Neural Machine Translators (NMTs) as Efficient Forward and Backward Arabic Transliterators

Toyib Ogunremi, Anthony Soronnadi, Olamide Shogbamu, Olubayo Adekanmbi
Data Science Nigeria
toyib, anthony, shogbamu, olubayo @datasciencenigeria.ai

Abstract

This study addresses the challenge in converting Romanized Arabic text back to its original Arabic script, a capability that remains largely unsupported by existing transliteration tools. We propose that both forward and backward transliteration tasks can be effectively approached as machine translation problems. To test this hypothesis, we fine-tune three HuggingFace transformer-based Neural Machine Translation (NMT) Pretrained Language Models (PLMs) on Arabic and Romanized script datasets. Experimental results demonstrate that these models perform well, achieving approximately *99 ROUGE score and 95 BLEU score*. Our findings underscore the potential of NMT models to accurately handle transliteration, offering a valuable resource for improving Arabic language accessibility and communication.

1 Introduction

The rise of digital communication has fueled the need for effective transliteration systems capable of converting Arabic text into the Latin alphabet and, crucially, performing the reverse operation to restore the original Arabic script. This dual capability is invaluable for Arabic-speaking communities and non-Arabic-speaking foreigners alike, facilitating cross-cultural communication, enhancing learning, and improving information accessibility. Although numerous tools exist for converting Arabic script to Latin text, To the best of our knowledge very few systemsif anyare available for accurately converting Latinized Arabic back into its original script. This gap presents a unique challenge, particularly in informal online contexts where Romanized Arabic is used, as there is no standardized orthography for this form, and writing conventions vary across users.

Romanized Arabic, commonly seen in social media, SMS messaging, and informal online interactions, represents a Latin-script version of Arabic where Arabic phonetics are approximated using Latin characters and, often, numerals. The primary challenge with this format is its inconsistency; Romanized Arabic lacks a unified system of rules, as users tend to employ their own interpretations of Arabic sounds using Latin characters. Additionally, regional dialects and personal writing styles further complicate the transliteration process, resulting in Romanized text that is highly variable and prone to phonetic ambiguities. These inconsistencies make it difficult to develop an accurate, automated system capable of translating back and forth between Arabic script and Romanized text.

Given these challenges, neural machine translation (NMT) models, especially those based on transformer architectures, present a promising solution. Transformer-based models are equipped with self-attention mechanisms that enable them to capture complex language patterns(Rashno et al., 2024), contextual dependencies, and phonetic nuances critical to transliteration tasks. Furthermore, pretrained language models (PLMs) can be fine-tuned on specific datasets that include Arabic and Romanized text pairs, allowing the models to learn the intricate relationships between the two scripts. This enables the system to better handle phonetic ambiguities, and non-standardized orthographies, resulting in more accurate transliteration outcomes.

Muslims in ML Workshop @ 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2 Related Works

The process of Romanizing non-Latin scripts, including Arabic has been widely explored within NLP to address challenges related to accessibility, cross-lingual alignment, and data processing compatibility. Romanization systems initially emerged as solutions to the ASCII-only environments of the early 1990s. In Arabic-speaking regions, conventions like Franco Arabic or Arabizi were adopted to enable the use of Roman characters for Arabic text. This convention facilitated text entry on English keyboards and increased Arabic content creation, yet it posed challenges due to a lack of standardization and parallel data. (Chalabi and Gerges, 2012) solved this problem by building a system inspired by basic phrase-based statistical machine translation to transliterate romanized text to colloquial Arabic.

The debate over a standardized Romanization system persists, with calls for a system that preserves Arabic phonemic integrity while enhancing global readability. Arabic's unique phonemes, often absent in English, pose challenges for creating a universally accessible Romanization system. The academic International Phonetic Alphabet (IPA) system, though recognized by linguists, has limited practical usage due to its complexity. Inconsistent Romanization practices have hindered Arabic language preservation and the teaching of Arabic to non-native speakers(Chakhachiro, 2010).

Romanization tools such as uroman were developed by (Hermjakob et al., 2018) to facilitate the forward only process of converting a wide range of non-latin scripts including Arabic to the romanized scripts. The tool, which relies on Unicode data, offers a uniform Latin-script representation.

