Benha Journal of Applied Sciences (BJAS) print: ISSN 2356–9751
Vol. (9) Issue (5) (2024), (45-48) online: ISSN 2356–976x
http://bjas.bu.edu.eg

Advancing Arabic Scientific Text Analysis: Evaluating Machine Learning Models for Named Entity Recognition

Nourhan M. Marzouk*, Hamada A. Nayel and Ahmed A. Elsawy
Department of Computer science, Faculty of Computers and Artificial intelligence, Benha University
E-mail: norhan.marzoq17@fci.bu.edu.eg

Abstract

The task of named entity recognition in Arabic text, particularly within the scientific and medical domains, presents unique challenges due to the language's rich morphology, the scarcity of resources, and dialectical diversity. This study evaluates the efficacy of Conditional Random Fields (CRF), Support Vector Machines (SVM), and Stochastic Gradient Descent (SGD) models for named entity recognition in Arabic scientific texts. These models have been implemented on a self-collected dataset consisting of Arabic abstracts of theses. The named entities identified in the dataset include proteins, DNA, RNA, cell types, and cell lines. Focusing on the scientific domain, our comparative analysis reveals significant performance differences among the models, with hybrid approaches showing promising results. SGD, SVM, and CRF achieved F1-scores of 0.96, 0.91, and 0.80, respectively. The results demonstrate the effectiveness of the proposed models. The research contributes to Arabic natural language processing by highlighting model strengths and guiding future selections and development of named entity recognition models.

Keywords: Arabic Named Entity Recognition, Entity Extraction, Arabic NLP, Machine Learning.

1. Introduction

Named Entity Recognition (NER) is a cornerstone of Natural Language Processing (NLP), playing an indispensable role in extracting and categorizing essential data elements such as personal names, organizations, and geographical locations from vast textual landscapes. Successful identification of these entities not only enhances the comprehension and retrieval of information but also lays the groundwork for advanced NLP applications including relationship extraction, sentiment analysis, machine translation, and knowledge extraction. The evolution of NER technology, primarily in English, has significantly contributed to the development of sophisticated tools that aid in the efficient processing of large datasets, highlighting the technology's potential to transform data into actionable insights across various sectors [1].

However, adapting NER to Arabic texts, especially those within the academic and scientific realms, presents unique challenges. Arabic's intricate morphological structure, syntactic complexity, and rich tapestry of dialects pose substantial obstacles for NER systems. The situation is further complicated in scientific literature, where the presence of domain-specific terminology and abbreviations necessitates NER models with heightened sensitivity to context and linguistic nuances [2], [3]. Moreover, the limited availability of specialized, annotated datasets for Arabic hampers the development and fine-tuning of robust NER models, thus stalling progress in this essential area of NLP [4].

Named Entity Recognition for Arabic (NERA) must navigate these linguistic intricacies to unlock the full potential of Arabic content, which spans across the Middle East and North Africa and is recognized by the United Nations for its global significance. The quest for efficient NERA systems is often hampered by the language's complex grammatical structures and the specificity of terminologies used in various professional fields, including medicine and engineering. This underscores the urgent need for effective NERA solutions that can cater to Arabic's unique attributes [5].

Acknowledging the pressing demand for advanced NERA capabilities, recent research endeavors have turned to deep learning methodologies. By harnessing expansive datasets and sophisticated computational models, these approaches aim to grasp the subtleties of Arabic's linguistic features and dialectal variations. The advancements in this domain are poised to revolutionize Arabic text processing applications, facilitating enhanced information retrieval, text mining, and machine translation, thereby bridging knowledge gaps and fostering global information exchange [6].

This study embarks on a comparative analysis of three widely adopted machine learning models, namely CRF, SVM, and SGD, utilizing a self-collected dataset. Our objective is to scrutinize the efficacy of these models in the domain of Arabic scientific NER. By providing a detailed evaluation of their performance, we aim to contribute valuable insights to the ongoing advancement of Arabic NLP. This paper explores the existing literature on NER with a specific focus on Arabic, addressing the inherent challenges and reviewing seminal works. We will outline our research methodology, encompassing data collection, model training, and evaluation, culminating in a comprehensive presentation and discussion of our findings. Through this analytical journey, we aspire to not only pinpoint the most effective models for Arabic scientific NER but also encourage continued innovation in NER strategies, particularly for languages that have traditionally been overlooked in NLP research.

2. Related work

In this section, we explore the advancements and future directions in NERA, highlighting the extraction of information from the burgeoning amount of Arabic scientific texts and the challenges posed by Arabic’s rich morphological features. Nayel et al. [5] delve into the realm of NER within Arabic medical documents, with a particular emphasis on identifying disease entities. They evaluated various deep learning approaches, including CRF, LSTM, LSTM-CRF, and BiLSTM, utilizing a specially curated dataset mentioned in [4]. Their findings highlight the enhanced efficacy of combined models such as LSTM-CRF and BiLSTM-CRF, which achieved F1-scores of 0.97 and 0.94, respectively.

Qu et al. [6] present a comprehensive overview of the progress in NERA, addressing the increasing importance of extracting information from the growing volume of Arabic content on the internet. This study also discusses the challenges posed by the rich morphology of the Arabic language and the future directions of NERA. Mahdhaoui and Mars [7] introduced an active learning approach to NERA using the pre-trained AraGPT2 model. They highlight the significance of utilizing state-of-the-art language models tailored to the Arabic language and the application of active learning techniques to improve the efficiency and accuracy of ANER systems.

Alsaaran and Alrabiah [2] investigate the performance of various deep neural network architectures in conjunction with BERT for NER of classical Arabic. They demonstrate the value of fine-tuning pre-trained language models for languages with limited resources like Arabic, highlighting significant improvements in NER tasks. In [8], Shaker et al. introduce a new dataset for NER in the Arabic language and propose the use of LSTM units and Gated Recurrent Units (GRU) to build NER models. The proposed models achieved satisfactory results, indicating the effectiveness of recurrent neural network units in Arabic NER tasks. Bazi and Laachfoubi [3] compared the performance of word representation methods for NERA. The evaluation using the AQMAR dataset shows that the feature of word representations significantly boosts the supervised NER system in Arabic. The performance is further improved when different approaches are combined.

3. Materials and Methods

In the proposed method, we introduce an innovative NLP pipeline specifically tailored for Arabic text, which addresses its rich linguistic intricacies. As depicted in Figure 1, our approach transcends traditional methodologies by incorporating a novel dataset collection phase focused on Arabic scientific discourse. This is complemented by an enhanced feature extraction process that accounts for Arabic's unique morphological features, and the strategic development of machine learning models. We employ a combination of CRF, SVM, and SGD, culminating in a rigorous evaluation phase. This phase is crucial for validating the model’s efficacy against the complex backdrop of Arabic semantics.

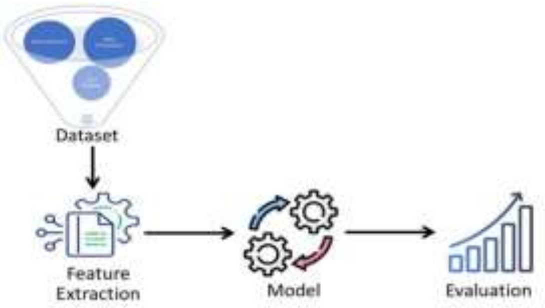

Fig. (1) The general structure of the proposed pipeline

3.1 Dataset

The dataset has been used in this study is a collection of theses' Arabic abstracts that have been extracted from Egyptian Universities Libraries Consortium (EULC)[1]. A set of Arabic keywords have been used as a seed for search related to microbiology.

Table (1) Search keywords.

Arabic keyword	الطحالب	البكتريا	الفطريات	الفيروسات	الاحياء الدقيقه
English translation	Algae	Bacteria	Fungi	Virus	Microbiology

[1] http://www.eulc.edu.eg/eulc_v5/libraries/start.aspx?ScopeID=1.&fn=SearchInterFace&flag=Thesis

Table (1) shows these keywords in Arabic and equivalent English terms.

Automatic annotation of these abstracts has been performed by inputting the English versions, available at the EULC, into a well-established biomedical named entity recognizer, ABNER [9]. This system is capable of recognizing five types of entities: proteins, DNA, RNA, cell types, and cell lines. The dataset comprises over 70,000 tokens, including 1,214 protein names, 47 DNA sequences, 30 RNA sequences, 45 cell types, and 15 cell lines.

3.2 Feature Extraction

A set of features has been extracted from the given dataset to discriminate among biomedical entities effectively. These features include the word itself, its Part of Speech (PoS) tag, word length, preceding word, and subsequent word.

3.3 Model Training

Our study aims to evaluate the performance of established machine learning models for Arabic scientific NER, focusing on CRF, SVM, and SGD to investigate how these models manage the complex linguistic structure of Arabic. CRF is celebrated for its adeptness in sequence labeling, capable of modeling the conditional probabilities of sequences, which makes it particularly suited to the contextual intricacies of Arabic texts [10]. SVM is noted for its ability to navigate high-dimensional feature spaces and utilizes the kernel trick for non-linear separations, advantageous for classifying the diverse linguistic characteristics of Arabic [11]. SGD is appreciated for its iterative parameter updates, making it an efficient optimizer for NER models, accommodating the vast datasets and complex feature spaces characteristic of Arabic NER tasks [12]. To implement the proposed model, an open-source Python implementation of CRF named **'sklearn_crfsuite'** has been used. The main features fed into the CRF model include PoS, predecessor word, current word, successor word, and digit information. For SVM and SGD, the **'sklearn'** package's optimized implementations have been employed. The linear kernel has been used for SVM, and a random seed of 42 has been set for all classifiers to ensure reproducibility.

3.4 Evaluation

To evaluate the models, we have randomly split the dataset into training, development, and test sets in a 6:2:2 ratio, respectively. Recall, Precision, and F1-score are the appropriate evaluation metrics for sequence tagging problems, including NER [13]. We calculated all three metrics for the implemented classification algorithms.

4. Results and Discussion

The experimental findings, presented in Table 2, demonstrate the efficacy of the machine learning models applied to NERA in scientific texts, with SGD showing superior performance. These results suggest the potential for further improvements through the exploration of diverse algorithms, expansion of datasets, and enhancement of preprocessing techniques.

Table (2) Results of Proposed Models

Algorithm	Data	P	R	F1-score
SGD	Dev	0.93	0.97	0.95
	Test	0.92	0.97	**0.96**
CRF	Dev	0.78	0.85	0.80
	Test	0.76	0.86	0.80
SVM	Dev	0.92	0.93	0.93
	Test	0.92	0.92	0.91

5. Conclusion and Future work

In conclusion, our research underscores the effectiveness of machine learning models, particularly SGD, in performing NER tasks within Arabic scientific texts. While CRF and SVM demonstrated commendable results, SGD's superior optimization capabilities make it especially suitable for the complex linguistic environment of Arabic texts. However, our study has certain limitations. The dataset used is restricted to scientific abstracts, which may not fully capture the broader linguistic and contextual diversity of the Arabic language. Additionally, the models may not effectively encompass all dialectical variations, impacting their generalizability. Looking ahead, integrating advanced deep learning and ensemble methodologies could potentially enhance the accuracy of our models. Expanding the dataset to include a wider variety of text sources and incorporating more sophisticated preprocessing techniques could help in addressing some of the noted limitations. Future research should also consider the application of contextual embeddings, which could improve the models' ability to grasp subtle linguistic nuances and dialectical variations in Arabic texts. Such advancements could significantly boost the performance of NER systems, paving the way for more comprehensive and accurate linguistic analysis in the realm of Arabic NLP.

References

[1] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural Architectures for Named Entity Recognition," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Knight, A. Nenkova, and O. Rambow, Eds., San Diego, California: Association for Computational Linguistics,

Jun. 2016, pp. 260–270. doi: 10.18653/v1/N16-1030.

[2] N. Alsaaran and M. Alrabiah, “Classical Arabic Named Entity Recognition Using Variant Deep Neural Network Architectures and BERT,” *IEEE Access*, vol. 9, pp. 91537–91547, 2021, doi: 10.1109/ACCESS.2021.3092261.

[3] I. El Bazi and N. Laachfoubi, “Arabic Named Entity Recognition using Word Representations,” *International Journal of Computer Science and Information Security,* vol. 14, pp. 956–965, Mar. 2016.

[4] N. Alshammari and S. Alanazi, “An Arabic Dataset for Disease Named Entity Recognition with Multi-Annotation Schemes,” *Data (Basel)*, vol. 5, p. 60, Mar. 2020, doi: 10.3390/data5030060.

[5] H. Nayel, N. Marzouk, and A. Elsawy, “Named Entity Recognition for Arabic Medical Texts Using Deep Learning Models,” in *2023 Intelligent Methods, Systems, and Applications (IMSA)*, IEEE, Jul. 2023, pp. 281–285. doi: 10.1109/IMSA58542.2023.10217658.

[6] X. Qu, Y. Gu, Q. Xia, Z. Li, Z. Wang, and B. Huai, “A Survey on Arabic Named Entity Recognition: Past, Recent Advances, and Future Trends,” *IEEE Trans Knowl Data Eng*, vol. 36, no. 3, pp. 943–959, Mar. 2024, doi: 10.1109/TKDE.2023.3303136.

[7] A. and Z. M. Mahdhaoui Hassen and Mars, “Active Learning with AraGPT2 for Arabic Named Entity Recognition,” in *Advances in Computational Collective Intelligence*, J. and G. L. and N. M. and T. J. and V. G. and K. A. Nguyen Ngoc Thanh and Botzheim, Ed., Cham: Springer Nature Switzerland, 2023, pp. 226–236.

[8] Shaker, Alaa, Alaa Aldarf, and Igor Bessmertny. "Using lstm and gru with a new dataset for named entity recognition in the arabic language." arXiv preprint arXiv:2304.03399 (2023).

[9] B. Settles, “ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text,” *Bioinformatics*, vol. 21, no. 14, pp. 3191–3192, Mar. 2005, doi: 10.1093/bioinformatics/bti475.

[10] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” in *Proceedings of the Eighteenth International Conference on Machine Learning*, in ICML ’01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[11] A. H. AbuElAtta, M. Sobhy, A. A. El-Sawy, and H. Nayel, “Arabic Regional Dialect Identification (ARDI) using Pair of Continuous Bag-of-Words and Data Augmentation,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 11, 2023, doi: 10.14569/IJACSA.2023.0141125.

[12] L. Bottou and Y. Cun, “Large Scale Online Learning,” in *Advances in Neural Information Processing Systems*, S. Thrun, L. Saul, and B. Schölkopf, Eds., MIT Press, 2003. [Online]. Available:https://proceedings.neurips.cc/paper_files/paper/2003/file/9fb7b048c96d44a0337f049e0a61ff06-Paper.pdf

[13] H. L. Shashirekha and H. A. Nayel, “A comparative study of segment representation for biomedical named entity recognition,” in *2016 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016*, 2016. doi: 10.1109/ICACCI.2016.7732182.