa-jussà. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Annette Rios Gonzales and Don Tuggener. 2017. Co-reference resolution of elided subjects and possessive pronouns in Spanish-English statistical machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 657–662.

David Graff, Mohamed Maamouri, Basma Bouziri, Sondos Krouna, Seth Kulick, and Tim Buckwalter. 2009. Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Nawar Halabi. 2016. *Modern standard Arabic phonetics for speech synthesis*. Ph.D. thesis, University of Southampton.

Google Translate Help. 2019. Get gender specific translations. `https://support.google.com/translate/answer/9179237`.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5. Citeseer.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W. Black. 2015. Character-based neural machine translation. *CoRR*, abs/1511.04586.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. Gender bias in neural natural language processing. *CoRR*, abs/1807.11714.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*.

Ngoc Quang Luong and Andrei Popescu-Belis. 2016. Improving pronoun translation by modeling coreference uncertainty. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, volume 1, pages 12–20.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, Philadelphia, USA.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1094–1101, Reykjavik, Iceland.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luís C. Lamb. 2018. Assessing gender bias in machine translation - A case study with google translate. *CoRR*, abs/1809.02208.

Ella Rabinovich, Shachar Mirkin, Raj Nath Patel, Lucia Specia, and Shuly Wintner. 2016. Personalized machine translation: Preserving original author traits. *arXiv preprint arXiv:1610.05461*.

Adam Sutton, Thomas Lansdall-Welfare, and Nello Cristianini. 2018. Biased embeddings from wild data: Measuring, understanding and removing. *CoRR*, abs/1806.06301.

Dima Taji, Jamila El Gizuli, and Nizar Habash. 2018. An Arabic dependency treebank in the travel domain. In *Proceedings of the Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*, Miyazaki, Japan.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008. Association for Computational Linguistics.

William Wright. 1955. *A grammar of the Arabic language*, volume I. Cambridge University Press.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. 2014. Large Scale Arabic Error Annotation: Guidelines and Framework. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *CoRR*, abs/1904.03310.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.

Ran Zmigrod, Sebastian J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Florence, Italy.