Normalization of Arabic Dialects into Modern Standard Arabic using BERT and GPT-2

Khalid Alnajjar[1] **and Mika Hämäläinen**[2]

[1]Rootroo Ltd, Finland
[2]Metropolia University of Applied Sciences, Finland

Corresponding author: Khalid Alnajjar , khalid@rootroo.com

Abstract

We present an encoder-decored based model for normalization of Arabic dialects using both BERT and GPT-2 based models. Arabic is a language of many dialects that not only differ from the Modern Standard Arabic (MSA) in terms of pronunciation but also in terms of morphology, grammar and lexical choice. This diversity can be troublesome even to a native Arabic speaker let alone a computer. Several NLP tools work well for MSA and in some of the main dialects but fail to cover Arabic language as a whole. Based on our manual evaluation, our model normalizes sentences entirely correctly 46% of the time and almost correctly 26% of the time.

Keywords

Arabic dialects, dialect normalization, dialect translation

I INTRODUCTION

Arabic, with its over 20 distinct dialects[1] spoken across North Africa and the Middle East, has a need for efficient NLP methods to tackle such a dialectal diversity within the language. Currently, even native speakers may face difficulties in understanding other varieties of Arabic due to significant linguistic differences such as vocabulary usage, grammar structure, pronunciation, and accent variation (see Benmamoun 2000). This is an even bigger problem for computational methods that mainly work on Modern Standard Arabic (MSA) and a handful of larger Arabic dialects (see Shoufan and Alameri 2015). Normalizing text written in an Arabic dialect makes it possible to employ existing NLP tools on dialectal text.

However, developing effective NLP tools for Arabic dialects poses several challenges. One major challenge is the lack of standardized orthographic representation for dialectal Arabic (see Eryani et al. 2020). Unlike MSA, which has a well-defined orthography, dialectal Arabic varieties consist of primarily spoken language that lacks consistent writing conventions. This means more diversity in the orthography which further makes the task more challenging for computational approaches.

Furthermore, the high level of dialectal diversity across different regions (see Horesh and Cotter 2016) adds another layer of complexity. Each Arabic dialect has its distinctive vocabulary, idioms and syntactic patterns, which makes it difficult to build comprehensive linguistic resources and models that can handle the entire spectrum of dialectal variation. Another challenge stems

[1] See https://iso639-3.sil.org/code/ara

ID	Variation	Text
182	MSA	هل ارتفع إجمالي المبيعات في أمريكا الشمالية بما يزيد عن خمسة بالمائة ؟
	Cairo	هي المبيعات في امريكا الشمالية بتزيد بأكتر من خمسه في الميه؟
320	MSA	كلا منا يريد أن يحاسب على وجبته.
	Cairo	نحب ندفع تمن اكلنا.
	Jerusalem	كل واحد فينا بدو يدفع عن وجبتو.
	Amman	احنا كل واحد حابب يدفع على وجبته.
	Rabat	كل واحد فينا بغا يخلص على راسو.
	Doha	كل واحد بيدفع حق وجبته.

Table 1: Example sentences from the MADAR parallel corpus, along with some of their dialectal variations

from the scarcity of labeled dialectal data for training NLP models (see Shoufan and Alameri 2015). While there are a lot of resources for Modern Standard Arabic, dialectal Arabic has a comparatively limited amount of resources, except for Levantine and Egyptian Arabic. Collecting and annotating dialectal text data is time consuming and labor intensive, and it requires contribution from native speakers with expertise in each specific dialect.

II RELATED WORK

Historical and dialectal text normalization has received quite a bit of research in the past. The commonly used methods in the recent past rely on statistical and neural machine translation on a character level. Recently, there have been some normalization attempts by using pre-trained Transformer based models. In this section, we will go through some of the recent approaches.

A relatively recent study conducted by Bollmann [2019] categorizes the current approaches into five groups when it comes to text normalization methods. These groups include substitution lists, such as VARD [Rayson et al., 2005] and Norma [Bollmann, 2012], rule-based methods proposed by Baron and Rayson [2008] and Porta et al. [2013], edit distance-based approaches [Hauser and Schulz, 2007, Amoia and Martinez, 2013], statistical methods, and the most recent addition of neural methods.

