The Shared Task on Gender Rewriting

Bashar Alhafni,[1] **Nizar Habash,**[1] **Houda Bouamor,**[2] **Ossama Obeid,**[1]
Sultan Alrowili,[3] **Daliyah Alzeer,**[4] **Khawlah M. Alshanqiti,**[5] **Ahmed ElBakry,**[6]
Muhammad ElNokrashy,[6] **Mohamed Gabr,**[6] **Abderrahmane Issam,**[7]
Abdelrahim Qaddoumi,[8] **K. Vijay-Shanker,**[3] **Mahmoud Zyate**[9]*
[1]New York University Abu Dhabi, [2]Carnegie Mellon University in Qatar,
[3]University of Delaware, [4]Taif University, [5]Umm Alqura University,
[6]Microsoft ATL Cairo, [7]Archipel Cognitive, [8]New York University, [9]Leyton
alhafni@nyu.edu

Abstract

In this paper, we present the results and findings of the Shared Task on Gender Rewriting, which was organized as part of the Seventh Arabic Natural Language Processing Workshop. The task of gender rewriting refers to generating alternatives of a given sentence to match different target user gender contexts (e.g., female speaker with a male listener, a male speaker with a male listener, etc.). This requires changing the grammatical gender (masculine or feminine) of certain words referring to the users. In this task, we focus on Arabic, a gender-marking morphologically rich language. A total of five teams from four countries participated in the shared task.

1 Introduction

The problem of gender bias in Natural Language Processing (NLP) systems has been receiving a lot of attention across a variety of tasks such as machine translation, co-reference resolution, and dialogue systems. Research has shown that NLP systems do not only have the ability to embed societal biases, but they also amplify and propagate them in ways that create representational harms and degrade users' experiences (Sun et al., 2019; Blodgett et al., 2020). The main cause of this problem is usually attributed to inherently biased data that is used to build these systems and which mirrors the inequalities of the world we live in. Therefore, many approaches were proposed to mitigate this problem by either using counterfactual data augmentation techniques (Lu et al., 2018; Hall Maudslay et al., 2019; Zmigrod et al., 2019) or by debiasing pretrained representation that is trained on biased data (Bolukbasi et al., 2016; Zhao et al., 2018; Manzini et al., 2019; Zhao et al., 2020). However, even the most balanced of models can still exhibit and amplify bias if they are designed to produce a single text output without taking their users' gender preferences into consideration (Habash et al., 2019; Alhafni et al., 2020, 2022b). Therefore, to provide the correct user-aware output, NLP systems should be designed to produce outputs that are as gender specific as the users preferences they have access to. Recently, Alhafni et al. (2022b) introduced the task of gender rewriting, which refers to generating alternatives of a given sentence to match different target user gender contexts. To encourage more researchers to work on this problem, we organized the Shared Task on Gender Rewriting. We focus on Modern Standard Arabic (MSA), a gender-marking morphologically rich language, in contexts involving two users.[1]

This shared task was organized as part of the Seventh Arabic Natural Language Processing Workshop (WANLP), collocated with EMNLP 2022. This is the first shared task at WANLP in seven years to target a language generation problem in Arabic. A total of five teams from four countries participated in the shared task. One team contributed to a system description paper which is included in the WANLP proceedings and cited in this paper. We provide a description of all submitted systems and the approaches they use. All of the datasets created for this shared task will be made publicly available to support further research on gender rewriting.

This paper is organized as follows. We first provide a description of the shared task (§2). We then describe the data used in the shared task, including a newly created set which we used for evaluation in §3. Next, we provide a description of all submitted systems in §4 and discuss the results in §5. Finally, we discuss the lessons we learned from running this shared task and provide recommendations to the (Arabic) NLP community in §6.

*The first four authors are the shared task organizers, listed in order of contribution. The remaining authors are the shared task participants in alphabetical order.