Recent advancements have focused on extending large language models (LLMs) to non-Romanized languages, particularly low-resourced ones. (Ogunremi et al., 2024) reported that non-Latin scripts exhibit very low ROUGE scores in headline generation tasks for African languages, even with a consistent setup across all languages. To address this, researchers have used Romanized text as an intermediary representation for LLMs, leveraging shared tokens and alignment properties with English. This method, demonstrated with models such as Llama 2 by (J et al., 2024), enables non-Roman script languages to effectively benefit from pretrained English LLMs. Results indicate that using Romanized text improves embedding alignment and significantly enhances cross-lingual transfer, yielding competitive performance across natural language understanding (NLU), generation (NLG), and machine translation (MT) tasks.

In summary, these works highlight both the practical and cultural implications of Romanization for Arabic, and other non-Latin scripts. As NLP systems continue to advance, the development of standardized and context-aware Romanization remains crucial for promoting linguistic inclusivity and facilitating cross-lingual transfer in language technologies.

3 Methodology

This section describes the whole experimentation process from data gathering, processing and model training.

3.1 Non-Latin Scripts

There are about 48 commonly used global writing system[1] of which Arabic is part with a unique script comprising 28 letters written from right to left. Unlike many languages that utilize the Latin alphabet, Arabic contains sounds that lack direct equivalents in Latin-based alphabets, which creates phonetic challenges in transliteration. Arabic also has diverse dialects and uses various phonetic and grammatical nuances, making its transliteration particularly complex. Transliteration is further complicated by Arabic's use of diacritical marks, which alter pronunciation and meaning, yet are often omitted in written texts.

The use of Romanized Arabic, where Arabic words are represented using Latin characters, is widespread in informal online communication, particularly on social media and messaging platforms. Romanized Arabic lacks a standardized orthography, leading to variations in spelling and representation that can differ between users and regions. This informal and often inconsistent ap-

[1] https://www.omniglot.com/writing/langalph.htm

proach makes the backward transliteration task i.e. converting Romanized script to Arabic script both challenging and essential for effective language processing.

3.2 Dataset

XL-Sum Arabic Corpus (Hasan et al., 2021) is the first publicly available abstractive summarization dataset containing 1 million news article-summary pairs in 44 languages scraped from BBC news covering wide range of domains or topics like sport, politics, business, health etc. We used the 46897 samples of arabic headline which serves as a one-liner summary of the full article, while also maintaining the 80%-10%-10% TRAIN-DEV-TEST split from the XL-Sum corpus.

Universal Romanization We used the uroman(Hermjakob et al., 2018) python library to generate the romanized arabic script. Uroman is a universal romanization tool that converts non-Latin scripts into a standard Latin alphabet form. This tool helps generate Romanized Arabic text from the Arabic corpus, creating parallel data pairs for training our models on both forward and backward transliteration tasks.

3.3 Pretrained Language Models (PLMs)

As the task is framed as a Translation task, we selected three transformer-based encoder-decoder pretrained language models (PLMs) available in the HuggingFace library.

mBART (Liu et al., 2020) is a multilingual seq2seq denoising auto-encoder model primarily intended for translation task, covering about 25 languages, including Arabic. We finetuned the $\sim$ 610M parameter Facebook/mBart-large-50[2](Tang et al., 2020) which covers additional 25 languages.

mT5 (Xue et al., 2020) is a multilingual variant of the Text-to-Text Transfer Transformer (T5) model that was pre-trained on a new Common Crawl-based dataset covering 101 languages. We finetuned the 580M parameter Google/mT5-base[3] model.

MarianMT (Junczys-Dowmunt et al., 2018) is an efficient and self contained Neural Machine Translation framework written entirely in C++ with minimal dependencies. We finetuned the Helsinki-NLP/opus-mt-en-ar[4] for the conversion of romanized script to Arabic text, and finetuned the Helsinki-NLP/opus-mt-ar-en[5] for the conversion of arabic text to romanized scripts.

3.4 Model Training

The models undergo supervised fine-tuning, where each model is trained to predict the correct Arabic script given a Romanized input and vice versa for forward transliteration tasks. The training setup for all the models was the same as we fine-tuned the models using a batch size of 4, number of epochs 5, and the default learning rate of $5e-5$. All models are fine-tuned on Kernel-Tesla P100 single GPU using the HuggingFace framework (Wolf et al., 2020).

4 Results

The results, as shown in Table 1, reveal notable variations in model performances across generated token length (gen_len), BLEU(Papineni et al., 2002) scores, and ROUGE(Lin, 2004) scores across the two directions (Arabic-to-Latin and Latin-to-Arabic).

More so, Table 2 also show a sample of the data and compare the predictions across the finetuned PLMs.