In the past, statistical methods gained significant attention, particularly in the form of various statistical machine translation (SMT) based approaches. These methods typically combine the normalization process with a standard translation process by training a character-level SMT model. They have been applied to normalize historical texts [Pettersson et al., 2013, Hämäläinen et al., 2018] as well as contemporary dialect normalization [Samardzic et al., 2015].

In recent times, there has been a growing trend of utilizing neural machine translation (NMT) for normalization methods, similar to the previous approaches based on statistical machine translation (SMT) at the character level. This is primarily due to the significantly improved capability of NMT in addressing the task. Bollmann and Søgaard [2016] employed a bidirectional long short-term memory (bi-LSTM) deep neural network to normalize historical German at the character level. They also investigated the effectiveness of incorporating additional auxiliary data during the training phase, known as multi-task learning. According to their comparative evaluations, the normalizations achieved using the neural network approach outperformed those achieved using conditional random fields (CRF) and Norma. Furthermore, models trained with the auxiliary data demonstrated the highest accuracy levels.

Previously, character-level NMT approaches have been applied in a similar fashion into normalizing contemporary Finnish [Partanen et al., 2019], historical Finnish [Hämäläinen et al., 2021] and Finnish Swedish dialects [Hämäläinen et al., 2020]. Furthermore, character-level NMT approaches have been shown to work in OCR post-correction tasks as well [Duong et al., 2021].

However, as Arabic dialects can differ from MSA in grammar and vocabulary to a higher degree than the European languages the character-level models have been used for, we believe that it is better to look into more robust methods that can take more context into consideration than just character sequences. Previously, BERT has been used to normalize Estonian dialects with moderate success [Hämäläinen et al., 2022b]. In a similar fashion, Bucur et al. [2021] fine-tune mBART model to conduct dialect normalization.

The methods described above rely on supervised machine learning. In other words, such models require pairs of standard and non-standard sentences. Unsupervised methods, however, are not quite as popular as their supervised counterparts, but there is some research in that vein as well. Costa Bertaglia and Volpe Nunes [2016] have used word embeddings to normalize Brazilian Portuguese, Rangarajan Sridhar [2015] have learned both word and phrase level embeddings to normalize English and Zalmout et al. [2019] normalize English neologisms based on word embeddings and lexical, semantic and phonetic similarity.

III DATA

We use the MADAR parallel corpus of Arabic dialects [Bouamor et al., 2018]. The corpus consists of dialects of 25 Arabic city in the travel domain, and it was built by taking English samples from the Basic Traveling Expression Corpus (BTEC) [Takezawa et al., 2007] and asking native speakers to translate them into their own dialect. The authors opted for this approach to reduce bias towards Modern Standard Arabic. Table 1 shows example sentences from the MADAR corpus.

We find, however, that the choice of using English instead of Modern Standard Arabic as the translation source is not ideal for training a normalization model due to the numerous ways of expressing the same message in different dialects, which introduces diversity in expression that confuses the model. For instance, the MSA sentence هل لك أن تساعدني في الوصول إلى حقيبتي ؟ (Can you help me reach my bag?) is ممكن تساعدني ألاقي شنتة سفري؟ (Can you help me find my travel bag?) in Jordanian dialect, in the corpus.

This is problematic as the model will need to learn to both normalize and paraphrase at the same time which makes the task more difficult. There are several examples in the corpus where the MSA normalization does not express the exact same thing as its dialectal counterparts, but rather something similar.

