Generative Approach for Gender Rewriting Task with ArabicT5

Sultan Alrowili
Department of Computer Science
University of Delaware
Newark, Delaware, USA
alrowili@udel.edu

K.Vijay-Shanker
Department of Computer Science
University of Delaware
Newark, Delaware, USA
vijay@udel.edu

Abstract

Addressing the correct gender in generative tasks (e.g., Machine Translation) has been an overlooked issue in the Arabic NLP. However, the recent introduction of the Arabic Parallel Gender Corpus (APGC) dataset has established new baselines for the Arabic Gender Rewriting task. To address the Gender Rewriting task, we first pre-train our new Seq2Seq ArabicT5 model on a 17GB of Arabic Corpora. Then, we continue pre-training our ArabicT5 model on the APGC dataset using a newly proposed method. Our evaluation shows that our ArabicT5 model, when trained on the APGC dataset, achieved competitive results against existing state-of-the-art methods. In addition, our ArabicT5 model shows better results on the APGC dataset compared to other Arabic and multilingual T5 models.

1 Introduction

In many generative downstream tasks in Arabic NLP, such as Machine Translation and chatbot applications, addressing the correct gender is crucial to increase the quality of the generated text to reach human-level performance. This also leads to having a generated text that is less biased and discriminating against specific gender. Moreover, when used in Translation and chatbot applications, generative models such as T5 (Raffel et al., 2019), and BART (Lewis et al., 2020) may adopt a gender bias, which they learn from the pre-training corpora. Thus, the Gender Rewriting downstream task has recently received more attention in Arabic NLP. This attention can be seen with the introduction of the Arabic Parallel Gender Corpus (APGC) dataset (Alhafni et al., 2022a).

Current state-of-the-art methods to address the Gender Rewriting task uses a multi-stage model consisting of rule-based, morphological analyzer, and encoder-decoder GRU model (Alhafni et al., 2022b). However, one issue with using a multi-stage model is that it increases the complexity of the model. This motivates us to seek a more simple alternative approach. In this work, we hypothesize that generative models such as T5 and BART could address the gender rewriting problem when trained on the APGC dataset. Thus, in this work, we introduce a novel method to address the Gender-Rewriting task through our ArabicT5 model, a model that we pre-trained on a collection of Arabic corpora.

Thus, our contributions in this work can be summarized in the following points:

- We introduce ArabicT5: a new Arabic T5 model pre-trained on a 17GB of Arabic corpora, including Arabic Wikipedia and Arabic News articles. This model has many applications beyond the scope of this work, such as Question Answering, Text Classification, Question Generation, Machine Translation, and Text Summarization. We also released our ArabicT5 model and our codes to the public community. [1].

- We introduce a new approach in the Arabic NLP that uses Seq2Seq models to address the Gender-Rewriting task.

- We evaluate and compare our approach with our ArabicT5 against AraT5 (Nagoudi et al., 2022) , mT5 model: the multilingual variant of T5 (Xue et al., 2021), and the multi-step gender rewriting model by Alhafni et al. (2022b). We also show through our analysis how design factors such as the pre-training corpora affect the evaluation performance.

[1]Our ArabicT5 models can be accessed at https://huggingface.co/sultan/ArabicT5-Base, https://huggingface.co/sultan/ArabicT5-Large, https://huggingface.co/sultan/ArabicT5-xLarge and our GitHub page https://github.com/salrowili/ArabicT5.

