Enhancing Arabic Biomedical Named Entity Recognition Using Transformer-Based Representations and CRF Sequence Labeling

Nassima Gannoune[1], Abdellah Madani[2], Mohamed Kissi[3]
Chouaib Doukkali University, El Jadida, Morocco[1, 2]
Hassan II University, Mohammédia, Morocco[3]

***Abstract*—Electronic health records have witnessed tremendous growth in recent years. To make these documents useful for decision-making, high-performance natural language processing (NLP) systems are essential. Named entity recognition (NER) is a critical task for many biomedical NLP applications that contribute to improving patient care, drug discovery, and disease surveillance. However, despite its status as an official language in more than 22 countries, Arabic is largely neglected in this field. Only limited work has been done and there is little well-annotated public dataset. This work tackles these issues by proposing an NER model capable of recognizing entities such as diseases, symptoms, and organs from biomedical Arabic text. To achieve this, an annotated dataset was first developed, followed by fine-tuning the CAMeLBERT model, a BERT-based model, in conjunction with a conditional random field (CRF) layer. The evaluation results indicate that the CAMeLBERT+CRF model achieves the best overall F1-score of 90%, surpassing other base models such as CAMeLBERT and AraBERT. This study demonstrates the effectiveness of the hybrid approach and underscores the importance of transfer learning techniques for low-resourced and morphologically rich languages like Arabic.**

***Keywords*—Named entity recognition; electronic health records; natural language processing; CAMeLBERT; CRF**

I. Introduction

Clinical patient records include personal details such as age, weight, past medical history and results of various laboratory tests, among others. Most of this information is presented as unstructured text, and extracting useful information from these records is a crucial and challenging task in biomedical natural language processing (NLP). Named entity recognition is a widely studied topic in the area of biomedical NLP that plays a significant role in information retrieval. The aim of the task is to identify biomedical entities that belong to certain types. Pioneering work in named entity recognition (NER) was presented at the sixth message understanding conference (MUC6) in 1995 [1], aiming to identify entities such as people, organizations, locations, and other relevant terms. Subsequently, this task has received considerable attention from researchers, particularly in the biomedical domain, to identify medical concepts such as diseases, drug names, and medical procedures, which is essential for a number of downstream tasks including text summarization [2], relation extraction [3], question answering [4], and clinical decision support [5]. Most works in biomedical NER have been devoted to the English language, making significant advancements in diagnostics, treatment, and research. However, limited research has been performed on the Arabic language. Most current systems have been developed and evaluated on general-domain. These models are optimized to recognize common entity types such as people, places, and organizations, but do not capture the specialized terms, and unique language found in biomedical texts. Designing high-performance Arabic biomedical NER systems is challenged by various factors, including the scarcity of annotated biomedical datasets in Arabic which limits the creation and evaluation of efficient models. For standard NLP pipelines, the morphological richness and syntactic complexity of Arabic, including its complex morphology, diacritical variations, and inherent ambiguity, present significant obstacles, especially in specialized fields like medicine. Recent advancements in pre-trained transformer-based models have improved Arabic NLP on a variety of tasks. However, these models are not specifically tailored for biomedical applications, they are typically trained on general-purpose models. As a result, there is still ongoing research into adapting them to domain-specific tasks like Arabic biomedical NER.

The objective of this study is to address the problem of improving Arabic biomedical NER performance in a low-resource setting by creating a custom Arabic biomedical dataset and exploiting transformer-based representations along with structured sequence labeling. Specifically, this paper evaluates CAMeLBERT, a pre-trained BERT-based [6] model tailored for modeling Arabic linguistic variation, when paired with a Conditional Random Field (CRF) layer [7] to capture label dependencies in biomedical entity recognition. The main contributions of this work can be summarized as follows:

- Construction of a newly annotated Arabic biomedical NER dataset to overcome the balance between quality of annotations and resources.
- Design and fine-tuning of transformer-based model with Conditional Random Fields (CRF) for capturing contextual and sequential dependencies in biomedical Arabic text.
- Experimental evaluation of the proposed transformer–CRF model on Arabic biomedical NER task, including comparison with baseline methods for examining their effectiveness and robustness.

The rest of the paper is structured as follows: the related work on named entity recognition is presented in Section II.

Section III describes the dataset procedures from collection to annotation. Section IV presents the model architecture in detail. The experimental results are shown and discussed in Section V. Finally, Section VI concludes the paper.

II. Related Work

Arabic NER can be divided into three main approaches: rule-based, machine learning-based, and deep learning-based. To identify pertinent entities, the rule-based approach uses manually created rules and patterns. It takes both domain-specific knowledge and a solid foundation of linguistics to construct these rules. The study described in [8] a new rule-based approach for named entity recognition (NER) in Classical Arabic texts. It tackles the challenges of Arabic morphology and script complexity by using a mix of trigger words, patterns, gazetteers, and blacklists to identify entities like names and locations. The results show about 90% precision and recall, aiming to improve automated rule updates and coverage over existing systems. However, the study [9] employed a supervised machine learning approach that integrates multiple types of features word-level, morphological, and knowledge-based to perform named entity recognition in Classical Arabic, using Naive Bayes classifier, the system learns to predict the entity type based on the extracted features. The study reports that, using a Naive Bayes classifier with various feature combinations, the system achieved promising results, with an overall precision of around 86%, recall of 75%, and an F-measure near 80%. Another study in [10] proposed a method that involves creating an extensive Arabic NER dataset that covers seven domains and has nine categories of named entities using BIOES tagging in order to better manage nested entities. Two recurrent neural network architectures, LSTM and GRU, were developed to build an NER model achieving an accuracy around 80%. On the other hand, Sadallah et al. [11] presented ANER a BERT-based model for Arabic named entity recognition able to recognize 50 different entity classes, built upon fine-tuning the pre-trained AraBERTv0.2-base model using the WikiFANE Gold corpus. The system achieved an F1 score of 88.7%, which beats CAMeL Tools' F1-score of 83% on the ANERcorp dataset. Moreover, the model exhibited robust generalization capabilities, attaining a 77.7% F1 score when assessed on out-of-domain data from the NewsFANE Gold dataset. Furthermore, ANER was deployed on a user-friendly web interface to make it accessible.

Before diving into the biomedical NER task in Arabic, it is important to acknowledge research conducted in other languages. English, as the most dominant language, and Chinese [12, 13, 14, 15], due to the large number of biomedical publications. Furthermore, languages such as Spanish [16, 17] have also been explored, driven by the extensive Spanish-speaking regions. Besides, French [18, 19] and German [20] have fewer works, but they still offer valuable resources. In order to overcome the difficulties caused by dataset limitations in low-resource languages, recent studies have used multilingual models like mBERT, XLM-R, and BioBERT-multilingual in the biomedical domain. These models make use of cross-lingual transfer, by training on high-resource languages like English and then applying the acquired knowledge to other languages [21]. Currently there are very few Arabic biomedical NER systems, due to the significant lack of annotated datasets.

Boudjellal et al. [22] proposed ABioNER, a BERT-based model designed to recognize biomedical named entities like diseases and treatments in Arabic texts. The model was pre-trained on a curated, manually annotated, silver-standard biomedical corpus with annotations for 13 different semantic categories, yet the study focused on "Disease or Syndrome" and "Therapeutic or Preventive Procedure". According to the results, ABioNER outperformed the AraBERT and multilingual BERT cased models, obtaining an F1-score of 85%, indicating improved effectiveness for biomedical NER tasks in Arabic. Although this work provides positive indications for the success of domain-oriented pre-training, it remains limited due to its applicability to a few entity types and by its lack of examination of linguistic variation.

Alanazi [23] presented NAMERAMA, a specialized NER tool based on Bayesian Belief Network (BBN), that extracts particular medical entities sush as disease names, symptoms, treatments, and diagnosis techniques. The authors used the AMIRA to perform tokenization and POS tagging, followed by manual correction due to the complex morphological structure of Arabic language. They extracted set of features such as the POS tag of the word, gazetteers, stopwords, lexical triggers and a set of grammar patterns, necessary to enhance the performance of the classifier. The system was tested on data containing 26 articles about cancer obtained from the King Abdullah Bin Abdulaziz Arabic Health Encyclopaedia (KAAHE) website. The results show that the BBN approach outperformed the baseline system by 21%, achieving an overall F-measure of 71.05%, with highest score in recognizing disease names with 98.10% and the lowest in recognizing symptoms with 41.66%. These results indicate that combining the probabilistic inference with the linguistic features, is effective for biomedical NER in Arabic. However, this approach is labor-intensive, not portable to other domains, and not easily scalable to larger collections of biomedical text.

Hamad and Abushaala [24] developed a biomedical NER model for extracting medical named entities of cancer in Arabic texts, utilizing an SVM model. The training corpus, constructed by Alanazi from the King Abdullah Bin Abdulaziz Arabic Health Encyclopedia (KAAHE), consists of various cancer types annotated with disease, symptoms, diagnosis, and treatment entity types. The authors extracted several features, including N-gram FastText, N-gram POS, TF-IDF, and non-medical word matching, fed into the SVM model as a vector. This approach achieved an overall F-score of 77.61% and showed high performance across most entity types, outperforming the BBN model. Although an improvement over a previous work, this method still requires significant feature engineering.

The study presented in [25] suggested two approaches for Arabic biomedical NER based on contextual embeddings. The first one fine-tuning BERT and AraBERT models to the NER task. The second uses AraBERT as a feature extractor, where the contextual embeddings are used as an input for several machine learning classifiers such as SVM, Decision Tree, and AdaBoost. The dataset employed in this work was devoted to NER tasks for diseases, with various annotation schemes. This hybrid approach leverages the contextual understanding of transformers and the simplicity of machine learning classifiers to improve recognition

accuracy. The fine-tuned AraBERT performed the best, particularly when handling single entities under the IO scheme and consecutive entities under the BIES or IOBES schemes. This study shows that transformer-based models outperform traditional approaches.

Overall, existing work on Arabic biomedical NER are based on either small domain-specific datasets, involves heavy manual feature engineering, or leverages general-purpose transformer models with minimal considerations on how suitable they are for biomedical Arabic text. In addition, decoding strategies at the sequence level and the reasons why some Arabic pre-trained models perform better over others for this particular task have received little attention. Motivated by these limitations, this study explores a CAMeLBERT+CRF based architecture in conjunction with a newly annotated biomedical Arabic dataset, with an aim of not only advancing the performance, but also drawing empirical insights concerning transfer learning for morphologically rich and low-resource biomedicine language.

III. Dataset

Due to the limited availability of annotated resources for Arabic biomedical NER, the work presented in this paper is based on a novel dataset tailored to the biomedical Arabic domain. The proposed dataset extends entity coverage by incorporating three clinically relevant entity types: disease, organ, and symptom. This entities design allows for richer biomedical information extraction as well as more realistic representation of medical texts. The dataset obtained is lexically and contextually diverse with respect to the biomedical topics, which allows for a challenging evaluation of models. The selection of this dataset is motivated by the fact that it contains more entities, is domain-specific, and more appropriate for assessing transformer-based models in Arabic biomedical NER.

This section outlines all the steps involved from data collection, processing, and annotation to create the dataset that will be used later to train and test the Arabic BioNER model. Fig. 2 illustrates the steps of dataset construction.

A. Data Collection

A corpus of unstructured texts was first gathered from a publicly accessible Arabic medical website, WebTeb [28], a well-known and trustworthy Arabic health information platform that offers a variety of medical content, including disease overviews, symptoms, treatments, medication information, and health articles. The focus was placed on a curated section called "أكثر الأمراض انتشاراً في الوطن العربي", The most common diseases in the Arab world. This section includes a variety of structured articles about common medical conditions in Arab countries. Special attention was given to the section containing diseases and symptoms. All of which support the objectives of entity recognition in the biomedical domain. The BeautifulSoup library was employed, a reliable web scraping tool, to extract all targeted documents from the website representing the most common diseases in the Arab world, such as Heart Disease, High Blood Pressure, Allergies, etc. The final corpus contains various diseases and their symptoms, providing a robust basis for disease NER in Arabic text.

B. Data Preprocessing

To prepare raw text for further processing, several steps were carried out to enhance the quality of the dataset and to facilitate effective model training. The first task is data cleaning and normalization, during this process, any irrelevant metadata like external links, hashtags and so on were remove. Then, Non-Arabic letters and unwanted characters like parentheses were excluded from the corpus to minimize noise. Then, diacritics were also eliminated from the corpus, they are very important to determine the correct pronunciation, but for text processing, their removal can improve word recognition. Afterwards, the Farasa segmenter was used [26], a fast and accurate Arabic segmentation tool, to tokenize and segment the sentences. All prefixes and suffixes were segmented like except the definite article "ال", and other suffixes, to preserve the correct semantic meaning of the word, and keep the BIO tagging coherent. For example, the word الالتهاب الرئوي (Pneumonia) can be detached to become ال+التهاب ال+رئوي, resulting in incorrect BIO tagging.

C. Annotation

For the annotation step, a semi-automatic approach was adopted. First, a rule-based matcher was used to identify entities predefined in the three lexicons manually created, categorized into diseases, symptoms, and organs. Then, the annotations were manually corrected using the IOB (Inside, Outside, Beginning) tagging scheme to enhance the quality of the annotations. Table II presents an example of an annotated sentence with the BIO scheme. The diverse biomedical contexts in the corpus make it possible to test the generalization ability of the models to different kinds of entities. Table I describes the statistics of the dataset including the total of tokens, labeled tokens, non-labeled tokens, and entity types. Fig. 1 illustrates the distribution of each entity type in the dataset. As expected, the majority of tokens about 87% of the corpus, are classified as O. Diseases has the highest share of annotated entities at 7.4%, followed by symptoms and organs, each at 2.8%. This illustrates the natural imbalance in biomedical texts, where disease mentions are tending to be prevalent than other entity types.

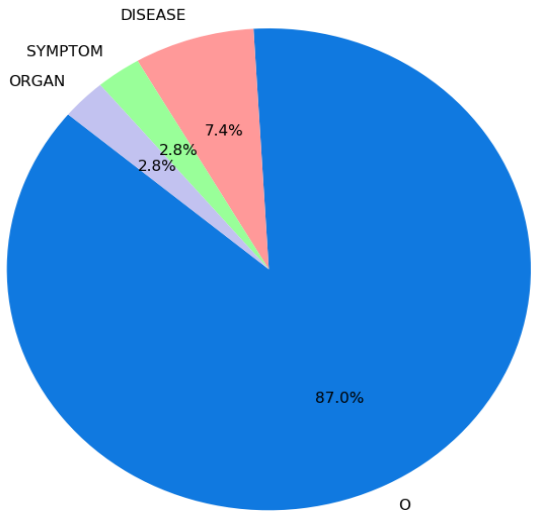

Fig. 1. Entity distribution.

TABLE I. DATASET OVERVIEW

Metric	Value
Total tokens (words)	12,682
Tokens labeled as entities	1,643
Non-entity tokens (O)	11,039
Entity types	O, B-DISEASE, I-DISEASE, B-ORGAN, I-ORGAN, B-SYMPTOM, I-SYMPTOM

TABLE II. EXAMPLE OF DATASET WITH BIO TAGS

Word	Translation	Tag
من	Whoever	O
يعاني	suffers	O
من	from	O
الذبحة	angina	B-DISEASE
الصدرية	pectoris	I-DISEASE
أو	or	O
من	from	O
احتشاء	infarction	B-DISEASE
عضلة	myocardial/muscle	I-DISEASE
القلب	heart	I-DISEASE
قد	may	O
يشعر	feel	O
بألم	pain	B-SYMPTOM
أو	or	O
ضغط	pressure	B-SYMPTOM
على	on	I-SYMPTOM
جدار	the wall	I-SYMPTOM
الصدر	chest	I-SYMPTOM

IV. METHODOLOGY

The CAMeL-Lab model bert-base-arabic-camelbert-mix [29] was chosen as the base encoder. CAMeLBERT is a pretrained Arabic BERT-based model introduced by study [6], trained on a diverse corpus of modern standard Arabic and dialectal Arabic, which enables it to learn richer lexical variations and morphological patterns. The model can learn more general lexical features along with richer morphological variations. This characteristic is extremely useful in biomedical Arabic, where the same medical concept may be written in different forms, informal and orthographic variations. That is why CAMeLBERT is more generalized for biomedical entities and more robust to lexical and morphological variations. Besides, CAMeLBERT provides efficient fine-tuning with a limited number of epochs, even with small amounts of annotated data, and thus is well suited to low-resource biomedical NER tasks. A Conditional Random Field (CRF) layer was also added to improve the sequence labeling capabilities of the base model, which is used to capture label dependencies between entity spans. The CRF helps model the structural constraints of valid label sequences and is particularly effective in improving performance on sequence tagging tasks such as NER. As shown in Fig. 3, input biomedical Arabic text is first tokenized into subword units, where special tokens [CLS] and [SEP] are automatically added by the tokenizer but not used during prediction. CAMeLBERT generates contextualized embeddings for each token, capturing semantic and syntactic dependencies across the sequence. Finally, the CRF layer decodes the most likely label sequence by modeling dependencies between output tags, ensuring valid BIO annotation. The overall architecture of the system is illustrated in Fig. 2.

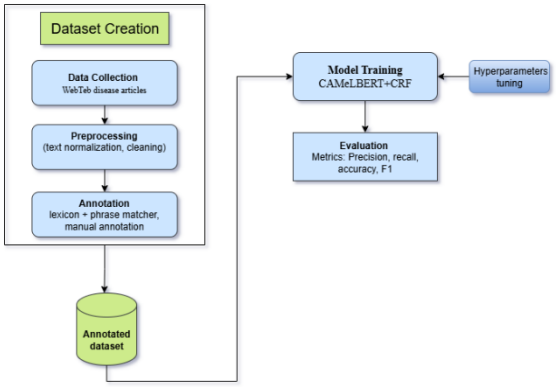

Fig. 2. Architecture of the proposed system.

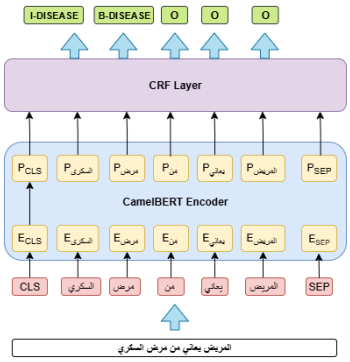

Fig. 3. Architecture of the proposed CAMeLBERT-CRF model for BioNER.

A. CRF Integration

Arabic BioNER is considered as a sequence labeling problem, so it's important for the model to learn valid transitions between labels, for example ensuring that I-ENTITY tag cannot occur without a preceding B-ENTITY tag. Thereby avoiding the independent decoding of each label. To capture these correlations between labels, a conditional random field was employed for sequence modeling, which avoids independent label decoding. The CRF layer computes the likelihood of the correct label sequence and is trained by maximizing this score over the training set using negative log-likelihood loss. A comparative analysis was carried out against a standard softmax layer, in order to evaluate the benefit of incorporating CRF into the model architecture. Table III illustrates the models predictive performance on the sentence "داء السكري من النوع الثاني", translated in English as Type 2 diabetes. It can be observed that the model without CRF incorrectly labels the token "من" as "O" when it is part of the name of the disease in Arabic, (I-DISEASE after O). Whereas the CRF layer correctly identifies it as part of the disease entity, instead of treating each token separately, it considers the sequence of labels for the entire sentence, improving contextual accuracy.

TABLE III. CRF VS. SOFTMAX FOR BIONER

Token	Prediction with CRF	Prediction without CRF
داء	B	B
السكري	I	I
من	I	O
النوع	I	I
الثاني	I	I

B. Training Procedure

Fine-tuning CAMeLBERT+CRF and other baseline models on the biomedical NER dataset containing entities such as diseases, symptoms, and organs. The Hugging Face Transformers library and Trainer API [27] were used for training. The Optuna framework was used to optimize the hyperparameters, allowing for efficient search with three training epochs, a batch size of eight, and a learning rate of 5e-5. Metrics like precision, recall, and F1-score were used to assess the model's performance. Evaluation considered exact span matches under the BIO labeling scheme and was reported per entity type as well as overall.

C. Baseline Comparison

The performance of the suggested CAMeLBERT+CRF architecture was compared with two baseline models in order to assess its efficacy:

- CAMeLBERT: A BERT-based model pre-trained on Arabic texts with different variants. For its fine-tuning, a standard softmax classifier was used to predict token labels. The impact of the CRF was evaluated using this model as a direct ablation study.
- AraBERT v2: An Arabic pretrained lanaguage model based on Google's BERT architechture. AraBERT was fine-tuned using the same dataset and training parameters. The impact of pretraining corpus specialization was evaluated using AraBERT as a comparative baseline.

V. RESULTS AND DISCUSSION

This section details the results obtained by the three models for the Arabic Biomedical NER task: AraBERT, CAMeLBERT, and the proposed CAMeLBERT + CRF model. All models were fine-tuned using the same annotated dataset on the disease, symptom, and organ entity types.

A. Overall performance

Table IV shows clearly that the CAMeLBERT+CRF approach outperformed the other baseline systems, achieving 90% in overall F1-score. AraBERT remains behind overall, confirming that CAMeLBERT is more suited to the Arabic biomedical corpus. The combination of the CRF layer leads to a significant improvement in precision with a relatively small increase in recall. It shows that the CRF layer can eliminate false positives more efficiently by limiting unreal label transfers. These improvements in performance translate into better generalization across entity types and more coherent tagging of multi-token entities.

TABLE IV. OVERALL PERFORMANCE OF CAMELBERT+CRF, CAMELBERT, AND ARAERT MODELS

Model	Precision (%)	Recall (%)	F1-score (%)
CAMeLBERT+CRF (proposed model)	91.93	88.22	90.03
CAMeLBERT	82.08	87.88	84.88
AraBERT	75.97	82.79	79.24

B. Entity-wise performance

Table V indicates that at the entity levels, CAMeLBERT+CRF consistently outperformed both baselines. Especially in symptoms (F1 = 89.72%) and organs (F1 = 88%), while maintaining strong results for disease (F1 = 91.08%). These findings suggest that the CRF layer yields the maximum benefits for the "Organ" and "Symptom" entities. For Organ, the F1 score increases by 6.34 points and for Symptom by 12.24 points. This indicates that such entity types are more susceptible to boundary ambiguity and complex word formation in Arabic text, which can be well resolved by sequence aware CRF layer. Broader analysis CAMeLBERT outperforms AraBERT in all entities where the difference is more pronounced in Organ and Symptom. This means that the pre-training corpus of CAMeLBERT provides more suitable representations for the domain and organ-related terms.

TABLE V. EVALUATION RESULTS OF CAMELBERT+CRF, CAMELBERT, AND ARABERT MODELS ACROSS DISEASE, ORGAN, AND SYMPTOM ENTITIES

Model	Entity Type	Precision (%)	Recall (%)	F1-score (%)
CAMeLBERT+CRF (proposed model)	Disease	91.93	90.24	91.08
	Organ	92.96	83.54	88.00
	Symptom	90.57	88.89	89.72
CAMeLBERT	Disease	87.13	90.85	88.96
	Organ	76.67	87.34	81.66
	Symptom	75.44	79.63	77.48
AraBERT	Disease	83.82	90.18	86.88
	Organ	65.00	73.39	68.94
	Symptom	69.64	73.58	71.56

C. Discussion

The comparative analysis proved that CAMeLBERT+CRF attained 90% in F1-score, which further verifies the strength of the model across different types of entities. On the entity level, the largest gains are for the symptom type with 89% in F1-score. This result surpasses those obtained with other models AraBERT and CAMeLBERT and shows the effectiveness of the proposed architecture, especially considering the fact the symptom expressions in Arabic are often multi-token expressions and they are often expressed in descriptive, verbose language. But low recall means a portion of them are still missed. The disease entity type, achieved the best performance, suggesting that the names of diseases are easier to detect. This improvement is due to higher proportion of disease mentions in the training set (7.4%), which enables the model to better identify entities in unseen text. Additionally, the vocabulary patterns for disease terms often differ from those for organ and

symptom terms, and so they are less ambiguous. Organ is the most difficult entity type of entity to identify, with its F1 score being the lowest among all models. This can be attributed to the high level of overlapping entity structures in Arabic biomedical texts. When organ names are embedded within larger symptom entities, as in "ألم شديد في الصدر", "الصدر" chest is not only part of the symptom expression "severe chest pain" but also an organ name itself. This creates problems with accurate predictions, resulting in lower F1-score, because the model is less confident in assigning a label. Another challenge is that they often appear with prepositions and definite articles, as in "في الكبد" in the liver. The CRF leyer helps the model learn that "في" in is typically followed by an organ entity, reducing false positives.

Compared to earlier Arabic biomedical NER systems in the literature namely, BBN based NAMERAMA [23] and SVM based cancer NER [24]. This improvement is not only quantitative: it the result of applying transformer-based contextual embeddings with sequence-level decoding which appears to be more effective in capturing the lexical and morphological complexities of biomedical Arabic text. In contrast to previous work that mainly reports baseline performance, the reasons why CAMeLBERT's diverse pretraining and CRF decoding yield better entity recognition are also discussed.

To better understand the limitations of the proposed model, a qualitative error analysis was conducted with a sample of the misclassified examples output by the CAMeLBERT+CRF model. The analysis shows there are a couple of repeating errors sources that could inform the linguistic and domain-specific difficulty of Arabic biomedical NER. Revealing that the majority of the remaining errors are due to ambiguous boundaries of entities in multi-token expressions. In some situations, the model predicts the right entity type but the span is not complete, which happens when the entities contain modifiers or attributive adjectives. The semantic overlap of diseases and symptoms, sometimes these entities are confused and misclassified as they share similar semantic context, since some Arabic terms can refer to symptom or disease according to context of use. The morphological complexity of Arabic, Arabic affixes, clitics and attached prepositions can hide entity boundaries, most notably in the case of organ names and symptom lists. While the CRF layer mitigates some of these problems by making the labels consistent, there are still errors in the cases of rare or highly inflected forms. Furthermore, rare and transliterated medical terms having inconsistent spelling continue to pose difficulties due to their negligible representation in the training data. These results suggest that CAMeLBERT+CRF provides marked performance enhancements for Arabic biomedical NER. However, future advances for this task will depend on larger datasets, more explicit annotation guidelines and stronger domain adaptation approaches.

In summary, these results demonstrate the advantage of using a CRF model on top of the transformer-based contextual embedding for complex, multi-token entities. This methodology provides a strong foundation for further domain-specific NER and the co-expansion of biomedical datasets.

VI. Conclusion and Future Work

This study investigated the problem of the challenge of detecting named entities in Arabic biomedical text, a topic which is still unexplored to a large extent despite its significance for applications in healthcare. By constructing a domain-specific annotated dataset and then implementing a hybrid NER model based on CAMeLBERT, an Arabic pre-trained model, combined with a CRF decoding layer that recognizes the entity types of disease, symptom, and organ. To show that performing effective biomedical NER in Arabic not only requires powerful pretrained representations, but also modeling strategies that bring the language's morphological richness and domain-specific variability into modeling. The proposed CAMeLBERT+CRF significantly outperforms robust baselines, such as CAMeLBERT and AraBERT for all entity types, achieving an F1-score of approximately 90%. The results demonstrate the effectiveness of pre-trained models with CRF in recognizing entities in specific domains with limited resources. Both the dataset and the model serve as valuable baselines for future work in biomedical natural language processing (NLP) in Arabic text. In general, this study provides strong evidence that transfer learning is beneficial for low-resource biomedical NLP, even if the annotated data are scarce. The CAMeLBERT+CRF model robustness to annotation noise indicates that pretrained language models can make up for the lack of data when finetuned with care. However, the performance gap observed among entity types reflects fundamental challenges of semantic ambiguity and morphological complexity which are still at the core of issues in processing Arabic biomedical text.

Future work will focus on enriching the dataset with diverse medical articles to include additional biomedical entity types such as drugs, treatments, and diagnostics from real-world EHR data, as this expansion may affect the model's generalizability. Moreover, more studies are needed explore to nested entities, considering the current limitations of CAMeLBERT+CRF model in this area. Additionally, focusing on other NLP tasks, such as text summarization, question-answering, and relation extraction within Arabic biomedical texts.

References

[1] R. Grishman and B. Sundheim, "Message Understanding Conference-6: A brief history," *Proc. 16th Int. Conf. Computational Linguistics (COLING)*, Copenhagen, Denmark, 1996.

[2] Z. Luo, Y. Ji, A. Gupta, Z. Li, A. Frisch, and D. He, "Towards accurate and clinically meaningful summarization of electronic health record notes: A guided approach," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informatics (BHI)*, Pittsburgh, PA, USA, 2023, pp. 1–5.

[3] N. Goyal and N. Singh, "Named entity recognition and relationship extraction for biomedical text: A comprehensive survey, recent advancements, and future research directions," *Neurocomputing*, vol. 618, p. 129171, 2025.

[4] K. Peng, C. Yin, W. Rong, C. Lin, D. Zhou, and Z. Xiong, "Named entity aware transfer learning for biomedical factoid question answering," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 19, no. 4, pp. 2365–2376, 2022.

[5] E. Hossain, R. Rana, N. Higgins, J. Soar, P. D. Barua, A. R. Pisani, and K. Turner, "Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review," *Comput. Biol. Med.*, vol. 155, p. 106649, 2023.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[7] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn. (ICML)*, San Francisco, CA, USA, 2001, pp. 282–289.

[8] Salah, R., Mukred, M., Zakaria, L. Q., Ahmed, R., and Sari, H., "A new rule-based approach for classical Arabic in natural language processing," 2022.

[9] R. Salah, M. Mukred, L. Q. Zakaria, and F. A. M. Al-Yarimi, "A machine learning approach for named entity recognition in classical Arabic natural language processing," *KSII Trans. Internet Inf. Syst.*, vol. 18, no. 10, pp. 2895–2919, 2024.

[10] A. Shaker, A. Aldarf, and I. Bessmertny, "Using LSTM and GRU with a new dataset for named entity recognition in the Arabic language," *arXiv preprint arXiv:2304.03399*, 2023.

[11] A. B. Sadallah, O. Ahmed, S. Mohamed, O. Hatem, D. Hesham, and A. H. Yousef, "ANER: Arabic and Arabizi named entity recognition using transformer-based approach," in *Proc. 2023 Intell. Methods, Syst., and Appl. (IMSA)*, 2023, pp. 263–268.

[12] Cui, X., Song, C., Li, D., Qu, X., Long, J., Yang, Y., and Zhang, H., "RoBGP: A Chinese Nested Biomedical Named Entity Recognition Model Based on RoBERTa and Global Pointer," *Computers, Materials & Continua*, vol. 78, no. 3, pp. 3603–3618, 2024.

[13] J. Li, S. Zhao, J. Yang, Z. Huang, B. Liu, S. Chen, H. Pan, and Q. Wang, "WCP-RNN: A novel RNN-based approach for Bio-NER in Chinese EMRs," *J. Supercomput.*, vol. 76, no. 3, pp. 1450–1467, 2020.

[14] Y.-L. Chiang, C.-H. Lin, C.-L. Sung, and K.-Y. Su, "Nested named entity recognition for Chinese electronic health records with QA-based sequence labeling," in *Proc. 33rd Conf. Comput. Linguistics and Speech Processing (ROCLING)*, 2021, pp. 18–25.

[15] P. Chen, M. Zhang, X. Yu, and S. Li, "Named entity recognition of Chinese electronic medical records based on a hybrid neural network and medical MC-BERT," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 315, 2022.

[16] G. G. Subies, Á. B. Jiménez, and P. M. Fernández, "ClinText-SP and RigoBERTa Clinical: a new set of open resources for Spanish Clinical NLP," *arXiv preprint arXiv:2503.18594*, 2025.

[17] R. A. A. Jonker, T. Almeida, R. Antunes, J. R. Almeida, and S. Matos, "Multi-head CRF classifier for biomedical multi-class named entity recognition on Spanish clinical notes," *Database*, vol. 2024, p. baae068, 2024.

[18] J. Copara, J. Knafou, N. Naderi, C. Moro, P. Ruch, and D. Teodoro, "Contextualized French language models for biomedical named entity recognition," in *Proc. 6e Conf. Conjointe Journées d'Études Sur La Parole (JEP), Traitement Automatique Des Langues Naturelles (TALN), Rencontre Des Etudiants Chercheurs En Informatique Pour Le Traitement Automatique Des Langues (RÉCITAL). Atelier Défi Fouille De Textes*, 2020, pp. 36–48.

[19] N. Bannour, C. Servan, A. Névéol, and X. Tannier, "A benchmark evaluation of clinical named entity recognition in French," *arXiv preprint arXiv:2403.19726*, 2024.

[20] Frei, J., and Kramer, F., "GERNERMED: An open German medical NER model," *Software Impacts*, vol. 11, p. 100212, 2022.

[21] P. Savaram, S. Tabassum, S. Bandu, and L. B, "Multilingual approaches to named entity recognition," in *Proc. 2024 3rd IEEE Delhi Section Flagship Conf. (DELCON)*, 2024, pp. 1–6.

[22] N. Boudjellal, H. Zhang, A. Khan, A. Ahmad, R. Naseem, J. Shang, L. Dai, and A. Khan, "ABioNER: A BERT-based model for Arabic biomedical named-entity recognition," *Complexity*, vol. 2021, pp. 1–6, 2021.

[23] S. Alanazi, "A named entity recognition system applied to Arabic text in the medical domain," in *Proc.*, 2017.

[24] R. M. Hamad and A. M. Abushaala, "Medical named entity recognition in Arabic text using SVM," in *Proc. 2023 IEEE 3rd Int. Maghreb Meeting Conf. Sci. Technol. Automat. Control Comput. Eng. (MI-STA)*, 2023, pp. 200–205.

[25] I. Ait Talghalit, H. Alami, and S. O. El Alaoui, "Exploring different annotation schemes for single and consecutive named entity recognition in the Arabic biomedical domain using transformer models and contextual semantic embeddings," *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 2, pp. 21854–21860, 2025.

[26] K. Darwish and H. Mubarak, "Farasa: A new fast and accurate Arabic word segmenter," in *Proc. 10th Int. Conf. Language Resources and Evaluation (LREC)*, 2016, pp. 1070–1074.

[27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., "Transformers: State-of-the-art natural language processing," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

[28] Webteb, "Medical Information and Health Articles," Webteb. Accessed: March 4, 2025. [Online]. Available: https://www.webteb.com/

[29] CAMeL-Lab, "BERT Base Arabic CAMeLBert Mix," Hugging Face Accessed: April 12, 2025. [Online]. Available: https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-mix