[1]http://gender-rewriting-shared-task.camel-lab.com/

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 98 - 107

Input Sentence	Target Speaker	Target Listener	Output Sentence
سعيدة حقا بمعرفتكن يا سيدات (Really glad to know you ladies)	Masculine	Masculine	سعيد حقا بمعرفتكم يا سادة (Really glad to know you gentlemen)
	Feminine	Masculine	سعيدة حقا بمعرفتكم يا سادة (Really glad to know you gentlemen)
	Masculine	Feminine	سعيد حقا بمعرفتكن يا سيدات (Really glad to know you ladies)
	Feminine	Feminine	سعيدة حقا بمعرفتكن يا سيدات (Really glad to know you ladies)

Table 1: Example of the gender rewriting task. The input sentence has four rewritten alternatives that match the different target user gender contexts. First person gendered words are in purple and second person gendered words are in red.

2 Task Description

The task of gender rewriting was introduced by Alhafni et al. (2022b) and it refers to generating alternatives of a given Arabic sentence to match different target user gender contexts. We focus on contexts involving two users (I and/or You) – first and second grammatical persons with independent grammatical gender preferences. This requires changing the grammatical gender (masculine or feminine) of certain words referring to the users (speaker/first person and listener/second person) in the input sentence. Therefore, given an Arabic sentence as an input, the goal is to generate four different gender rewritten alternatives to match the different target user gender contexts (i.e., female speaker with a male listener, a male speaker with a male listener, a male speaker with a female listener, and a female speaker with a female listener). Table 1 shows an example of the gender rewriting problem where the input sentence is rewritten to its four gender alternatives that match the four target user gender contexts.

Notation We use the notation that is defined by Alhafni et al. (2022b). Namely, we use four elementary symbols to facilitate the discussion of this task: 1M, 1F, 2M and 2F. The digit part of the symbol refers to the grammatical person (1^{st} or 2^{nd}) and the letter part refers to the grammatical gender (Masculine or Feminine). Additionally, we use B to refer to invariant/ambiguous gender.

2.1 Shared Task Restrictions

We provided the participants with a set of restrictions for building their systems to ensure a common experimental setup and fair comparison. Participants were asked not to use any external manually labeled datasets. However, the use of publicly available unlabeled data was allowed. Participants were also not allowed to use the publicly available development and test sets of the shared task corpus for training their systems. Moreover, we provided the participants with a new blind test set that was manually annotated for this shared task. The participants were provided with the input sentences and they did not have access to the gold references. We discuss the properties and statistics of this new test set in more detail in §3.2.

2.2 Evaluation Metrics

We follow Alhafni et al. (2022b) by treating the gender rewriting problem as a user-aware grammatical error correction task and use the MaxMatch (M^2) scorer (Dahlmeier and Ng, 2012) as our evaluation metric. The M^2 scorer computes the Precision (P), Recall (R), and $F_{0.5}$ by maximally matching phrase-level edits made by a system to gold-standard edits. The gold edits are computed by the M^2 scorer based on provided gold references. We also report BLEU (Papineni et al., 2002) scores which are obtained using SacreBLEU (Post, 2018). We report the gender rewriting results in a normalized space for Alif, Ya, and Ta-Marbuta (Habash, 2010).

3 Shared Task Data

In this section, we describe the data we use in the shared task.

3.1 The Arabic Parallel Gender Corpus

We use the publicly available Arabic Parallel Gender Corpus (APGC) – a parallel corpus of Arabic sentences with gender annotations and gender rewritten alternatives of sentences selected from

OpenSubtitles 2018 (Lison and Tiedemann, 2016). The corpus comes in three versions: APGC v1.0 (Habash et al., 2019), APGC v2.0 (Alhafni et al., 2022a), and APGC v2.1 (Alhafni et al., 2022b). In this shared task, we use APGC v2.1 which contains 80,326 gender-annotated parallel sentences (596,799 words) of contexts involving first and second grammatical persons covering singular, dual, and plural constructions.

