Lotus at WojoodNER Shared Task: Multilingual Transformers: Unveiling Flat and Nested Entity Recognition

Jiyong Li[1], Dilshod Azizov[2], Hilal AlQuabeh[2], Shangsong Liang[1,2,*]
[1]Sun Yat-sen University
[2]Mohamed bin Zayed University of Artificial Intelligence
lijy373@mail2.sysu.edu.cn,{dilshod.azizov, hilal.alquabeh}@mbzuai.ac.ae
[*]Corresponding author, liangshangsong@gmail.com

Abstract

We introduce our systems developed for two subtasks in the shared task "WOJOOD" on Arabic NER detection, part of ARABICNLP 2023. For Subtask 1, we employ the XLM-R model to predict Flat NER labels for given tokens using a single classifier capable of categorizing all labels. For Subtask 2, we use the XLM-R encoder by building 21 individual classifiers. Each classifier corresponds to a specific label and is designed to determine the presence of its respective label. In terms of performance, our systems achieved competitive *micro-F1* scores of **0.83** for Subtask 1 and **0.76** for Subtask 2, according to the leaderboard scores.

1 Introduction

Named Entity Recognition (NER) is crucial for Natural Language Processing (NLP), enabling the extraction of entities like names and locations from texts. Given the rich linguistic diversity and varied dialects of Arabic, NER becomes especially challenging (Guellil et al., 2021).

Arabic, spoken by 420 million natives, is one of the top ten global languages (Guellil et al., 2021; Cheng et al., 2021; Qu et al., 2023). Its lack of capital letters amplifies its morphological complexity, contrasting NER ease in English due to its varied dialects and rich history.

Arabic can be broadly classified into three distinct forms (Benajiba and Rosso, 2008): Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA) (Elgibali, 2005). Although CA is the esteemed language of most religious Arabic texts, MSA, recognized as an official language by the United Nations, finds its presence in contemporary media, formal correspondences, and the academic sphere. DA, on the contrary, dominates informal day-to-day communications (Qu et al., 2023).

The increasing volume of Arabic content on digital platforms, driven by the proliferation of social media, has led to a surge in the demand for Arabic NER. Beyond its general applications, NER serves specialized domains, enabling tasks such as relation extraction (Cheng et al., 2021), entity linking (Gu et al., 2021), event extraction (Zhu et al., 2021), coreference resolution (Clark and Manning, 2016), and machine translation (Ugawa et al., 2018),

Historically, most Arabic NER research has focused on direct, flat entity recognition techniques. However, the introduction of the Wojood corpus (Jarrar et al., 2022b) marks a pivotal shift. This corpus, which forms the foundation of the ARABICNLP 2023 WOJOOD (Jarrar et al., 2023) shared task, stands out for its extensive reach, encompassing more than 550k tokens from MSA and its respective dialects. All of these are carefully annotated across a spectrum of 21 different entity types.

The shared task (Jarrar et al., 2023) highlights two principal NER challenges:

(i) Wojood-Flat: This traditional method assigns each token to a single well-defined entity type.

(ii) Wojood-Nested: A more complex approach where tokens can be linked to multiple overlapping entity labels, highlighting the intricacies of languages as depicted in Figure 1.

Two significant challenges arise in this context. First, despite progress in NLP, Nested NER (Wang et al., 2022; Jarrar et al., 2022a; Straková et al., 2019) remains relatively uncharted compared to its flat NER counterpart, which has been deeply explored through cutting-edge linguistic, statistical, and neural techniques (Li et al., 2019; Zirikly and Diab, 2015; Shaalan and Raza, 2009). Second, there exists a conspicuous lack of detailed and expansive datasets designed specifically for nested NER. Consequently, addressing the demands of Subtask 2 poses a more significant challenge compared to Subtask 1.

Our paper offers the following contributions:

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 765–770

Figure 1: An example of a nested NER scheme, where some tokens have more than one entity type assigned. Source: www.sina.birzeit.edu

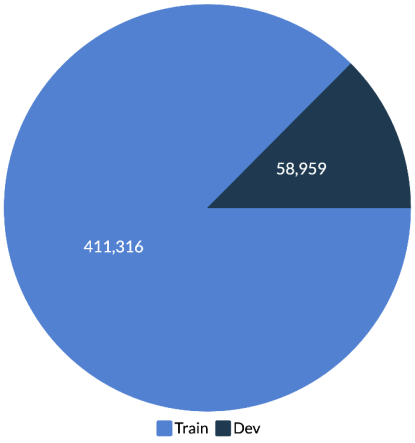

Figure 2: Statistics about tokens distribution in train and development sets.

- We introduce an automated system that uses the XLM-R (Conneau et al., 2019) architecture for both subtasks, but with distinct number of classifiers.
- We discuss and compare the performance of XLM-R, AraBERT(Antoun et al., 2020), and MARBERT (Abdul-Mageed et al., 2020) in our datasets.

The organization of this paper is as follows. In Section 2, we present prior and recent research on arabic NER. Section 3 presents a comprehensive analysis of the dataset. Section 4 describes our proposed system, experimental setup, and results. In Section 5, we conclude and point out ideas for future research.

2 Related Work

Arabic NER has seen notable advancements with various corpora, such as Ontonotes 5, which features 18 entity types from MSA (Weischedel et al., 2011), and others such as ANERCorp, CANER-Corpus, and the expansive AQMAR corpus (Benajiba et al., 2007; Salah and Zakaria, 2018; Mohit et al., 2012). The Wojood corpus stands out, supporting named entities over 21 types and spanning both MSA and dialects (Jarrar et al., 2022b, 2023). Although NER methods have transitioned from rule-based to deep learning, the integration of LSTM-CRF (Lample et al., 2016) and pre-trained embeddings has been transformative. The introduction of BERT highlighted the potential of transformers in NER (Devlin et al., 2018). Nested NER challenges persist, but innovations such as multi-layer BiLSTM and pyramid architectures signal progress (Katiyar and Cardie, 2018; Ju et al., 2018; Wang et al., 2020).

3 Data

In this section, we provide a detailed description of the dataset released by the WojoodNER organizers, which comprises Arabic tokens, where less than half of the dataset consists of named entities.

The Wojood corpus encompasses a comprehensive and diverse array of flat and nested named entities, representing a new split distinct from the established Wojood paper (Jarrar et al., 2022b).

Data Attributes:

Each token in the Wojood corpus is associated with one of the predefined named entities.

- Token: single-word or sub-word unit.
- Entity types: 21 predefined entity types (e.g., location, organization, event).

Data Size:

The set Wojood corpus comprises a significant number of tokens for each named entity. In total, the dataset has around 550k tokens. Figure 2 illustrates the distribution of tokens in the train and development sets. Figure 3 shows the distribution of the named entities in the train and development sets. We observe that the majority, which constitute almost 60% of the dataset, are "Not named entities". The second most common

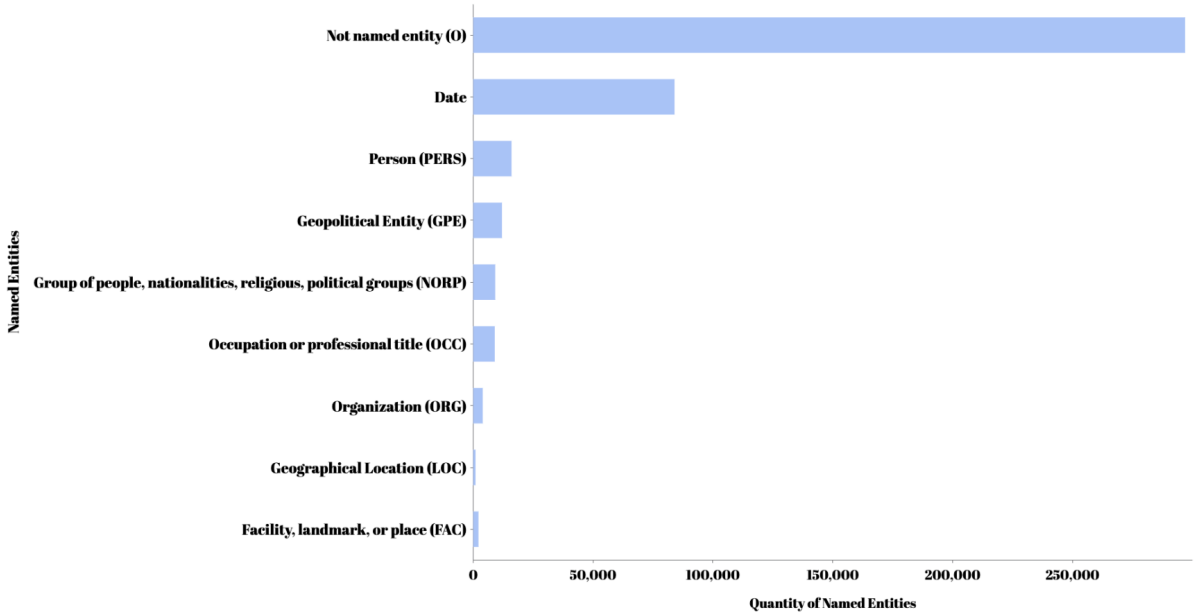

Figure 3: Namedy entities distribution over the train and development from our Wojood corpus.

entity is "Date", followed by "Person (PERS)", "Geographical Entity (GPE)". The least frequent entities in the dataset are "Organization", "Facility, landmark or place (FAC)", and "Geographical Location (LOC)". The rest of the entities named not shown in Figure 3, their numbers are significantly low (e.g., "Currency").

4 Systems Description and Results

4.1 System Description

For evaluation, we use the official evaluation scorers provided for the shared task. The primary measure for both subtasks is the micro-F1 score. However, the scorers also provide data on precision and recall. Our model training was executed on 2 NVIDIA Tesla T4 (16GB) GPU.

Subtask 1. For the Subtask 1 system, we used a configuration with a sequence length of 256 and a batch size of 8. The model was trained for 5 epochs, which is the optimal duration to prevent overfitting and with a learning rate of 2e-5. Measures are captured every 500 steps, and gradients are limited to a maximum norm of 1.0. The ADAMW variant was chosen for optimization. Model checkpoints are saved every 500 steps and at the end of each epoch. We did not employ a warm-up phase, as indicated by both the warm-up ratio and the step count set to zero. To counteract overfitting, we applied a weight decay of 0.01.

Subtask 2. For Subtask 2, we trained the XLM-R model on Wojood corpus with parameters including a batch size of 16 and a learning rate of 1e-5, and the rest of the hyperparameters are similar to Subtask 1.

4.2 Results

During the initial stages, we experimented with AraBERT, MARBERT, and XLM-R with the default parameters. We experimented with the development set, since we used it as a test set, and from the train set we cut 10% out of the total tweets for the development set.

The comparative evaluation of the three frameworks, MARBERT, AraBERT, and XLM-R, on two distinct subtasks showcased varied performances. For Subtask 1, XLM-R emerged as the leading model with the highest micro-F1 score of 0.829, precision of 0.803, and recall of 0.857. AraBERT was followed with a micro-F1 of 0.713, and Precision and Recall values of 0.695 and 0.731, respectively. MARBERT, on the other hand, demonstrated a comparatively lower performance, recording a micro-F1 of 0.563. In Subtask 2, XLM-R maintained its superior performance, achieving the highest micro-F1 of 0.879 and a precision of 0.882. However, in terms of recall, MARBERT led with a score of 0.884. AraBERT showed decent performance with a micro-F1 of 0.848, a precision of 0.826, and a recall of 0.871.

In addition, in Subtask 1, data processing emerged as a critical component, dictating how

	Subtask 1			Subtask 2		
	Micro F1	**Precision**	**Recall**	**Micro F1**	**Precision**	**Recall**
MARBERT	0.663	0.675	0.610	0.870	0.857	**0.884**
AraBERT	0.713	0.695	0.731	0.848	0.826	0.871
XLM-R	**0.829**	**0.803**	**0.857**	**0.879**	**0.882**	0.877

Table 1: Experimental results of MARBERT, AraBERT, and XLM-R on the development sets for *Subtask 1* and *Subtask 2*.

well the subsequent stages would proceed. Meanwhile, in Subtask 2, the structure of the classifier piqued interest. Specifically, a simplistic approach to the nested NER – treating it as a standard classification problem and differentiating "I-XX" and "B-XX" as separate labels – would likely lead to suboptimal results.

Subtask 1. For our Named Entity Recognition (NER) task, we employed a careful preprocessing approach on our Wojood corpus using the AraBERT preprocessor and tokenizer. A key component in this process is ensuring accurate label alignment. To handle the challenge posed by tokenization splitting words into fragments, we introduced a strategy: any token resulting from either padding or representing a fragment of a word is assigned a label of -100. For instance, the word "responding" would be tokenized into two parts: "respond" and "-ing". In this case, "-ing" would be assigned the label -100.

Subsequently, our NER task was framed as a multi-class classification problem. It is important to note that we treat "I-XX" and "B-XX" as separate labels. We used the XLM-R model equipped with a single classifier capable of categorizing all labels. Based on the leaderboard scores, our system achieved a competitive micro-F1 score of 0.83.

Subtask 2. For this subtask, we use official scripts for processing, resulting in slight procedural variations compared to Subtask 1. Similar data processing methods were employed; however, in this case, padding tokens were assigned tags corresponding to the "O" label index. We conceptualized this nested NER task as a two-tier classification. After initial input processing, the system generates 21 different classifiers, each specifically related to a unique label, such as "CARDINAL", etc.

Each of these classifiers has the role of categorizing input tokens into one of three categories: "I-", "B-", or "O". To elucidate with an example: should an input be classified as "B-" by the "CARDINAL" classifier, it would translate into a prediction "B-CARDINAL". The performance measures on the leaderboard indicate that our system achieved a micro-F1 score of 0.76.

5 Conclusion and Future Work

In this paper, we detail our XLM-R based systems for two subtasks in the ARABICNLP 2023 Wojood NER shared task. Subtask 1 utilized a single classifier, while Subtask 2 developed 21 label-specific classifiers. Our models achieved micro-F1 scores of **0.83** and **0.76** for Subtasks 1 and 2, respectively, according to official leaderboard scores. We also compared our systems with state-of-the-art AraBERT and MARBERT models.

In future work, we plan to incorporate data augmentation methods, including sentence mixing and back-translation. Additionally, we would adopt a Meta-Ensembling approach, integrating models such as AraBERT, MARBERT, XLM-R, and ARBERT, to enhance performance on the unique and diverse Wojood corpus.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Yassine Benajiba and Paolo Rosso. 2008. Arabic named entity recognition using conditional random fields. In *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*, volume 8, pages 143–153.

Yassine Benajiba, Paolo Rosso, and José Miguel Benedíruiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In *Computational Linguistics and Intelligent Text Processing: 8th International Conference, CICLing 2007, Mexico City, Mexico, February 18-24, 2007. Proceedings 8*, pages 143–153. Springer.

Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. 2021. Hacred: A large-scale relation extraction dataset toward hard cases in practical applications. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2819–2831.

Kevin Clark and Christopher D Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. *arXiv preprint arXiv:1606.01323*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alaa Elgibali. 2005. *Investigating Arabic: Current parameters in analysis and learning*, volume 42. Brill.

Yingjie Gu, Xiaoye Qu, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Xiaolin Gui. 2021. Read, retrospect, select: An mrc framework to short text entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12920–12928.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5):497–507.

Mustafa Jarrar, Muhammad Abdul-Mageed, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, Nagham Hamad, and Alaa Omer. 2023. WojoodNER: The Arabic Named Entity Recognition Shared Task. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022a. Named entity recognition, multi-task learning, nested entities, bert, arabic ner corpus. *arXiv preprint arXiv:2205.09651*.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022b. Wojood: Nested arabic named entity corpus and recognition using bert. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3626–3636.

Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A neural layered model for nested named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1446–1459.

Arzoo Katiyar and Claire Cardie. 2018. Nested named entity recognition revisited. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2019. A unified mrc framework for named entity recognition. *arXiv preprint arXiv:1910.11476*.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A Smith. 2012. Recall-oriented learning of named entities in arabic wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173.

Xiaoye Qu, Yingjie Gu, Qingrong Xia, Zechang Li, Zhefeng Wang, and Baoxing Huai. 2023. A survey on arabic named entity recognition: Past, recent advances, and future trends. *arXiv preprint arXiv:2302.03512*.

Ramzi Esmail Salah and Lailatul Qadri Binti Zakaria. 2018. Building the classical arabic named entity recognition corpus (canercorpus). In *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pages 1–8. IEEE.

Khaled Shaalan and Hafsa Raza. 2009. Nera: Named entity recognition for arabic. *Journal of the American Society for Information Science and Technology*, 60(8):1652–1663.

Jana Straková, Milan Straka, and Jan Hajič. 2019. Neural architectures for nested ner through linearization. *arXiv preprint arXiv:1908.06926*.

Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural machine translation incorporating named entity. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3240–3250.

Yu Wang, Hanghang Tong, Ziye Zhu, and Yun Li. 2022. Nested named entity recognition: a survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–29.

Yuesong Wang, Tao Guan, Zhuo Chen, Yawei Luo, Keyang Luo, and Lili Ju. 2020. Mesh-guided multi-view stereo with pyramid architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2039–2048.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0.

LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium.

Tong Zhu, Xiaoye Qu, Wenliang Chen, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Min Zhang. 2021. Efficient document-level event extraction via pseudo-trigger-aware pruned complete graph. *arXiv preprint arXiv:2112.06013*.

Ayah Zirikly and Mona Diab. 2015. Named entity recognition for arabic social media. In *Proceedings of the 1st workshop on vector space modeling for natural language processing*, pages 176–185.