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 491 - 495

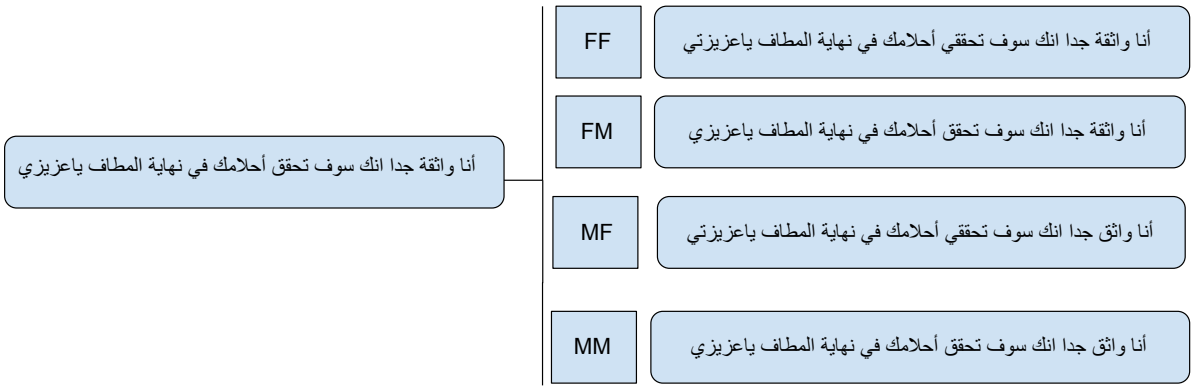

Figure 1: Example of the Gender-Rewriting task where we address different targeted genders. [FF: Female-to-Female, FM: Female-to-Male, MF: Male-to-Female, MM: Male-to-Male] .

2 Background

In this section, we will first explain the APGC dataset. Then, we will have an overview of the current state-of-the-art model; the multi-step gender rewriting model (Alhafni et al., 2022b). Then we will explain the T5, mT5 (Xue et al., 2021), and AraT5 models (Nagoudi et al., 2022).

2.1 Arabic Parallel Gender Corpus

Arabic Parallel Gender Corpus (APGC) (Alhafni et al., 2022a) is a new dataset introduced recently to address gender bias in natural language processing (NLP) applications. This dataset aims to address gender identification and rewriting sentence where the context involves one or two users (I and/or you). In Figure 1, we illustrate the structure of the APGC dataset.

2.2 The Multi-step Model Approach

The Multi-step Model (Alhafni et al., 2022b) represents the state-of-the-art model to address the Arabic Gender-Rewriting task. The Multi-step Model consists of multiple-stages including: (1) Gender Identification (GID), (2) Corpus-based Rewriter (CorpusR) (3) Morphological Rewriter (MorphR), and (4) NeuralR. The Gender Identification component aims to classify the word-level gender label for each word in the sentence using Arabic Transformer-Based models. The Corpus-based Rewriter (CorpusR) uses a bigram maximum likelihood estimator that uses the context to re-write desired word-level target gender. On the other hand, Morphological Rewriter (MorphR) component uses the morphological generator included in the CAMeL Tools. The last component in this Multi-step Model is the Neural Rewriter (NeuralR), a character-level attention-based encoder-decoder model. For both the encoder and decoder, it uses a GRU model (Chung et al., 2014).

2.3 T5

There are two common approaches where language models address downstream tasks. The first approach is the extractive approach, where we fine-tune the language model to extract specific spans (e.g., Question Answering) or predict a class in the Text Classification problem. Language Models that follow the extractive approach are models such as BERT (Devlin et al., 2019), ELECTRA (Clark et al., 2020), and ALBERT (Lan et al., 2019). On the other hand, generative models such as BART (Lewis et al., 2020), T5 (Raffel et al., 2019), and XLENT (Yang et al., 2019) are built to generate the target text to address the downstream task. For example, in T5, the Text-to-Text Transfer Transformer model, instead of extracting the spans that define the answer boundary, it generates the answer from the model parameters.

2.4 mT5

The mT5 model (Xue et al., 2021) is a multilingual variant of T5, which was pre-trained on the new Common Crawl-based dataset that consists of 6.3T tokens covering 101 languages. The mT5 model also uses a large vocabulary file that consists of 250K tokens.

2.5 AraT5

AraT5 (Nagoudi et al., 2022) is a newly introduced Arabic Language Model that pre-trains T5 on a collection of Arabic Corpora. AraT5 was pre-trained

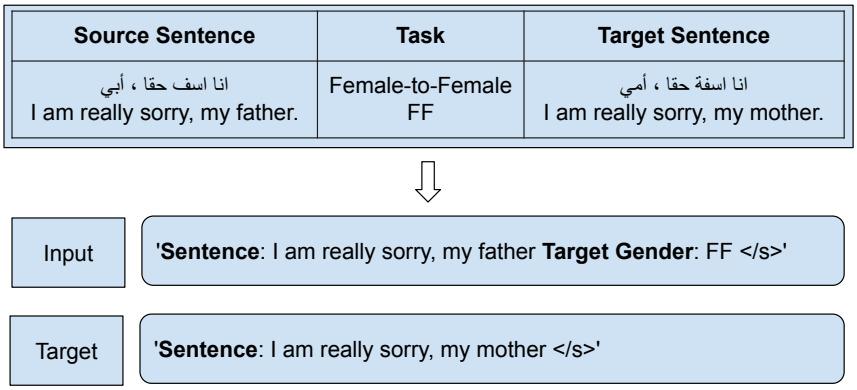

Figure 2: Example of our proposed method to address the Gender-Rewriting task. Both "Sentence" and "Target Gender" are used as tags and they are part of the input and target sentences.

for 80 days on the TPUv3-8 unit with a maximum sequence length of 128. AraT5 shows promising results on downstream tasks against Multi-lingual mT5 model. In our evaluation, we use three variants of AraT5, including:

- **AraT5-MSA$_{\text{Base}}$**: pre-trained on Modern Standard Arabic (MSA) corpora (70GB) which include a collection of Arabic News articles and Arabic websites.
- **AraT5-Twitter$_{\text{Base}}$**: pre-trained on Arabic Twitter Dataset (178GB).
- **AraT5$_{\text{Base}}$**: pre-trained on 248GB of Arabic Corpora including Modern Standard Arabic (MSA) corpora (70GB) and dataset from Twitter (178GB).

3 Method

In this section, we will first explain how we build our new T5 model. Next, we will explain our method to address the Gender Rewriting task and the details of our environmental and evaluation setup.

3.1 Pre-training our ArabicT5 model

We build our ArabicT5 model by pre-training T5 model on a collection of Arabic Corpora including Arabic Wikipedia, News Articles (El-Khair, 2016), Hindawi Books [2] and Marefa encyclopedia [3]. We pre-train our ArabicT5 model using an efficient T5 implementation (Tay et al., 2021), which reduces pre-training and fine-tuning costs by studying T5 design factors (e.g., hidden size layers, attention heads, attention layers). We build our vocabulary using the SentencePiece model (Kudo and Richardson, 2018) and choose our vocabulary size as 32K tokens. In contrast to the AraT5 model, which only introduces the base model, we introduce based, large and xlarge models.

Our ArabicT5$_{\text{base}}$ model has 512 hidden size layers, eight attention heads, and 20 attention layers. We pre-train our ArabicT5$_{\text{base}}$ for 256K steps with a batch size of 256 (131,072 tokens) on TPUv3-32 unit. On the other hand, our ArabicT5$_{\text{large}}$ model uses 768 hidden size layers, 12 attention heads, and 16 attention layers. Moreover, we pre-train ArabicT5$_{\text{xlarge}}$ model which differ from ArabicT5$_{\text{large}}$ that it has more attention layers (36). We pre-train both our ArabicT5$_{\text{large}}$ and ArabicT5$_{\text{xlarge}}$ for 512K steps with a batch size of 512 (262,144 tokens) on TPUv3-128. For all models, we maintain all other settings set by (Tay et al., 2021) (e.g., learning rate, warm-up steps). We use the official TensorFlow implementation of T5 to pre-train our base and large models.

3.2 Preparing The Dataset

T5 models use a unified Text-to-Text framework that addresses all downstream tasks in Text-to-Text format as an input text and target text. For example, to address Text Classification problems such as Sentiment Analysis, we will add the sentence as the input text and the class (positive/negative) as the target text. To address the Gender Rewriting problem, we add the original sentence and targeted Gender (e.g., FF, FM, MF, MM) in the input text. Then we will add the output sentence which addresses the targeted gender in the target text. We will also add the flag </s> to mark the end of the sequence in both the input and target text. We illustrate our

[2] https://www.hindawi.org/books
[3] https://www.marefa.org/

Model	P	R	$F_{0.5}$	B
The Multi-Step Gender Rewriting Model (Alhafni et al., 2022b)	**88.8**	86.8	**88.3**	**98.1**
$mT5_{Base}$	71.6	82.0	73.4	97.5
$AraT5_{Base}$	72.8	83.6	74.7	97.7
$AraT5\text{-}MSA_{Base}$	72.6	83.8	74.6	97.7
$AraT5\text{-}Twitter_{Base}$	72.2	82.1	74.0	97.6
$ArabicT5_{Base}$ (ours)	72.1	85.5	74.4	97.7
$ArabicT5_{Large}$ (ours)	72.7	86.2	74.4	98.0
$ArabicT5_{XLarge}$ (ours)	73.0	**87.1**	75.4	98.0

Table 1: Evaluation Result of mT5, AraT5, ArabicT5 on the DEV set of APGC v2.1. [P: Precision, R: Recall, B: BLEU score] . We use reported results for the Multi-Step Gender Rewriting Model and generate the result for all other models.

method in Figure 2.

3.3 Experimental Setup

We fine-tune our ArabicT5, mT5, and AraT5 using the PyTorch XLA library https://github.com/pytorch/xla, which allows us to use Torch code on the TPUv3-8 unit. We fine-tune all models for 70 epochs with a learning rate of 1e-4. For evaluation, we follow a similar approach to (Alhafni et al., 2022b) by using the BLEU (Bilingual Evaluation Understudy) and MaxMatch (M2) scorer (Dahlmeier and Ng, 2012)[4]. We also adapt the same normalization script adapted by Gender Rewriting Shared Task [5].

4 Results and Discussion

In Table 1, we show the evaluation of both AraT5, mT5, and our ArabicT5 model with different scales (base, large, xlarge). In addition, we show the evaluation score of the current state-of-the-art model: The Multi-step Model by Alhafni et al. (2022b). We explain The Multi-step Model in detail in Section 2.2. This evaluation in Table 1 aims to compare the performance between single-stage seq2seq T5-based models against the current multi-stages state-of-the-art model.

We can observe from the results that there is a significant gap in performance between the Multi-Step Model and other Seq2Seq T5-based models. This gap is caused by the fact that these Seq2Seq models use a single-stage sentence-level approach. However, observe the close gap in blue score between all models in Table 1, which may caused by the fact that in the Arabic language, we only change a few letters in the sentence to address the right gender. In addition, we can attribute the significant gap in both Precision and $F_{0.5}$ scores between The Multi-Step Model and other Seq2Seq models to the multi-stage components used by Alhafni et al. (2022b). It also worth noting that our largest ArabicT5 models achieve the best recall score among all models showing the potential of seq2seq models.

On the other hand, the evaluation comparison between T5-based models, including mT5, AraT5, and our ArabicT5, shows how pre-training corpora significantly affect the performance in the Gender-Rewriting task. Our ArabicT5, pre-trained on modern classical Arabic corpora (Arabic Encyclopedias and Arabic news articles), shows superiority against other models that use Arabic website collection and Twitter Datasets.

We use our best-performing model $ArabicT5_{xLarge}$ to submit our prediction for the blind test dataset of Gender Rewriting task (Alhafni et al., 2022c) at the Seventh Arabic Natural Language Processing Workshop (WANLP 2022).

5 Conclusion

In this paper, we introduced a new Arabic T5 model pre-trained on 17GB of Arabic corpora. Also, we illustrate how our ArabicT5 model shows a competitive evaluation performance against the current state-of-the-art model and other Seq2Seq T5 models. For future work, we plan to add further stages to our ArabicT5 model to improve the evaluation performance on the Gender-Rewriting task.

[4] Alhafni et al. (2022b) states that "The M2 scorer computes the precision (P), recall (R), and F0.5 by maximally matching phrase-level edits made by a system to gold-standard edits"

[5] The normalization script can be accessed through this link https://github.com/CAMeL-Lab/gender-rewriting-shared-task/blob/master/utils/normalize.py

Acknowledgements

The authors would like to acknowledge the ultimate support from Google Research Cloud TRC for providing access to Tensor Processing Unit TPU. We use resources given from TRC to pre-train and fine-tune our ArabicT5 model. The author also would like to thank Patil Suraj who release a suite of codes and examples with T5 model to the public community which make our work easier.

References

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.

Bashar Alhafni, Nizar Habash, Houda Bouamor, Ossama Obeid, Sultan Alrowili, Daliyah Alzeer, Khawlah M. Ashnqiti, Ahmed ElBakry, Muhammad ElNokrashy, Mohamed Gabr, Abderrahmane Issam, Abdelrahim Qaddoumi, K. Vijay-Shanker, and Mahmoud Zyate. 2022c. The shared task on gender rewriting. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop*, Abu Dhabi, UAE. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

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. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ibrahim Abu El-Khair. 2016. 1.5 billion words arabic corpus. *arXiv preprint arXiv:1611.04033*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. 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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding.