TF-IDF or Transformers for Arabic Dialect Identification? ITFLOWS participation in the NADI 2022 Shared Task

Fouad Shammary[1], **Yiyi Chen**[2,3], **Zsolt T. Kardkovács**[1], **Haithem Afli**[1], **Mehwish Alam**[2,3]
[1] Department of Computer Science, Munster Technological University, Cork, Ireland
[2] FIZ-Karlsruhe - Leibniz Institute for Information Infrastructure, Karlsruhe, Germany
[3] Karlsruhe Institute of Technology, Karlsruhe, Germany
{firstname.lastname}@mtu.ie, {firstname.lastname}@fiz-karlsruhe.de

Abstract

This study targets the shared task of Nuanced Arabic Dialect Identification (NADI) organized with the Workshop on Arabic Natural Language Processing (WANLP). It further focuses on Subtask 1: the identification of the Arabic dialects at the country level. More specifically, it studies the impact of a traditional approach such as TF-IDF and then moves on to study the impact of advanced deep learning based methods. These methods include fully fine-tuning MARBERT as well as adapter based fine-tuning of MARBERT with and without performing data augmentation. The evaluation shows that the traditional approach based on TF-IDF scores the best in terms of accuracy on TEST-A dataset, while, the fine-tuned MARBERT with adapter on augmented data scores the second on Macro F1-score on the TEST-B dataset. This led to the proposed system being ranked second on the shared task on average.

1 Introduction

Arabic is a Semitic language spoken in more than 26 countries by more than 350 million people with at least 30 dialects[1]. Some previous studies attempted to use hierarchical deep learning for a fine-grained dialect classification (de Francony et al., 2019). Arabic has its own, letter based writing system which is used mostly for only those consonants which could denote a wide range of pronunciation alternatives. A single letter in this alphabet can have various forms depending on the context, and its position within the word which are encoded by different characters. There are also single character ligatures which are formed by two or more characters (e.g., from this corpus: U+FEFB ("AL") denotes U+0627 (A) and (U+0644 (L), or words like U+FD71 (“aspired") or U+FDF2 (“Allah") are also represented by a single character). Similarly to the Latin alphabet, Arabic letters can denote, e.g., Urdu, Ottoman Turkish, Sindhi, Malay, Uyghur, or even English and French words which are not uncommon.

This article targets the Nuanced Arabic Dialect Identification (NADI) 2022 Shared Task (Abdul-Mageed et al., 2022). It more specifically focuses on Subtask 1 aimed at identifying country-level dialects by providing ∼20k Twitter data which are labeled by geo-location (i.e. country) from where the tweets were posted. In this Shared Task no external *labeled* data sources were allowed to be used, however, a large unlabeled dataset was also provided. The training set remains relatively small to encourage competitors to use few or zero-shot learning models. Solutions were tested on two datasets using macro-averaged F1-score:

- TEST-A: ∼5k tweets with all previously provided dialects,
- TEST-B: ∼1.5k tweets with an undisclosed number of country-level dialects.

According to the systems developed and presented in this work, dialect identification can be modeled at the character, word, expression, or phrase level. Each of these levels was modeled by the traditional TF-IDF method, a pre-trained transformer called MARBERT (Abdul-Mageed et al., 2021), and by using MARBERT with word level augmentation. These methods were analyzed individually as well as by using their combination in order to find the most informative parameter regarding the dialects. On TEST-A dataset the traditional approach produced the best accuracy, while on TEST-B dataset our approach won the runner-up award on macro-averaged F1-score.

2 Data

The NADI 2022 Shared Task Subtask 1 dataset contained a total of 20,398 tweets in the training set,

[1] ISO 639-3 identified dialects: https://iso639-3.sil.org/code/ara

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 420 - 424
December 8, 2022 ©2022 Association for Computational Linguistics

4,872 validation samples from 18 dialects, while unlabeled test sets, TEST-A and TEST-B contained 4,758 and 1,473 tweets respectively. Dialects are identified based on geolocation data where the tweets were originated instead of the linguistic analysis. This in itself leads to the contamination of the data since people might reside in a country other than their country of origin. Moreover, sometimes the words are used in or borrowed from other languages, e.g., English or Urdu. Additionally, there is an imbalance in the class distribution in the training dataset (see Table 1).