IV ARABIC DIALECT NORMALIZATION

This section outlines our approach to normalizing dialectal Arabic sentences. Our approach utilizes an encoder-decoder architecture, with the encoder being a BERT model and the decoder being a GPT-2 model. Instead of training these models from the beginning, we employ pre-

Dialectal sentence	Model prediction	Gold standard	Is correct?
واش رايك في قهوة؟	ما رأيك في احتساء بعض القهوة ؟	ما رأيك في بعض القهوة ؟	yes
خرج إذا سمحت تبدل هالفستان بواحد قياس أكبر نمرة؟	هل يمكن أن تغير هذا الفستان عالي الساق بآخر أصغر ؟	هلا بدلت هذا الفستان بآخر من مقاس أوسع ؟	almost
بغيت نبدل لي ريزيرفاسيو ديالي.	أود أن أغير حجزي.	أريد أن أغير حجوزاتي .	yes
رجاءا انطي هذا الشراب لاولادك، مقياس ويحد، بعد كل وجبي.	من فضلك، أعطني مشروبا أخر. إن مقدار الرسوم للوجبات الخفيفة، من فضلك.	من فضلك أعط طفلك هذا الدواء الشراب ، جرعة واحدة بعد كل وجبة .	no
في شي جولة صباحية؟	هل هناك جولة بعد الظهر ؟	هل توجد أية جولة صباحية ؟	almost
هالكاري حر .	هذه الجعة عديمة النكهة.	هذا الكاري حار .	no
شكون تحب تشوف؟	من الذي تود رؤيته ؟	من الذي تريد أن تراه ؟	yes

Table 2: Examples of the model output and our manual error analysis

trained language models and adapt them for the task of dialect normalization through transfer learning. We selected a BERT-based model as the encoder due to its great capability of capturing contextual meaning. On the other hand, GPT-2 is widely recognized for its language generation abilities, making it an ideal choice as a decoder for generating text.

First we have to pick the suitable pretrained models. As we use Transformers library [Wolf et al., 2020], we select our models from their repository. The current state-of-the-art dialectal BERT model for Arabic is AraBERTv0.2 [Antoun et al., 2020] which is based on the BERT [Devlin et al., 2019]. We use the base model that is also trained on Tweets in Arabic, as conversations in social media platforms tend to be dialectal.

As for the selection of the GPT-2 model, we use AraGPT2 [Antoun et al., 2021] which is trained on the OSCAR [Abadji et al., 2022], OSIAN [Zeroual et al., 2019] and 1.5 billion words Arabic [El-Khair, 2016] corpora among others. This means that the model has a wide coverage over different genres, dialects and levels of formality.

We combine the two models into an encoder-decoder architecture similarly to the work described in Hämäläinen et al. [2022a]. To ensure proper configuration of the new model and a correct mapping between the encoder and decoder, we defined the mapping of special characters such as beginning of sentence, padding, unknown and end of sentence tokens. Furthermore, we apply pre- and post-processing steps to the new architecture similar to the ones followed in the base BERT and GPT-2 models, respectively.

We grouped the sentences in the training data by ID so that MSA sentences along with their dialectal variants are in the same group. Then, we split the groups into three parts with portions of 80%, 10%, 10% for training, validation and testing, in the mentioned order. This way, the model is validated and tested using sentences that the model was never exposed to during the training phase. We train the model for three epochs, use a batch size of 4 and validate the model with generation length penalty of 2, repetition penalty of 3 and 4 beams.

V RESULTS AND EVALUATION

For the automatic evaluation, we calculated the word error rate (WER) between the MSA variant and the generated normalization by the model. The average WER across all dialects is 83.80, which is a high error rate.

On a closer inspection of the results, we found that many of the normalization results of our model were actually correct but different to how they were normalized in the gold standard. For this reason, we sampled 100 sentences from the test data and did error analysis manually. Examples of the dialectal sentences, the output of our model and the gold standard together with our manual annotation can be seen in Table 2.

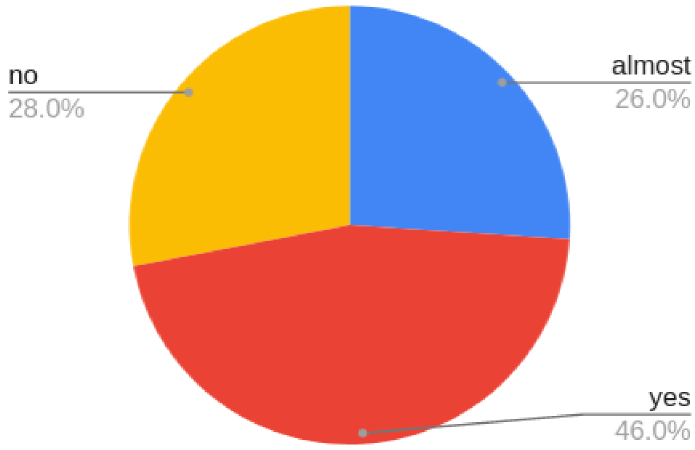

Figure 1: Results of the manual evaluation, where the labels are *yes* for a correct normalization, *almost* for normalizations with a single semantic mistake, *no* for a wrong normalization

