Named Entity Recognition for Algerian Arabic Dialect in Social Media

Abdelhalim Hafedh Dahou[1(✉)] and Mohamed Amine Cheragui[2]

[1] Lorraine University, 54000 Nancy, France
abdelhalim-hafedh.dahou2@etu.univ-lorraine.fr
[2] Mathematics and Computer Science Department, Ahmed Draia University, 01000 Adrar, Algeria
m_cheragui@univ-adrar.edu.dz

Abstract. Named entities recognition (NER) has become, over time, a potentially helpful pre-processing for several NLP tasks. It mainly identifies and classifies entities in the text into predefined categories, such as a person, location, and organization. Most of the work done in Arabic named entity recognition (ANER) has focused on Modern Standard Arabic (MSA). However, most scripts on the internet, especially in social media (which are currently a source for corpus development), are in a dialectal form that does not follow standard writing rules. This paper investigates the possibility of deep learning based on named entity recognition in Algerian dialect script through a comparative study between 5 models: AraBERT, DziriBERT, MARBERT, ARBERT, and mBERT. We chose these five models for two significant reasons; the first one, that these models are already pre-trained on the MSA and Arabic dialect text, the second reason, that they have proved their efficiency for other tasks such as Part of Speech Tagging, sentiment analysis, and ANER.

Keywords: NLP · Deep learning · Algerian dialect · Named Entity Recognition · Social media

1 Introduction

Over the last three decades, the expansion of the internet has engendered an exponential explosion in terms of the quantity of information stored, making finding relevant information a particular task. Several applications of natural language processing (NLP) are interested in developing methods and tools to deal with this challenge, such as information extraction [1], information retrieval [3], question answering [2], and machine translation [4]. In these different topics, the task of named entities recognition plays a transversely role.

The concept of named entity (NE) was introduced in the mid-1990s[1] [5], as a subtask of information extraction activity. It consists of identifying certain textual entities such as names of persons, organizations, and locations.

[1] In 1996 at the Message Understanding Conference (MUC).

M. R. Laouar et al. (Eds.): ICISAT 2022, LNNS 624, pp. 135–145, 2023.
https://doi.org/10.1007/978-3-031-25344-7_13

The NER is now facing new difficulties inherent to the characteristics of the modality or the type of text to be treated. It must be said that the large amount of information circulating on the web, mainly that found in social media such as Twitter and Facebook, is dialectal, especially for the Arabic language, which has a variety of dialects derived from Modern Standard Arabic (MSA).

In this context, the work we present aims to detect and extract, in Algerian dialect texts, the relevant named entities based on a deep learning approach, through a comparative study between several models (AraBERT, DziriBERT, MARBERT, ARBERT, and mBERT), but also the elaboration of a corpus dedicated to Algerian Named Entity Recognition (ALNER) which will be very useful for the training and evaluation phases.

The structure of the paper is as follows. Section 2 is devoted to some specificities of Algerian dialect. A literature review of ANER is given in Sect. 3. Section 4 describes the method for our comparative study. Section 5 presents in brief the models used in this study. The experimental process are discussed in Sect. 6. Results and discussion are respectively mentioned in Sect. 7 and 8. Section 9 concludes this paper and gives directions for future work.

2 Algerian Dialect: Overview and Specificities

The Algerian dialect (also called darijaa [8] or daArjah [6]) is a living language used daily by 70–80% of the Algerian population in all societal behaviors (press, television, social communication, internet exchanges, SMS, etc.) and family dialogues and others. The Algerian dialect is mainly based on modern standard arabic (65%) and also other languages such as French (19%), Turkish, Spanish, and Berber (16%) [7].

With an estimated land area of 2.382 million km^2, it is clear that even the Algerian dialect can have variants. According to [6, 8, 9], we can identify four sub-variants, which are:

- Algiers dialect: which covers the whole central area of the country (Algiers and neighboring towns);
- Oran dialect: It is located between the Algerian-Moroccan border and the town of Tenes (Chlef);
- Rural dialect: spoken in the east of the country (from Constantine to the Algerian-Tunisian border);
- Saharan dialect: Dialect used by the population living in the south of Algeria.