Annotations Each sentence in APGC v2.1 has one of nine labels: 1M/2M, 1M/2F, 1F/2M, 1F/2F, 1M/B, B/2M, 1F/B, B/2F, and B. Each of these labels indicates the existence (or lack thereof) of first and/or second persons gendered references in the sentence. APGC v2.1 also contains two types of word-level gender labels: basic and extended. The basic schema labels each word as B, 1F, 2F, 1M, or 2M. The basic labels refer to the *primary* person-gender marking signal in the word, which could come from the base form if gendered or the pronominal enclitic if the base form is not gendered.[2] The extended schema marks the person-genders of both the base words and their pronominal enclitics. This results in 25 word-level gender labels (e.g., B+1F, 1F+2M, etc.). All sentences containing gender-specific words have gender-rewritten parallels. The parallels of B-labeled sentences are trivial copies. Out of the 80,326 sentences in APGC v2.1, 54% (43,346) contain gendered words. In terms of word-level statistics, only 9.7% (58,066) are gender specific.

APGC v2.1 is organized into five parallel corpora that are fully aligned (1-to-1) at the word level: Input, Target 1M/2M, Target 1F/2M, Target 1M/2F, and Target 1F/2F. All five corpora are balanced in terms of gender, i.e., the number of 1F and 1M words is the same; and the number of 2F and 2M words is the same. The Input corpus contains sentences with all possible word types (B, 1F, 2F, 1M, 2M). The Target 1M/2M corpus contains sentences that consist of B, 1M, 2M words; the Target 1F/2M corpus contains sentences that consist of B, 1F, 2M words; the Target 1M/2F corpus contains sentences that consist of B, 1M, 2F words; and the Target 1F/2F corpus contains sentences that consist of B, 1F, 2F words.

[2]Changing the grammatical gender of Arabic words involves either changing the form of the base word, changing the pronominal enclitics that are attached to the base word, or a combination of both (Alhafni et al., 2022b)

Splits We use Alhafni et al. (2022a)'s splits: 57,603 sentences (427,523 words) for training (TRAIN), 6,647 sentences (49,257 words) for development (DEV), and 16,076 sentences (120,019 words) for testing (TEST).

3.2 Blind Test Set

To ensure fair comparison between all participants, we manually annotated a new blind test set to evaluate their systems. We plan on making this new test set publicly available. We will refer to this set as *Blind Test* throughout the paper.

Data Selection We followed the same procedure that was used in (Habash et al., 2019) and (Alhafni et al., 2022a) to create the APGC. We selected sentences from the English-Arabic OpenSubtitles 2018 dataset (Lison and Tiedemann, 2016) by extracting sentence pairs that include first or second pronouns on the English side. We annotated 5,000 sentences such that 1,061 (21.2%) include first and second person pronouns, 2,116 (42.3%) include only first person pronouns, and 1,823 (36.5%) include only second person pronouns. The sentences were selected such: (a) they do not overlap with any of the sentences that are in APGC; and (b) their proportions approximate the distribution of the Arabic-English pairs in the OpenSubtitles 2018 dataset that have first or second persons pronouns on the English side (Alhafni et al., 2022a).

Data Annotation We conducted the annotation through a linguistic annotation firm that hired professional linguists to complete the task.[3] We provided them with the same annotation guidelines that were defined in Alhafni et al. (2022a) and used to annotate the APGC. That is, the annotators were asked to identify the genders of the first and second person references in each sentence. In the case a gendered reference exists, the annotators were asked to copy the sentence and modify it to obtain the opposite gender forms. As was done when creating the APGC, the modifications are strictly limited to morphological reinflections and word substitutions. Therefore, the total number of words is maintained along with a perfect alignment between each sentence and its parallel opposite gender forms. This allowed us to obtain basic and extended word-level gender annotations automatically as was done by Alhafni et al. (2022a,b).