[2]https://huggingface.co/facebook/mbart-large-50

[3]https://huggingface.co/google/mt5-base

[4]https://huggingface.co/Helsinki-NLP/opus-mt-en-ar

[5]https://huggingface.co/Helsinki-NLP/opus-mt-ar-en

Table 1: Evaluation results

Model	Arabic - Latin			Latin - Arabic		
	gen_len	**Bleu**	**Rouge** (R1/R2/RL)	**gen_len**	**Bleu**	**Rouge** (R1/R2/RL)
Facebook/mBart-large-50	30.9	**98.7**	**99.5 / 99.0 / 99.5**	20.0	**90.5**	**95.5 / 91.4 / 95.5**
Google/mT5-base	**30.6**	96.9	98.6 / 97.3 / 98.6	22.4	85.1	92.8 / 86.4 / 92.8
Helsinki-NLP/opus-mt-ar-en	35.4	97.4	98.9 / 97.9 / 98.9	-	-	-
Helsinki-NLP/opus-mt-en-ar	-	-	-	**14.9**	89.5	95.1 / 90.5 / 95.1

Facebook/mBart-large-50 model exhibited the best overall performance across both transliteration directions possibly due to it's higher parameter size, with consistently high BLEU and ROUGE scores, making it the most reliable for converting between these scripts. While Google/mT5-base and Helsinki-NLP models also performed competitively, It is also worthy of note that the difference in the generated token length can be attributed to the Out of the box vocabularies found in the corpus when tokenizing the texts due to transliteration being a language the models has not been directly exposed to.

The findings underscore the feasibility of neural machine translation models in transliteration tasks, specifically the ability of fine-tuned PLMs to support the backward conversion task traditionally underserved by existing transliteration tools. Future work will seek to expand the transliteration to cater to full romanization which involves the representation of both the graphemes and the phonemes of Arabic and other non-latin scripts.

References

Chakhachiro, R. (2010). Standard romanisation conventions from arabic into english: A matter of language protection. In *Romanization of Arabic Names: Proceedings of the International Symposium on Arabic Transliteration Standard: Challenges and Solutions*, pages 3--18, Abu Dhabi, U.A.E. Ministry of Culture, Youth and Community Development.

Chalabi, A. and Gerges, H. (2012). Romanized Arabic transliteration. In Bali, K., Choudhury, M., and Okuno, Y., editors, *Proceedings of the Second Workshop on Advances in Text Input Methods*, pages 89--96, Mumbai, India. The COLING 2012 Organizing Committee.

Hasan, T., Bhattacharjee, A., Islam, M. S., Mubasshir, K., Li, Y.-F., Kang, Y.-B., Rahman, M. S., and Shahriyar, R. (2021). XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693--4703, Online. Association for Computational Linguistics.

Hermjakob, U., May, J., and Knight, K. (2018). Out-of-the-box universal Romanization tool uroman. In Liu, F. and Solorio, T., editors, *Proceedings of ACL 2018, System Demonstrations*, pages 13--18, Melbourne, Australia. Association for Computational Linguistics.

J, J., Dabre, R., M, A., Gala, J., Jayakumar, T., Puduppully, R., and Kunchukuttan, A. (2024). RomanSetu: Efficiently unlocking multilingual capabilities of large language models via Romanization. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15593--15615, Bangkok, Thailand. Association for Computational Linguistics.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. *CoRR*, abs/1804.00344.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74--81, Barcelona, Spain. Association for Computational Linguistics.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210.

Ogunremi, T., sessi Akojenu, S., Soronnadi, A., Adekanmbi, O., and Adelani, D. I. (2024). AfriHG: News headline generation for african languages. In *5th Workshop on African Natural Language Processing*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311--318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rashno, E., Eskandari, A., Anand, A., and Zulkernine, F. (2024). Survey: Transformer-based models in data modality conversion.

Tang, Y., Tran, C., Li, X., Chen, P., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (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.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.

A Appendix / Supplemental Material

Table 2: A Sample of the Romanized-Arabic Data for Comparison

PLMs	Arabic	Latin
Target	نضال حسن يمثل أمام محكمة عسكرية أمريكية	ndal hsn ymthl amam mhkma 'skrya amrykya
Facebook/mBart-large-50	نضال حسن يمثل أمام محكمة عسكرية أمريكية	ndal hsn ymthl amam mhkma'skrya amrykya
Google/mT5-base	نضال حسن يمثل أمام محكمة عسكرية أمريكية	ndal hsn ymthl amam mhkma'skrya amrykya
Helsinki-NLP/opus-mt-en-ar	نضال حسن يمثل أمام محكمة عسكرية أمريكية	-
Helsinki-NLP/opus-mt-ar-en	-	ndal hsn ymthl amam mhkma'skrya amrykya