Figure 1 shows the results of our manual analysis. Our model works better than what the automatic evaluation initially shows. Many times our model indeed normalizes sentences correctly. We found that the gold standard often does not have an exact word to word normalization but rather conveys the same idea using a different wording and structure, whereas our normalization model tends to perform a word-to-word normalization.

In terms of errors, a very common source of errors was numbers. Our model normalizes sometimes numbers incorrectly to wrong numbers. Another type of error was normalization to semantically similar, yet wrong, words such as رحلات جوية (airplane trip) instead of قطار (train), صباحية (morning) instead of يوم كامل (full day), دي جي (dj) instead of ديسكو (disco) and so on. These cases where the model makes one small error in otherwise correct sentence are marked as almost in Figure 1.

VI DISCUSSION AND CONCLUSIONS

In this paper, we have presented our approach on normalizing dialectal Arabic into Modern Standard Arabic using BERT and GPT-2. Our method resulted in low performance on the gold standard data, but after conducting a thorough manual inspection, we uncovered that our model works relatively well. The most problematic cases are numbers and semantically similar words.

After inspecting the shortcomings of the initial corpus, we can conclude that the similar word error is an artifact that is a result of the annotation practices of the corpus. The corpus has several instances of sentences that are not complete translations which makes the model also predict normalizations that are not completely identical to the content of the dialectal sentence.

We foresee that the number problem can be solved by introducing more numbers in the training data. This could be done by first training dialect specific dialectalization models that are trained to convert Modern Standard Arabic into different dialects. We could pass MSA sentences that have numbers through these dialectalization models and generate this way parallel dialect-MSA sentences that contain numbers and use these to retrain the normalization model.

VII LIMITATIONS

The model has severe issues with numbers so it should not be used in any contexts where numbers are of a great importance. Furthermore, the model works rather well, but it still makes mistakes with similar words. This means that at the current state, the model is well suited to be used as an auxiliary tool for manual normalization, where a person fixes the mistakes the model makes.

The model itself is not computationally heavy, and we trained it overnight on a desktop computer running an RTX3090 GPU. This means that the model can be trained at home without using an HPC. The model has been trained on an openly available dataset and language models, which means that our results are replicable.

ACKNOWLEDGMENTS

This research was funded by a grant from Niilo Helanderin säätiö.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv e-prints*, art. arXiv:2201.06642, January 2022.

Marilisa Amoia and Jose Manuel Martinez. Using comparable collections of historical texts for building a diachronic dictionary for spelling normalization. In *Proceedings of the 7th workshop on language technology for cultural heritage, social sciences, and humanities*, pages 84–89, 2013.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9, 2020.

Wissam Antoun, Fady Baly, and Hazem Hajj. AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, Kyiv, Ukraine (Virtual), April 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.wanlp-1.21.

Alistair Baron and Paul Rayson. VARD2: A tool for dealing with spelling variation in historical corpora. In *Postgraduate conference in corpus linguistics*, 2008.

Elabbas Benmamoun. *The feature structure of functional categories: A comparative study of Arabic dialects*. Oxford University Press, 2000.

Marcel Bollmann. (Semi-)automatic normalization of historical texts using distance measures and the Norma tool. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, Lisbon, Portugal, 2012. URL https://marcel.bollmann.me/pub/acrh12.pdf.

Marcel Bollmann. A large-scale comparison of historical text normalization systems. 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 3885–3898, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1389. URL https://www.aclweb.org/anthology/N19-1389.

Marcel Bollmann and Anders Søgaard. Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 131–139, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1013.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. The madar arabic dialect corpus and lexicon. In *LREC*, 2018.

Ana-Maria Bucur, Adrian Cosma, and Liviu P Dinu. Sequence-to-sequence lexical normalization with multilingual transformers. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 473–482, 2021.

Thales Felipe Costa Bertaglia and Maria das Graças Volpe Nunes. Exploring word embeddings for unsupervised textual user-generated content normalization. In *Proceedings of the 2nd Workshop on Noisy User-generated*

Text (WNUT), pages 112–120, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/W16-3916.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019.

Quan Duong, Mika Hämäläinen, and Simon Hengchen. An unsupervised method for ocr post-correction and spelling normalisation for finnish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 240–248, 2021.

Ibrahim Abu El-Khair. 1.5 billion words arabic corpus. *ArXiv*, abs/1611.04033, 2016.

Fadhl Eryani, Nizar Habash, Houda Bouamor, and Salam Khalifa. A spelling correction corpus for multiple Arabic dialects. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4130–4138, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.508.

Mika Hämäläinen, Tanja Säily, Jack Rueter, Jörg Tiedemann, and Eetu Mäkelä. Normalizing early English letters to present-day English spelling. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 87–96, 2018.

Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar. Normalization of different swedish dialects spoken in finland. In *Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities*, pages 24–27, 2020.

Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar. Lemmatization of historical old literary finnish texts in modern orthography. In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1: conférence principale*, pages 189–198, 2021.

Mika Hämäläinen, Khalid Alnajjar, and Thierry Poibeau. Modern french poetry generation with roberta and gpt-2. In *13th International Conference on Computational Creativity (ICCC) 2022*, 2022a.

Mika Hämäläinen, Khalid Alnajjar, and Tuuli Tuisk. Help from the neighbors: Estonian dialect normalization using a finnish dialect generator. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 61–66, 2022b.

Andreas W Hauser and Klaus U Schulz. Unsupervised learning of edit distance weights for retrieving historical spelling variations. In *Proceedings of the First Workshop on Finite-State Techniques and Approximate Search*, pages 1–6, 2007.

Uri Horesh and William M Cotter. Current research on linguistic variation in the arabic-speaking world. *Language and Linguistics Compass*, 10(8):370–381, 2016.

Niko Partanen, Mika Hämäläinen, and Khalid Alnajjar. Dialect text normalization to normative standard finnish. In *The Fifth Workshop on Noisy User-generated Text (W-NUT 2019)*. The Association for Computational Linguistics, 2019.

Eva Pettersson, Beáta Megyesi, and Jörg Tiedemann. An SMT approach to automatic annotation of historical text. In *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18*, number 087, pages 54–69. Linköping University Electronic Press, 2013.

Jordi Porta, José-Luis Sancho, and Javier Gómez. Edit transducers for spelling variation in Old Spanish. In *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18*, number 087, pages 70–79. Linköping University Electronic Press, 2013.

Vivek Kumar Rangarajan Sridhar. Unsupervised text normalization using distributed representations of words and phrases. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 8–16, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-1502. URL https://aclanthology.org/W15-1502.

Paul Rayson, Dawn Archer, and Nicholas Smith. VARD versus WORD: a comparison of the UCREL variant detector and modern spellcheckers on english historical corpora. *Corpus Linguistics 2005*, 2005.

Tanja Samardzic, Yves Scherrer, and Elvira Glaser. Normalising orthographic and dialectal variants for the automatic processing of Swiss German. In *Proceedings of the 7th Language and Technology Conference*, 2015. ID: unige:82397.

Abdulhadi Shoufan and Sumaya Alameri. Natural language processing for dialectical Arabic: A survey. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 36–48, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3205. URL https://aclanthology.org/W15-3205.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics &*

Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006, pages 303–324, September 2007. URL https://aclanthology.org/O07-5005.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 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, 2020.

Nasser Zalmout, Kapil Thadani, and Aasish Pappu. Unsupervised neologism normalization using embedding space mapping. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 425–430, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5555. URL https://aclanthology.org/D19-5555.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. OSIAN: Open source international Arabic news corpus - preparation and integration into the CLARIN-infrastructure. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4619. URL https://aclanthology.org/W19-4619.