[3]https://www.ramitechs.com/

(a) Original Test Set						(b) Balanced Test Set					
Sentences		Label	Rewriting Label			Input	Target 1M/2M	Target 1F/2M	Target 1M/2F	Target 1F/2F	Sentences	
2,818	56.4%	B				B	B	B	B	B	2,818	38.5%
91	1.8%	1F/B	1M/B			1F/B	1M/B	1F/B	1M/B	1F/B	263	3.6%
172	3.4%	1M/B	1F/B			1M/B	1M/B	1F/B	1M/B	1F/B	263	3.6%
559	11.2%	B/2F	B/2M			B/2F	B/2M	B/2M	B/2F	B/2F	1,851	25.3%
1,292	25.8%	B/2M	B/2F			B/2M	B/2M	B/2M	B/2F	B/2F	1,851	25.3%
8	0.2%	1F/2F	1M/2F	1F/2M	1M/2M	1F/2F	1M/2M	1F/2M	1M/2F	1F/2F	68	0.9%
21	0.4%	1F/2M	1M/2M	1F/2F	1M/2F	1F/2M	1M/2M	1F/2M	1M/2F	1F/2F	68	0.9%
13	0.5%	1M/2F	1F/2F	1M/2M	1F/2M	1M/2F	1M/2M	1F/2M	1M/2F	1F/2F	68	0.9%
26	1.4%	1M/2M	1F/2M	1M/2F	1F/2F	1M/2M	1M/2M	1F/2M	1M/2F	1F/2F	68	0.9%
5,000											7,318	

Table 2: **Sentence-level** statistics of the original (a) and the balanced Blind Test set (b) with its five versions.

(a) Original Test Set						(b) Balanced Test Set					
Words		Label	Rewriting Label			Input	Target 1M/2M	Target 1F/2M	Target 1M/2F	Target 1F/2F	Words	
32,548	91.8%	B				B	B	B	B	B	46,550	88.3%
138	0.4%	1F	1M			1F	1M	1F	1M	1F	452	0.9%
241	0.7%	1M	1F			1M	1M	1F	1M	1F	452	0.9%
738	2.1%	2F	2M			2F	2M	2M	2F	2F	2,624	5%
1,805	5.1%	2M	2F			2M	2M	2M	2F	2F	2,624	5%
35,470											52,702	

Table 3: **Word-level** statistics of the original (a) and the balanced Blind Test set (b) with its five versions.

Data Statistics Table 2(a) includes the statistics of the newly annotated sentences. This constitutes the Original Blind Test set. Out of all sentences in this set, 2,818 (56.4%) are labeled as B. There are 1,851 sentences (37%) that include only second-person gendered references (B/2F and B/2M). This is about five times more than sentences with only first-person gendered references (1F/B and 1M/B), which accounts for 5.3% (263 sentences) of all sentences. Moreover, the number of sentences including first or second person masculine references is more than the ones including feminine references (1,292 B/2M vs 559 B/2F, and 172 1M/B vs 91 1F/B). There are 68 (1.4%) sentences that have both first and second gendered references. These results are consistent with APGC v2.0 (Alhafni et al., 2022a). The basic word-level statistics of the Original Blind Test set are presented in Table 3(a). We evaluated inter-annotator agreement (IAA) on 500 sentences between two annotators. The IAA in terms of nine sentence-level labels (B, M, F, for 1st and for 2nd persons, e.g., 1M/2F or 1B/2M) was 98.0%. Agreement in exact match on gender rewriting alternatives was 96.2%.

Similarly to Habash et al. (2019) and Alhafni et al. (2022a), to ensure equal gender representation in our dataset, we force balance the corpus by adding the manually rewritten sentences to the test set and using their original forms as their rewritten forms. This constitutes the Balanced Blind Test set. The sentence-level statistics of the balanced set are presented in Table 2(b). This corpus has 7,318 sentences in total. Out of all sentences, 38.5% (2,818) are marked as B, whereas sentences with gendered references constituted 61.5% (4,500 sentences). Moreover, we organize the data into five balanced corpora as was done in APGC v2.0 (§3.1). The basic word-level statistics of the Balanced Blind Test set are presented in Table 3(b). The extended word-level statistics of the Balanced Blind Test set are in Table 4.

Word Gender Label		Words	
Basic	Extended		
B	B	46,550	88.3%
1M	1M+B	445	0.8%
	B+1M	7	0.01%
1F	1F+B	445	0.8%
	B+1F	7	0.01%
2M	2M+B	2,464	4.7%
	B+2M	144	0.3%
	2M+2M	16	0.03%
2F	2F+B	2,464	4.7%
	B+2F	144	0.3%
	2F+2F	16	0.03%
		52,702	

Table 4: Statistics of the extended word-level gender of the Blind Test set.

Team	Affiliation
Cairo Team	Microsoft ATL Cairo, Egypt
CasaNLP	Archipel Cognitive, and Leyton, Morocco
Distinguishers	Taif University, and Umm Alqura University, KSA
Qaddoumi	New York University, USA
UDEL-NLP	University of Delaware, USA

Table 5: List of the five teams who participated in the gender rewriting shared task.

Team	Gender ID	Special Preprocessing	Pretrained Models
Cairo Team	Word		CAMeLBERT MSA + AraT5-MSA
CasaNLP	Word	Word Side Constraints	CAMeLBERT MSA + AraT5-MSA
Distinguishers	Word	Morphological Features	CAMeLBERT MSA + AraBERT
Qaddoumi		Romanization	T5
UDEL-NLP		Sentence Side Constraints	ArabicT5

Table 6: Approaches and techniques used by the participants. Gender ID refers to gender identification. Special Preprocessing refers to any form of preprocessing done to modify the data (e.g., adding side-constraints, morphological processing, transliteration, etc.). Pretrained Models indicates the usage of pretrained models as part of the system.

4 Participants and Systems

Five teams from four countries participated in the shared task. Table 5 presents the names of the participating teams and their affiliations. Next, we describe the approaches the participants took to develop their gender rewriting systems.

4.1 Systems Descriptions

All participants leveraged pretrained language models such as AraBERT (Antoun et al., 2020), CAMeLBERT (Inoue et al., 2021), T5 (Raffel et al., 2020), and AraT5 (Nagoudi et al., 2022), when developing their systems. Some systems consisted of multiple components to do gender identification and then rewriting as was done in Alhafni et al. (2022b), while others treated the problem as a traditional sequence-to-sequence (Seq2Seq) task. Table 6 presents a summary of the different approaches used to develop the different systems.

Cairo Team The system developed by Cairo Team was a multi-step system consisting of the following components: (a) a word-level gender identification classifier; (b) a word-level person identification classifier; and (c) sentence-level gender rewriting Seq2Seq models. The word-level classifiers were built by fine-tuning CAMeLBERT MSA (Inoue et al., 2021), on the training data of APGC v2.1. Cairo Team used the *basic* word-level annotations in the corpus to build these two classifiers. Concretely, the gender identification component was trained to identify the gender of each word as M, F, or B, whereas the person identification component was trained to classify the person which the word refers to as 1st, 2nd, or none. For the sentence-level Seq2Seq models, Cairo Team built four different models, one for each target user gender context (i.e., 1M/2M, 1F/2M, 1M/2F, 1F/2F), by fine-tuning AraT5-MSA$_{\text{BASE}}$ (Nagoudi et al., 2022).

During inference, the input sentence is passed to the word-level classifiers to get the gender and person labels for each word. These predicted labels indicate which words need to be rewritten based on the compatibility between the labels and the target user gender contexts. Then, the same input sentence is passed to each Seq2Seq model to get its rewritten forms. After that, Cairo Team uses a simple heuristic to reduce the noise that could be generated in the outputs of the Seq2Seq models and to ensure that only the necessary gendered words are changed. To do so, Cairo Team generates all subsets of possible trigrams for each gendered word that needs to be changed in the input. Then, they search for partial matches of these trigrams in the Seq2Seq model generated sentences and pick the generated words that have the highest match. The intuition behind this approach is that: (a) the Seq2Seq model would benefit from seeing the entire sentence to apply in-context word gender rewriting; and (b) most of the gendered words in the APGC v2.1 (96.9%) are due to morphological

inflections, which allows the matching heuristic to have a high coverage.

The fine-tuning of the models was done using Hugging Face's Transformers (Wolf et al., 2020). Both the word-level gender and person identification classifiers were fine-tuned on a single GPU for 10 epochs with a maximum sequence length of 128, a batch size of 32, and a learning rate of 1e-4. The sentence-level gender rewriting component was fine-tuned on a single GPU for 30 epochs with a maximum sequence length of 128, a batch size of 16, and a learning rate of 1e-3. Checkpoints were saved every 1000 steps and at the end of fine-tuning, the best checkpoint was picked based on the development set.

CasaNLP The system introduced by CasaNLP was also a multi-step system that consists of word-level gender identification and sentence-level gender rewriting. For gender identification, the team used the gender identification model that was developed and released by Alhafni et al. (2022b).[4] The gender identification component takes the input sentence and assigns an *extended* gender label to every word in the input. After that and based on the compatibility between the labels and the target user gender contexts, CasaNLP adds word-level target gender labels as *side-constraints* (Sennrich et al., 2016) to the words that need to be rewritten in the input sentence (e.g., سعيد [2F] أنا). They do this preprocessing step across all sentences in APGC v2.1. Then, they fine-tune AraT5-MSA$_{\text{BASE}}$ on the preprocessed sentences in TRAIN. The intuition here is that the model should learn to only rewrite the words that are marked in the input. The team follows the same procedure during inference to generate the gender rewritten alternatives.

The fine-tuning of the models was done using Hugging Face's Transformers. The sentence-level gender rewriting system was fine-tuned for 10 epochs with a maximum sequence length of 64, a batch size of 32, and a learning rate of 1e-3 with 4 gradient accumulation steps.

Distinguishers This team introduced a multi-step system that does word-level gender identification and out-of-context word-level gender rewriting. For gender identification, they used the model that was developed and released by Alhafni et al. (2022b).[4] For gender rewriting, the team developed an out-of-context word-level Seq2Seq model. The model followed the approach introduced in BERT-fused (Zhu et al., 2020), where they first use AraBERT (Antoun et al., 2020) to extract representations for the input word, and then the representations are fused with each layer of the encoder and decoder of a standard Transformer model (Vaswani et al., 2017). The model was trained on gendered words present in APGC v2.1. They also explored adding morphological features to their Seq2Seq model. They used CAMeLTools (Obeid et al., 2020) to do morphological tokenization on the words and get their part-of-speech tags. They added the tags as side-constraints to each word. During inference, they first run the gender-identification component over the input sentence to get predicted gender labels for each word. Then for each word that needs to be rewritten, they pass it to the Seq2Seq model to get its gender alternative.

The out-of-context word-level gender rewriting model was built using Simple Transformers.[5] The model was fine-tuned on a single GPU for 5 epochs with a maximum sequence length of 25, a learning rate of 1e-5, and a batch size of 32.

Qaddoumi The approach this team took to build their gender rewriting system relied on romanizing the Arabic text and using an *English* pre-trained model. The team preprocessed the data in APGC v2.1 by using the Safe Buckwalter transliteration scheme (Buckwalter, 2002; Habash, 2010). They continue fine-tuning a grammatical error correction model that was originally built by fine-tuning T5 (Raffel et al., 2020) on the JFLEG corpus (Napoles et al., 2017).[6] When producing the final outputs, they convert the text back to Arabic script.

The sentence-level gender rewriting system was fine-tuned using the Happy Transformer library on a single GPU for 5 epochs with a maximum sequence length of 1024, a batch size of 32, and a learning rate of 5e-5.[7]

UDEL-NLP The system developed by UDEL-NLP was at the sentence-level and based on T5. The team introduced a new Arabic T5 model called ArabicT5 (Alrowili and Vijay-Shanker, 2022),

[4] https://github.com/CAMeL-Lab/gender-rewriting/

[5] https://github.com/ThilinaRajapakse/simpletransformers

[6] https://huggingface.co/vennify/t5-base-grammar-correction

[7] https://github.com/EricFillion/happy-transformer

Team	Precision	Recall	$F_{0.5}$	BLEU
Cairo Team	**76.26 (1)**	72.27 (3)	**75.42 (1)**	**94.89 (1)**
CasaNLP	51.05 (4)	**84.60 (1)**	55.45 (4)	86.06 (4)
Distinguishers	20.93 (5)	19.03 (5)	20.52 (5)	84.89 (5)
Qaddoumi	56.49 (3)	77.06 (2)	59.68 (2)	88.53 (3)
UDEL-NLP	57.10 (2)	68.61 (4)	59.08 (3)	91.02 (2)
Alhafni et al. (2022b)	**88.50**	**84.98**	**87.78**	**97.62**

Table 7: Results on the Blind Test set. Numbers in parentheses are the ranks.

which was pretrained on MSA by using an efficient T5 implementation (Tay et al., 2021). They fine-tuned the ArabicT5 model by adding side-constraints to the beginning of each sentence to indicate the target users' gender, and appending an <eos> to each sentence. The team follows the same preprocessing steps during inference.