According to the studies done on Algerian dialect by [6, 10], it is clear that the differences between Algerian dialect and standard Arabic are multiple and occur on different levels of processing (phonological, morphological, orthographic, and lexical); we will limit ourselves in this section to listing those, which are the most obvious:

- Since the Algerian dialect is a descendant of MSA, it inherits its challenges, such as agglutination, lack of diacritics, and syntactic flexibility.

- The Algerian dialect lacks codification and standardization. It is written either in Latin letters (Arabizi) or Arabic letters; a word can have several orthographic representations.
  Example: the word walnut can have an orthographic representation "جوز" in the Algiers dialect and "قرقاع" in the Oran dialect.
- The Algerian dialect never uses the dual, only the singular and plural.
- An orthographic variation is since the dialect is based more on phonetics. Example: the word "chkoun?" which means "who?" can be written: chkouun? or chkoune?
- In the Algerian dialect we can find words coming from other languages, but written in Arabic letters. Example: The word "بوجي" (Bouji) means "move" in French.

3 Related Work

Since 1998 [11, 12], several works have been done on ANER, most of them focusing on Modern Standard Arabic. These works can be grouped into three categories depending on the approach adopted: rule-based approach (Requires the intervention of a linguist to set up rules that will be converted into regular expressions) [13, 14], machine learning approach (Example of techniques used: Conditional Random Fields, Support Vector Machines, Maximum Entropy, Decision Trees and HMM) [4, 15–18] or a hybrid approach (combining between rule-based and machine learning to improve the NER task) [19–21]. However, for the Arabic dialect, works are scarce; there are only three articles, the first one on the Egyptian dialect by [22], which is based on a machine learning approach using the CRF technique. The system developed detects person (F1 Score = 49.18%) and location (F1 Score = 91.429%) entities. The second one, in the Tunisian dialect, was developed by [23]. The system adopts a rule-based approach accompanied by the translation from Tunisian into standard Arabic. The system achieves an overall score in recognizing person, location, and organization entities equal to 86.81%. The last one is [34], a corpus of Algerian Arabizi–French code-switching acquired from Facebook pages, annotated by three native speakers for several tasks including NER (persons and places) and tested just in two tasks: Algerian city identification and Algerian region dialect identification. To the best of our knowledge, there is no work on the Algerian dialect for the NER task.

It is important to note that if work in the Arabic dialect is still behind, it is mainly due to the lack of linguistic resources (specialized corpora in the Arabic dialect NER).

4 Methodology

In this research, we use models that have already been trained on MSA and Arabic Dialect texts and are based on the BERT architecture. Deep Learning models have been shown in the literature to perform well on the NER task.

A comparative study of models was conducted on an Algerian dialect dataset (DzNER)[2], which two native Arabic speaker annotators created to satisfy this study's objective.

5 Arabic Pre-trained Models

In order to employ the weights and architecture obtained and apply the learning to the problem statement, the majority of state-of-the-art models are now built on pre-trained models that have previously been trained on huge datasets. The architecture, training set, and vocabulary size of the Arabic pre-trained models will be briefly discussed in this section.

5.1 AraBERT

AraBERT [24] is a pre-trained Arabic language model based on the BERT architecture, till now appears in different versions such as AraBERT v0.2, AraBERT v2, and arabert v02-twitter in order to replace the old version v1. In this study, we will use two recent versions from AraBERT, which are AraBERT v0.2 and AraBERT v02-twitter.

AraBERTv 0.2 was trained on 77 GB of data that contains 200 million sentences constructed from Arabic news websites and two publicly available large Arabic corpora [25] and [26].

AraBERT v02-twitter was trained on the same datasets as AraBERT v0.2, in addition to 60M Multi-Dialect Tweets filtered from a collection of 100M with the existence of emojis on their vocabulary.

5.2 MARBERT

MARBERT [27] is an Arabic pre-trained language model that uses the same network architecture as BERT-base and was trained on Dialectal Arabic (DA). The training was done on 1 Billion Arabic short tweets taken from a large in-house dataset of about 6 Billion tweets. In total, the training dataset makes up 128 GB of text (15.6 Billion tokens) and a vocabulary of size 110K.

5.3 ARBERT

ARBERT [27] was published with MARBERT and is also a large-scale pre-trained masked language model focused on MSA and Egyptian dialect. It have the same architecture as BERT-base and pass through the same train configuration as MARBERT, in addition to the next sentence prediction (NSP). The model was trained on a collection of Arabic datasets comprising 61 GB of text (6.2B tokens) and had a vocabulary of 100K word pieces.

[2] https://github.com/Dahouabdelhalim/DzNER-Corpus.

5.4 DziriBERT

DziriBERT [28] is the first Transformer-based Language Model pre-trained dedicated just for the Algerian Dialect. It can handle both Arabic and Latin characters in Algerian text content. Even though it was pre-trained on significantly less data (one million tweets), it achieves new state-of-the-art performance on Algerian text classification datasets.

5.5 mBERT

mBERT [29] is a model that has been pre-trained on the top 104 languages, including Arabic, and has a shared vocabulary of 110k tokens. This model was trained with the largest Wikipedia using a masked language modeling (MLM) objective to predict the next words and sentences.

6 Experiments

This section provides details about the dataset used in the fine-tuning and evaluation of the models, the configuration of each model, and the environment that holds the experiments.

6.1 Dataset

We use an Algerian dialect dataset annotated by two Arabic native speaker annotators who follow the same tagging guidelines. The data was gathered from two sources:

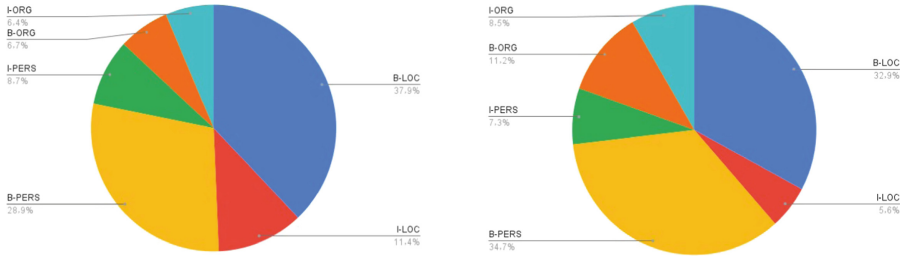
- Algerian dialect corpus [30]: a corpus built on top of the NArabizi treebank by [31]. The Algerian dialect corpus, which consists of 731 phrases and 11763 tokens, was collected from an Algerian newspaper's web forum and song lyrics.
- Algerian Facebook pages: the data was selected from a collection of Algerian Facebook pages with 1462 phrases totaling 15868 tokens that were automatically retrieved from post comments.

Both corpora are concatenated together, and the final version, DzNER, was annotated using the IOB2 format [32].

Table 1 and Fig. 1 display dataset information, including the number of tokens and named entity types PER, LOC, and ORG.

Table 1. Evaluation data statistics.

	Tokens	Person	Organization	Location
Train	22320	699	327	642
Test	3504	112	39	147
Total	25824	811	366	789



(a) Test data (b) Train data

Fig. 1. Entities distribution in train and test data.

In the first stage, a pre-processing phase was applied to clean the dataset by removing empty spaces, noisy punctuation such as "_", and some typo words that have no meaning like "يااااه" ,"الهاهاهاها" and "أووووووووو". After cleaning, we construct three versions from the cleaned dataset, which are:

- Segmented dataset: we applied a pre-segmentation process to evaluate the models in this text format. An example of this, "الجزاير" with a label B-LOC becomes "جزاي ال","ر" with labels B-LOC, I-LOC respectively.
- Unsegmented dataset: this dataset is the inverse of the previous version, which means that the complete word is kept without the prefix being separated.
- Latin chars dataset: we applied code-switching to the cleaned version, which led to converting all the words existing in the dataset from Arabic chars to Latin chars. The code-switching process was done by taking into account the correspondence presented in [33].

6.2 Experimentation Setup

We used the same API as described in the AraBERT documentation, and all of the experiments were done on the Google Colab platform with a GPU Tesla P100-PCIE16GB. Using the validation dataset, we fine-tuned the hyper-parameter to identify the ideal configuration for each pre-trained model to obtain its best performance.

Table 2 lists the hyper-parameters for each model and the rest of parameters are exist in our repository[3].

Table 2. Hyper-parameters values for each used model.

Models	Parameters	Value
AraBERT	Epochs	20
	learning_rate	5e−5
	warmup_steps	42
DziriBERT	Epochs	15
	learning_rate	5e−5
	warmup_steps	0
MARBERT	Epochs	15
	learning_rate	3e-5
	warmup_steps	42
ARBERT	Epochs	30
	learning_rate	3e−5
	warmup_steps	42
M-BERT	Epochs	15
	learning_rate	5e−5
	warmup_steps	42

7 Results

Table 3, 4 and 5 illustrates the experimental results of applying the existing Arabic pre-trained models on the three existing versions of the DzNER dataset and comparing them in terms of precision, recall, and F1 score measures.

Table 3. Results of the models on the segmented dataset.

Metrics	AraBERT v0.2	AraBERT v0.2-T	mBERT	ARBERT	MARBERT	DziriBERT
P	0.830	0.804	0.734	**0.860**	0.780	0.780
R	0.750	0.768	0.627	**0.781**	0.759	0.700
F1	0.790	0.786	0.676	**0.819**	0.771	0.740

Table 3 presents the details of the model results in the segmented DzNER corpus.

The ARBERT model fared better than the other models in terms of precision, recall, and F1 score, as evidenced. The DziriBERT model, which is the most specialized for this kind of text, was outperformed by both AraBERT variations and MARBERT. We noticed that the results between the AraBERT versions are 0.004 lot closer. Precision, Recall, and F1 Score data show that mBERT is the least effective model when compared to the others.

[3] The other parameter are available in: https://github.com/Dahouabdelhalim/NER-model-onthe-DzNER-corpus.

Table 4. Results of the models on the unsegmented dataset.

Metrics	AraBERT v0.2	AraBERT v0.2-T	mBERT	ARBERT	MARBERT	DziriBERT
P	0.875	0.846	0.776	**0.888**	0.799	0.831
R	0.803	0.803	0.698	**0.803**	0.780	0.767
F1	0.838	0.824	0.735	**0.844**	0.789	0.798

Table 4 presents the models' results in the unsegmented DzNER corpus. Although the two AraBERT versions had similar recall rates, it is clear that the ARBERT model outperformed the other models in terms of F1 score. DziriBERT outperformed MARBERT and mBERT in terms of precision and F1-score. mBERT is still the last model in terms of outcomes in comparison to the others.

Table 5. Results of the models on the Latin chars dataset.

Metrics	AraBERT v0.2	AraBERT v0.2-T	mBERT	ARBERT	MARBERT	DziriBERT
P	0.692	0.663	**0.730**	0.720	0.670	0.680
R	0.616	0.566	**0.630**	0.529	0.520	0.593
F1	0.652	0.610	**0.676**	0.610	0.586	0.634

Table 5 resume the model's results in the Latin chars DzNER corpus. In this experiment, the mBERT fared better than the other models in terms of F1 score, recall, and precision. We can see that there is a 0,042 difference in the results between the AraBERT versions. In terms of the experiment's outcomes, the model MARBERT comes in last.

8 Discussion

As illustrated in the results section, ARBERT achieved the highest performance on segmented and unsegmented DzNER dataset versions, and the mBERT achieved the highest performance just in the Latin chars DzNER dataset version. The ARBERT model shows that training on different sources of data and the vocabulary size is a clear indicator for the boost in performance and also the intersection between the MSA and Algeria dialect such as: "**الملعب**" means stadium, "**كارثة**" means catastrophe, "**الناس**" means people, etc. mBERT, on the other hand, outperformed all of the models in the Latin chars DzNER dataset due to its shared vocabulary and the fact that the Algerian dialect has a large number of French terms such as: "**Cv**" which is the abbreviation of "**ca va**" and means how are you, also there is "**tranquille**" means calm, especially in the sport category we find the frequent French word match which means game. Those French words making it difficult for models trained solely on the Arabic language to represent them.

The performance improvement occurred in the second experiment with the unsegmented DzNER corpus, confirming the conclusion [24] that pre-segmentation

reduces a model's performance in the NER task, as seen in Table 3 against Table 4. The primary reasons why the DziriBERT model can't compete with other models are because it's trained on fewer data and has a tiny vocabulary. Furthermore, after comparing the writing styles of Algerian users on Facebook and Twitter, we discovered that the content on Twitter is significantly more clear and understandable than the content on Facebook, which may provide some details regarding the models' challenges in this dataset.

9 Conclusion

We presented our efforts to compare the existing Arabic pretrained language models on the Algerian dialect for the task of NER. We chose those models because they are trained on large MSA and dialectal Arabic datasets covering different domains and including social media text. The ARBERT models per form better than DziriBERT, MARBERT, AraBERT, and mBERT in the Arabic chars, and the mBERT outperformed all models in Latin chars. DzNER, on the other hand, is the first dataset for the Algerian dialect NER task that comes in three versions: segmented, unsegmented, and Latin chars. The DzNER dataset is open to the public in order to expand the work for the Algerian dialect.

We intend to expand the DzNER dataset in terms of entity quantity and domain diversity in the future. In the NER task, we also intend to investigate the impact of normalization and data augmentation on the Algeria dialect.

References

1. Mesmia, F., Haddar, K., Friburger, N., Maurel, D.: CasANER: Arabic named entity recognition tool. In: Intelligent Natural Language Processing: Trends and Applications, pp. 173–198 (2018)
2. Helwe, C., Elbassuoni, S.: Arabic named entity recognition via deep co-learning. Artif. Intell. Rev. **52**(1), 197–215 (2019). https://doi.org/10.1007/s10462-019-09688-6
3. Shaalan, K.: A survey of Arabic named entity recognition and classification. Comput. Linguist. **40**, 469–510 (2014)
4. Youssef, A., Elattar, M., El-Beltagy, S.: A Multi-embeddings approach coupled with deep learning for Arabic named entity recognition. In: 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES), pp. 456–460 (2020)
5. Grishman, R., Sundheim, B.: Message understanding conference-6: a brief history. In: COLING 1996 Volume 1: The 16th International Conference On Computational Linguistics (1996)
6. Saadane, H., Habash, N.: A conventional orthography for Algerian Arabic. In: Proceedings of the Second Workshop on Arabic Natural Language Processing, pp. 69–79 (2015)
7. Meftouh, K., Bouchemal, N., Smaili, K.: A study of a non-resourced language: the case of one of the Algerian dialects. In: The Third International Workshop on Spoken Languages Technologies for Under-resourced Languages-SLTU 2012, pp. 1–7 (2012)
8. Saadane, H.: Le traitement automatique de l'arabe dialectalisé: aspects méthodologiques et algorithmiques. Université Grenoble Alpes (2015)

9. Guellil, I.: Sentiment analysis approach for Arabic dialects texts analysis based on automatic translation: application to the Algerian dialect. Doctoral thesis. Ecole nationale Supérieure d'informatique (2021)
10. Kerras, N., Lahssan Baya, M.: Standard Arabic and Algerian languages: a sociolinguistic approach and a grammatical analysis. Íkala, Revista De Lenguaje Y Cultura, p. 24 (2019)
11. Gahbiche-Braham, S., Maynard, H., Yvon, F.: Traitement automatique des entit´es nommées en arabe : détection et traduction. Traitement Autom. Des Lang. (TAL) **54**, 101–132 (2014)
12. Maloney, J., Niv, M.: TAGARAB: a fast, accurate Arabic name recognizer using highprecision morphological analysis. In: Computational Approaches to Semitic Languages (1998)
13. Shaalan, K., Raza, H.: NERA: named entity recognition for Arabic. J. Am. Soc. Inform. Sci. Technol. **60**, 1652–1663 (2009)
14. Elsebai, A., Meziane, F., Belkredim, F., et al.: A rule based Persons names Arabic extraction system. Commun. IBIMA **11**, 53–59 (2009)
15. Helwe, C., Dib, G., Shamas, M., Elbassuoni, S.: A semi-supervised BERT approach for Arabic named entity recognition. In: Proceedings of the Fifth Arabic Natural Language Processing Workshop, pp. 49–57 (2020)
16. Sa'a, D., Tawalbeh, S., Al-Smadi, M., Jararweh, Y.: Using bidirectional long short-term memory and conditional random fields for labeling Arabic named entities: a comparative study. In: 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 135–140 (2018)
17. Benali, B., Mihi, S., El Bazi, I., Laachfoubi, N.: New approach for Arabic named entity recognition on social media based on feature selection using genetic algorithm. Int. J. Electr. Comput. Eng. **11**, 1485 (2021)
18. Balla, H., Delany, S. Exploration of approaches to arabic named entity recognition. In: CLEOPATRA@ ESWC, pp. 2–16 (2020)
19. Abdallah, S., Shaalan, K., Shoaib, M.: Integrating rule-based system with classification for arabic named entity recognition. In: Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Processing. CICLing 2012. Lecture Notes in Computer Science, vol. 7181, pp. 311–322. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28604-9_26
20. Shaalan, K., Oudah, M.: A hybrid approach to Arabic named entity recognition. J. Inf. Sci. **40**, 67–87 (2014)
21. Hkiri, E., Mallat, S., Zrigui, M.: Arabic-English text translation leveraging hybrid NER. In: Proceedings of The 31st Pacific Asia Conference on Language, Information And Computation, pp. 124–131 (2017)
22. Zirikly, A., Diab, M.: Named entity recognition system for dialectal Arabic. In: Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), pp. 78–86 (2014)
23. Torjmen, R., Haddar, K.: The automatic recognition and translation of Tunisian dialect named entities into modern standard Arabic. In: Bekavac, B., Kocijan, K., Silberztein, M., Šojat, K. (eds.) NooJ 2020. CCIS, vol. 1389, pp. 206–217. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-70629-6_18
24. Antoun, W., Baly, F., Hajj, H.: AraBERT: transformer-based model for Arabic language understanding. ArXiv Preprint ArXiv:2003.00104. (2020)
25. El-Khair, I.: 1.5 billion words Arabic corpus. ArXiv Preprint ArXiv:1611.04033. (2016)
26. Zeroual, I., Goldhahn, D., Eckart, T., Lakhouaja, A.: OSIAN: open source international Arabic news corpus-preparation and integration into the CLARIN infrastructure. In: Proceedings of The Fourth Arabic Natural Language Processing Workshop, pp. 175–182 (2019)

27. Abdul-Mageed, M., Elmadany, A., Nagoudi, E.: ARBERT MARBERT: deep bidirectional transformers for Arabic. ArXiv Preprint ArXiv:2101.01785 (2020)
28. Abdaoui, A., Berrimi, M., Oussalah, M., Moussaoui, A.: DziriBERT: a pre-trained language model for the Algerian dialect. ArXiv Preprint ArXiv:2109.12346 (2021)
29. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv Preprint ArXiv:1810.04805 (2018)
30. Touileb, S., Barnes, J.: The interplay between language similarity and script on a novel multi-layer Algerian dialect corpus. In: Findings of the Association For Computational Linguistics: ACL-IJCNLP 2021, pp. 3700–3712 (2021)
31. Seddah, D., et al.: Building a user-generated content North-African Arabizi treebank: tackling hell. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1139–1150 (2020)
32. Ratnaparkhi, A.: Maximum entropy models for natural language ambiguity resolution. University of Pennsylvania (1998)
33. Yaghan, M.: Arabizi: a contemporary style of Arabic Slang. Design Issues **24**, 39–52 (2008)
34. Abainia, K.: Dzdc12: a new multipurpose parallel Algerian Arabizi–French code switched corpus. Lang. Resour. Eval. **54**, 419–455 (2020)