3 System

Three models were proposed for this subtask out of which the first model was a *traditional approach* without using any language models or deep neural network architecture, i.e., TF-IDF based. In the second approach, the data augmentation was performed with fine-tuned MARBERT (Abdul-Mageed et al., 2021) with and without adapters (Pfeiffer et al., 2020a).

3.1 System 1: Traditional Approach

In order to capture relevant differences between dialects, one can look for particular linguistic alterations of similar characters, words, phrases, or expressions. TF-IDF method has a long history in detecting such meaningful differences in texts, especially for detecting topics in large texts. This study considers all the texts with the same label as a single document. This way, dialects can be identified as common sub-word patterns (in our case 1 to 7-grams) which are frequent enough (i.e. $f(w) > t_f$) within a document (dialect), but they are not universal, i.e. at least $k > 0$ documents shall not contain this pattern at all. Since the dataset had the same topic for all dialects TF-IDF method most likely identifies dialects rather than topics. These (t_f, k)-patterns could be used as fingerprints for dialects. The most likely fingerprint using maximum likelihood determines the outcome of the prediction. The best accuracy was achieved using $(3, 9)$-patterns as fingerprints.

However, using N-grams could lead to a wide variety of errors. The appearance of words and encoding could be misleading using Arabic enabled, modern operating systems. For example, ﺣاجه and حاجه appear to be the same, however, their underlying Unicode characters are completely different (e.g. the first letter is `U+FEA3` with respect to `U+062D`). To avoid such a problem, one can introduce a transliteration module that maps these differences into a common alphabet. While it sometimes helps differences between words like السلامه and السلامہٕ which are hardly noticeable in transliteration (both translate to "AlslAmh", the latter is Urdu and means "peace be upon you", the former is Arabic (means "safety"). Both appear in the NADI 2022 corpus. In the current study, it was noticed that the transliteration based approach tends to over-perform traditional character-based approaches when using (t_f, k)-fingerprints.

Since the training dataset was small and unbalanced, this approach favors more sampled dialects over small ones. A randomly sampled balanced set worsened the overall accuracy because of the small training samples.

3.2 Data Augmentation based Approach

Data augmentation is a technique where the amount of data is increased by adding slightly modified copies of the existing data. Several kinds of data augmentation techniques are generally used in NLP such as word level, and sentence level. This paper uses the word insertion technique from (Wei and Zou, 2019) combined with Transformers by inserting a word randomly based on context. This technique is performed on all tweets from the countries that represent less than 10% of the data. Each tweet is augmented by inserting one or two words randomly based on the contextualized embeddings from MARBERT. The entire dataset, which is comprised of both the newly augmented tweets dataset and the original tweets dataset, is checked for any duplicates which are then removed. For instance, there were 642 tweets from the UAE (labeled as "uae"), which increased to 1284 after augmentation and removing duplicates Table 1.

3.3 System 2: Fine-tuning MARBERT

The data was tokenized in the preprocessing step no other preprocessing was used. In this system, MARBERT embeddings were fed into the max pooling layer, and then dense layers. MARBERT was fine-tuned for 5 epochs. Early stopping was employed when there was no improvement in the validation metric (balanced accuracy).

3.4 System 3: Adapter-based Approach

In order to leverage the multilinguality and improve the transferability of MARBERT, while at the same

Table 1: Distribution of dialects within the NADI 2022 Shared Task Subtask 1 challenge training dataset before and after applying augmentation techniques.

Label	Nr. Samples (%)		Label	Nr. Samples (%)	
	Original	Augmented		Original	Augmented
egypt	4,283 (20.99%)	4,283 (13.55%)	libya	1,286 (6.31%)	2,571 (8.13%)
kuwait	429 (2.10%)	857 (2.71%)	iraq	2,729 (13.38%)	2,719 (8.60%)
tunisia	859 (4.21%)	1,715 (5.43%)	yemen	429 (2.10%)	858 (2.71%)
ksa	2,140 (10.49%)	2,139 (6.77%)	morocco	858 (4.21%)	1,715 (5.43%)
palestine	428 (2.10%)	855 (2.70%)	algeria	1,809 (8.86%)	3,606 (11.41%)
lebanon	644 (3.16%)	1,287 (4.07%)	bahrain	214 (1.05%)	430 (1.36%)
oman	1,501 (7.35%)	3,002 (9.50%)	uae	642 (3.15%)	1,284 (4.06%)
qatar	215 (1.05%)	430 (1.36%)	syria	1,287 (6.31%)	2,573 (8.14%)
jordan	429 (2.10%)	858 (2.71%)	sudan	215 (1.05%)	430 (1.36%)

time, being more computationally efficient, the fine-tuning strategy Adapter (Houlsby et al., 2019) is used. Transformer layers are connected using skip-connections with adapter layers, which are composed of a down-projection and an up-projection. For fine-tuning the model, only the parameters of the adapter layers are trained, while the pre-trained transformer layers are frozen. In (Pfeiffer et al., 2020b), the authors propose an adapter-based framework for multi-task cross-lingual transfer (MAD-X), in which the language adapters and task adapters are trained separately. Task adapters can be trained with datasets for specific tasks, while language adapters are task-agnostic. For country-level dialect detection, the augmented dataset was used to train task adapters based on MARABERT, using the configuration of PfeifferConfig[2] by leaving out the adapter in the last transformer layer (MAD-X 2.0), which proves to be superior than original MAD-X in zero-shot transfer (Pfeiffer et al., 2021). The hyper-parameters used for training are learning rate $1e-4$, batch size 16, and training epoch 6. The fine-tuned model performs the best at step 4500, which is used for testing.

4 Results

As shown in Table 2, the model that performs the best on the DEV dataset in every metric is MARBERT fine-tuned on the augmented dataset using adapters, i.e. Fine-tuned-Adapter-MARBERT (AUG). Surprisingly, the regarding model performs the worst on the TEST-A dataset. In comparison, the TF-IDF approach scores the best in all metrics other than the Macro-F1 score. Since MARBERT-based models are pre-trained on a much larger corpus, and fined-tuned for this specific task, one would expect the contrary. On the DEV dataset, it can be clearly seen that TF-IDF cannot model properly small sampled dialects which leads to poor macro-F1 performance. That is, the TF-IDF based solution can capture enough information for some dialects for which transformers can't. The only reasonable explanation is that information on dialects is most likely encoded at a sub-word level which MARBERT by design could not see.

On the TEST-B dataset, where the number of country-level dialects is unknown, Fine-tuned-Adapter-MARBERT (AUG) performs the best in every metric. However, the performance difference among transformer-based approaches with or without augmented data tested on either TEST-A or TEST-B dataset is not as noticeable as the difference between the traditional approach and the transformer-based approaches tested on TEST-B. This indicates the superiority of zero-shot transfer of the pre-trained transformer.

5 Discussion

Results show noticeably high variance in precision and overall accuracy between development and test data sets, regardless of which submitted model one cross-references. Under-sampling could explain that because in small samples words can either be interpreted as dialectal use of another, more common concept or simply another topic, stance, or key communication element which focuses the attention. In both cases, the word embedding, and TF-IDF could see clear alternatives for the same concept which is the basis of the classification.

Moreover, the traditional approach suffers sig-

[2] https://tinyurl.com/c6vwrmyt

Table 2: Results on DEV and TEST datasets. Aug indicates that the model trained on the augmented training dataset. Digits in bold indicate the best results for the corresponding dataset.

Dataset	**DEV**		**TEST-A**		**TEST-B**	
Models	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy
Traditional TF-IDF	0.1275	0.3437	0.0466	0.1642	0.0555	0.1906
Full Fine-tuned MARBERT	0.3329	0.5272	**0.1862**	**0.3218**	0.1668	0.3338
Fine-tuned MARBERT (Aug)	0.3192	0.5066	0.0495	0.1127	0.1702	0.3372
Fine-tuned-Adapter-MARBERT (Aug)	**0.3462**	**0.5293**	0.0485	0.1152	**0.1767**	**0.3392**

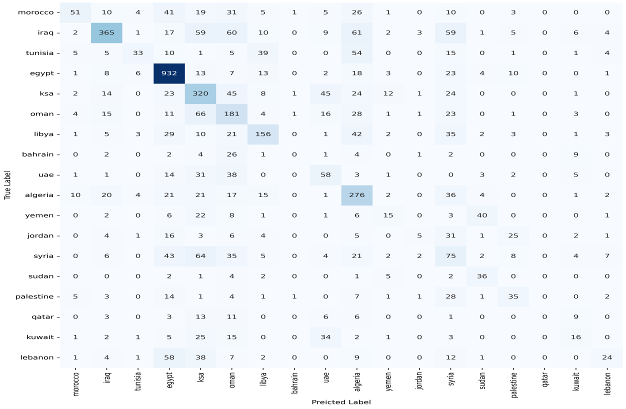

Figure 1: Confusion matrix of Fine-tuned-Adapter-MARBERT on the DEV set

nificantly less in accuracy in comparison with transformers-based models, and its performance is relatively consistent between the two differently sampled test sets. While TF-IDF models have no background knowledge of the language, there is no pretraining available, it still can outperform transformers in terms of accuracy, especially for dialects with large samples. In that sense, the TF-IDF approach is more stable, and therefore its power for generalization is stronger which means it can grab some important features of dialectal Arabic. There is a strong indication to improve or to create a subword based, or a transliteration and sub-word based transformer for Arabic.

Further analysis of predictions made be the best performing model show an expected over prediction of dialects with higher presence within the training data Figure 1. The over prediction showed a tendency towards dialects that are more similar. For instance, UAE was predicted more as Oman or KSA rather than Egypt. On the other hand, countries with small presence such as Qatar and Bahrain had no correct predictions on the DEV set.

6 Conclusion

This paper targets the problem of Arabic dialect detection based on a traditional approach as well as the pre-trained transformers in a dataset where few-shot learning was encouraged, and no large training set was provided. While the TF-IDF based approach performs less than the pre-trained transformer based approach on the NADI 2022 corpus which was expected, the accuracy of the TF-IDF approach surprisingly remained competitive on the whole (TEST-A) test set. TF-IDF obviously underperforms as compared to the MARBERT-based approach for low sampled dialects due to a lack of enough data for stable fingerprinting which explains TEST-B results. Since TF-IDF and MARBERT target different levels of the written language, so the most reasonable explanation is that dialect is more likely determined at the sub-word level. This hypothesis, however, needs further investigation.

Acknowledgements

This work is a part of the ITFLOWS project funded by the EU's Horizon 2020 research and innovation program under Grant Agreement No 882986 and the Science Foundation Ireland research Centre ADAPT through Grant 13/RC/2106_P2.

References

Muhammad Abdul-Mageed, AbdelRahim A. Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: deep bidirectional transformers for arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7088–7105. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The Third Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of the Seven Arabic Natural Language Processing Workshop (WANLP 2022)*.

Gael de Francony, Victor Guichard, Praveen Joshi, Haithem Afli, and Abdessalam Bouchekif. 2019. Hierarchical deep learning for Arabic dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 249–253, Florence, Italy. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulic, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 46–54. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing"*.