The sentence-level gender rewriting system was built by fine-tuning ArabicT5 using Hugging Face's Transformers on a single GPU for 70 epochs, a maximum sequence length of 512, a batch size of 32, and a learning of 1e-4.

5 Results

Table 7 presents the results on the newly annotated Blind Test set. The last row is for the state-of-the-art system by Alhafni et al. (2022b). The best result in terms of $F_{0.5}$ is achieved by the Cairo Team (75.42), the official winner of the shared task. This is mainly due to their high score in precision (76.26). Qaddoumi comes in second place achieving an $F_{0.5}$ of 59.68, followed by UDEL-NLP in third place with 59.08 in $F_{0.5}$. In fourth place, CasaNLP achieves an $F_{0.5}$ score of 55.45 with the highest recall of 84.60. Distinguishers comes in fifth place, achieving 20.52 in $F_{0.5}$. It is worth noting that none of the systems is able to beat the previously published system by Alhafni et al. (2022b) applied to the new Blind Test.

Error Analysis We conducted a simple error analysis over the outputs of all system on the Blind Test set. Given that most teams employed sentence-level Seq2Seq models when developing their gender rewriting systems, we suspected that the outputs will be noisy since sentence-level models will not guarantee that changes are only applied to gendered words, or maintain the word-level parallelism between the input and output. Table 8(a) presents the relative difference in the number of generated words for each team in comparison with the Blind Test reference; and Table 8(b) presents their correlation with the shared task metrics. None of the teams maintained the total number of words. We observe a strong negative correlation between the absolute value of relative word count differences and the evaluation metrics – almost -51% correlation with $F_{0.5}$, and -78% correlation with recall.

(a)

Team	Word Δ
Cairo Team	0.80%
CasaNLP	-0.02%
Distinguishers	1.28%
Qaddoumi	-0.63%
UDEL-NLP	0.05%

(b)

Metric	Correl
Precision	-42.95%
Recall	-77.56%
$F_{0.5}$	-50.86%
BLEU	-11.86%

Table 8: (a) The relative difference in the number of generated words for each team in comparison with the Blind Test reference. (b) The Pearson correlation of the shared task metrics in Table 7 with the *absolute* values of Word Δ.

After inspecting the outputs of the submitted systems, we noticed that much of the noise was due to not handling punctuation correctly. We removed the punctuation from all the outputs and evaluated the systems in this space. Table 9 shows the results on the Blind Test set after removing the punctuation. The scores of all teams went up significantly, with the exception of Distinguishers. The highest increase of 31.6 points in $F_{0.5}$ is in the case of CasaNLP. In terms of the ranks of the systems in this unofficial evaluation space, CasaNLP is the best performer and they achieve 87.04 in $F_{0.5}$. They also have the highest precision, recall, and BLEU scores. The Cairo Team comes in second place with an $F_{0.5}$ of 83.76, followed by UDEL-NLP who achieves an $F_{0.5}$ of 70.22. Qaddoumi and Distinguishers are in fourth and fifth places, achieving 63.35 and 20.41 in $F_{0.5}$, respectively.

Team	Precision	Recall	$F_{0.5}$	BLEU
Cairo Team	87.34 (2)	71.98 (3)	83.76 (2)	95.74 (2)
CasaNLP	**87.72 (1)**	**84.45 (1)**	**87.04 (1)**	**97.18 (1)**
Distinguishers	20.81 (5)	18.96 (5)	20.41 (5)	84.11 (5)
Qaddoumi	60.68 (4)	76.90 (2)	63.35 (4)	89.06 (4)
UDEL-NLP	70.67 (3)	68.50 (4)	70.22 (3)	91.99 (3)
Alhafni et al. (2022b)	**88.38**	**84.87**	**87.65**	**97.30**

Table 9: Results on the Blind Test set of after removing the punctuation. Numbers in parentheses are the ranks.

6 Outlook and Lessons Learned

We organized this shared task on gender rewriting for Arabic to raise awareness in the Arabic NLP community of the problem of gender bias in Arabic NLP systems, and to encourage the community to come up with new approaches to alleviate this problem. Although the shared task received some interest from the community, the participation was limited[8] when compared to other shared tasks organized at recent editions of WANLP[9] or OSACT.[10] We believe that this is due to a couple of factors.

First is the **skewed interest towards sentence-level classification** tasks within the Arabic NLP community and the lack of novel open-vocabulary sequence transduction tasks. For instance, most of the shared tasks organized at WANLP over the past few years focused on sentence-level classification to tackle dialect identification: MADAR and NADI (Bouamor et al., 2019; Abdul-Mageed et al., 2020, 2021); or Arabic sarcasm detection: ArSarcasm (Abu Farha et al., 2021). The last shared task that featured a generation problem in Arabic was the QALB shared task on grammatical error correction (Rozovskaya et al., 2015).

We acknowledge the importance of working on sentence-level classification problems, but there are many natural language generation tasks where Arabic is still lagging behind compared to other languages. Examples of such tasks include dialectal machine translation, grammatical error correction, text simplification, and style transfer, to name a few. We envision that the development of resources and models for such tasks would re-spark the interest of the Arabic NLP community in a wide range of exciting, yet unsolved problems in Arabic NLP.

Second is the **novelty and difficulty** of the gender rewriting problem compared to other conventional sequence transduction tasks. Approaching the problem correctly requires developing controlled generation models that are able to make subtle, yet complex and grammatically correct, edits at the word level. In retrospect, we recognize that we could have organized this shared task as two subtasks: one on gender identification at the word or sentence levels, and the other on sentence-level gender rewriting. This could have served as a bridge between classification and generation tasks, too, and allowed more people to participate for part if not the whole of the task. As such, we recommend that organizers of novel and nontraditional tasks to break the problem into subtasks to encourage more participation.

Lastly, the main goal of participating in a shared task is to learn about a new problem by introducing an interesting solution, which could benefit the community as a whole, as a positive or negative result. Being on top of the leaderboard should not be the only motive; we encourage organizers within the community to echo this sentiment when running their shared tasks.

Limitations and Ethical Considerations

Our intention of organizing this shared task is to increase the inclusiveness of NLP applications that deal with gender-marking morphologically rich languages. However, we acknowledge that, like all NLP technologies, developing systems for gender identification and rewriting could be used in malicious ways to discriminate against, or erase, certain identities in certain contexts. We also acknowledge that by limiting the choice of gender expressions to grammatical gender, we exclude alternatives such as non-binary gender or no-gender expressions. We are not aware of any sociolinguistics published research that discusses such alternatives for Arabic. We stress on the importance of adapting Arabic NLP models to new gender alternative forms as they emerge as part of the language usage.

[8] While 15 teams registered for the shared task initially, only five of them ended up participating.

[9] http://www.arabic-nlp.net/

[10] https://osact-lrec.github.io/

References

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 296–305, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2020. Gender-aware reinflection using linguistically enhanced neural models. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 139–150, Barcelona, Spain (Online). Association for Computational Linguistics.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022a. The Arabic parallel gender corpus 2.0: Extensions and analyses. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1870–1884, Marseille, France. European Language Resources Association.

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022b. User-centric gender rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 618–631, Seattle, United States. Association for Computational Linguistics.

Sultan Alrowili and K. Vijay-Shanker. 2022. Generative approach for gender-rewriting task with ArabicT5. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop*, Abu Dhabi, UAE. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

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*, volume 29. Curran Associates, Inc.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

Tim Buckwalter. 2002. Buckwalter Arabic morphological analyzer version 1.0. Linguistic Data Consortium (LDC) catalog number LDC2002L49, ISBN 1-58563-257-0.

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy.

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

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, Hong Kong, China.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

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.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and*

Short Papers), pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Alla Rozovskaya, Houda Bouamor, Nizar Habash, Wajdi Zaghouani, Ossama Obeid, and Behrang Mohit. 2015. The second QALB shared task on automatic text correction for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 26–35, Beijing, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. 2021. Scale efficiently: Insights from pre-training and fine-tuning transformers.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy.