WojoodNER 2023: The First Arabic Named Entity Recognition Shared Task

Mustafa Jarrar[1] **Muhammad Abdul-Mageed**[2,3] **Mohammed Khalilia**[1] **Bashar Talafha**[2]
AbdelRahim Elmadany[2] **Nagham Hamad**[1] **Alaa' Omar**[1]
[1]Birzeit University, Palestine
[2]Deep Learning & Natural Language Processing Group, The University of British Columbia
[3]Department of Natural Language Processing & Department of Machine Learning, MBZUAI
mjarrar@birzeit.edu muhammad.mageed@ubc.ca

Abstract

We present WojoodNER-2023, the first Arabic Named Entity Recognition (NER) Shared Task. The primary focus of WojoodNER 2023 is on Arabic NER, offering novel NER datasets (i.e., Wojood) and the definition of subtasks designed to facilitate meaningful comparisons between different NER approaches. WojoodNER-2023 encompassed two Subtasks: FlatNER and NestedNER. A total of 45 unique teams registered for this shared task, with 11 of them actively participating in the test phase. Specifically, 11 teams participated in FlatNER, while 8 teams tackled NestedNER. The winning teams achieved F_1 scores of 91.96 and 93.73 in FlatNER and NestedNER, respectively.

1 Introduction

NER is a fundamental task in Natural Language Processing (NLP), especially in information extraction and language understanding (Jarrar et al., 2023a). The objective of NER is to identify and classify named entities in a given text into predefined categories, such as "*person*", "*location*", "*organization*", "*event*", and "*occupation*". NER is also a critical task for many NLP applications, such as question-answering systems (Shaheen and Ezzeldin, 2014), knowledge graphs (James, 1991), and semantic search (Guha et al., 2003), interoperability (Jarrar et al., 2011) among others. Named entities can either be flat or nested. For instance, in the sentence "`Cairo Bank announces its profit in 2023`", there are two flat entities: "`Cairo Bank`" is tagged as `ORG` (i.e., organization) and "`2023`" as `DATE`. In nested NER, entity mentions contained inside other entity mentions are also considered named entities. In this case, "`Cairo`", is tagged as `GPE` (i.e., geopolitical entity). Section 3 illustrates more examples. As will be discussed in Section 2, research in Arabic NER is currently limited, particularly in the context of nested entities. This limitation is not exclusive to Modern Standard Arabic (MSA) but extends to various Arabic dialects across diverse domains and NER subtypes. The majority of existing research on Arabic NER primarily emphasizes flat entities to cover a limited set of entity types, mainly "person", "organization", and "location".

Figure 1: Topics in the Wojood NER corpus.

In this paper, we provide an overview of the WojoodNER-2023 Shared Task[1], which represents a significant step forward in advancing NER research in the Arabic language. The shared task encompasses subtask1 (FlatNER) and subtask2 (NestedNER). For this competition, we grant participants access to the Wojood corpus (Jarrar et al., 2022)[2], a substantial and diverse Arabic NER dataset known as Wojood. As shown in Figure 1, Wojood is particularly notable for its scale, containing approximately 550K tokens. About 12% of the corpus was collected from social media in *Palestinian* and *Lebanese* dialects Curras and Baladi corpora (Haff et al., 2022). The remaining $\sim$ 88% is in MSA, covering multiple domains, including

[1]SharedTask Call: https://dlnlp.ai/st/wojood/
[2]Wojood Corpus: https://sina.birzeit.edu/wojood/

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 748–758

health, *finance*, *politics*, *ICT*, *terrorism*, *migration*, *history and culture*, and *law and elections*, making it a rich resource for various research purposes. Wojood was annotated manually using 21 entity types, offering a rich Arabic NER corpus.

The primary objective of this shared task is to encourage participants to explore different NER methodologies. Teams were invited to experiment with various approaches, ranging from classical machine learning to advanced deep learning and transformer-based techniques, among others. The shared task generated a remarkably diverse array of submissions. A total of 45 teams registered to participate in the shared task. Among these, 11 teams successfully submitted their models for evaluation on the blind test set during the final phase of the competition. As a result, we received 11 papers that provide detailed insights into the results achieved by these teams for either one or both of the subtasks.

The rest of the paper is organized as follows: Section 2 provides a brief overview of Arabic NER. We describe the two subtasks and WojoodNER-2023 restrictions in Section 3. Section 4 introduces shared task datasets and evaluation setup. We present participating teams and shared task results and provide a high-level description of submitted systems in Section 5. We conclude in Section 6.

2 Literature Review

NER has been a long-standing research area, with significant advances made in recent years. As will be discussed in this section, early NER approaches focused on identifying and classifying flat named entities, and recent research has focused on nested NER. In this section, our primary focus is exclusively on Arabic NER research, encompassing corpora, methodologies and shared tasks.

Corpora. Most of the available Arabic NER corpora are annotated as flat NER. `ANERCorp` (Benajiba et al., 2007), sourced from the news domain (MSA text), comprises $\sim$ 150k tokens. Its main emphasis is directed towards four distinct entity types. `CANERCorpus` (Salah and Zakaria, 2018) is dedicated to Classical Arabic (CA) and encompasses a dataset of 258K tokens. This corpus is annotated for a total of 14 entity types, all of which pertain to religious entities. `ACE2005` (Walker et al., 2005) is a multilingual corpus that incorporates Arabic text encompassing *five* distinct types of entities. `Ontonotes5` (Weischedel et al., 2013) dataset consists of approximately 300K tokens, meticulously annotated with 18 distinct entity types. Nevertheless, these corpora were collected a long time ago and mainly cover the media and politics domains; hence, may not be representative of the current state of Arabic language use. This is especially the case since language models are known to be sensitive to temporal and domain shifts. Recently, Jarrar et al. (2022) proposed `Wojood`, the largest Arabic NER corpus. It is distinctive for its support of both flat and nested entity annotations, making it a crucial resource utilized in this shared task. It comprises roughly 550K tokens encompassing a diverse range of 21 unique entity types, spanning both MSA and two dialectal Arabic forms (the Palestinian Curras2 and Lebanese Baladi corpora (Haff et al., 2022).

Methodologies. Various studies explore Arabic NER by employing various approaches, with some researchers focusing on rule-based (Shaalan and Raza, 2007; Jaber and Zaraket, 2017) and machine learning (Settles, 2004; Abdul-Hamid and Darwish, 2010; Zirikly and Diab, 2014; Dahan et al., 2015; Darwish et al., 2021) strategies. Recent researches embrace deep learning methodologies including character and word embeddings with Long-Short Term Memory (LSTM) networks (Ali et al., 2018), BiLSTM followed by Conditional Random Field (CRF) models (El Bazi and Laachfoubi, 2019; Khalifa and Shaalan, 2019), Deep Neural Networks (DNN) (Gridach, 2018), and pretrained Language Models (LM) (Jarrar et al., 2022; Liqreina et al., 2023). Wang et al. (2022) proposed a survey that extensively explores different approaches to nested entity recognition, encompassing rule-based, layered-based, region-based, hypergraph-based, and transition-based methodologies. Fei et al. (2020) proposed a multitask learning approach for nested NER that employs a dispatched attention model. Ouchi et al. (2020) proposed an approach for nested NER that involves enumerating all region representations from the contextual encoding sequence and then assigning a category label to each of them.

Shared tasks. While there are multiple shared tasks for NER in various languages and domains, such as the `MultiCoNER` for multilingual complex NER (Malmasi et al., 2022), the `HIPE-2022` for NER and linking in multilingual historical documents (Ehrmann et al., 2022), the `RuNNE-2022` for nested NER in Russian (Artemova et al., 2022),

and the NLPCC2022 for extracting entities in the material science domain (Cai et al., 2022). To the best of our knowledge, there has been no dedicated shared task for Arabic NER. Therefore, we initiate this shared task with the aim of being the inaugural event in this specific domain.

3 Task Description

To the best of our knowledge, WojoodNER-2023 is recognized as the inaugural shared task in Arabic NER. In this competition, we present two distinct subtasks—one for "**FlatNER**" and the other for "**NestedNER**". These subtasks are of paramount importance in addressing the challenges inherent in Arabic NER processing. We now describe each subtask in detail.

3.1 Subtask1 – FlatNER

In FlatNER, each token in the data is labeled with only one tag. The participants in this subtask are expected to develop models to classify each token as a multi-class classification problem. An example of the FlatNER data is shown in Figure 2. The Wojood annotation guidelines were designed for nested entities only, therefore, the flat entities were derived from the nested entities by taking the top-level entity mentions (i.e., topmost tags).

Figure 2: Flat NER example

3.2 Subtask2 – NestedNER

In the NestedNER subtask, each token can have one or more tags. In this data, we will find entity mentions inside other entity mentions as demonstrated in Figure 3. For instance, the phrase "مؤسسة إدوارد سعيد" is annotated as ORG, which is the same as the flat annotation in Figure 2. However, in nested NER, it contains another entity mention "إدوارد سعيد" tagged with PERS.

Figure 3: Nested NER example

3.3 Restrictions

This section outlines the stipulations and directives that govern participants' engagement in the WojoodNER 2023 Shared Task. These regulatory directives and guidelines establish an equitable competitive environment for all participants, ensuring transparency and impartiality throughout the duration of the WojoodNER 2023 Shared Task. They also ensure the credibility of the task's assessment procedure, which was published on the shared task official website frequently asked question page.

External data. Participants are strictly prohibited from using external data from previously labeled datasets or employing taggers that have been previously trained to predict named entities. The use of any resources with prior knowledge related to NER is not allowed.

Data format constraints. The submission to the task consists of one file containing the model prediction in CoNLL format. The CoNLL format should include multiple columns space-separated. The first column is reserved for the tokens, while all subsequent columns are used for the tags. In the case of nested NER, the tag columns have a predefined order, which we specified on the shared task webpage[3]. The IOB2 (Sang and Veenstra, 1999) scheme is used for the submission, which is the same format used in the Wojood dataset. Finally, text segments are separated by a blank line.

Pretrained models. The participants are allowed to utilize pretrained transformer models such as "*BERT*" (Devlin et al., 2018) and word representations like "*Word2Vec*" (Church, 2017) and "*ELMo*" (Peters et al., 2018) for the purpose of transfer learning. It is worth noting that our baseline model is based on BERT.

Linguistic features. When considering the incorporation of linguistic features to enhance the dataset, participants are permitted to include part-of-speech tagging and syntactic layers within their code.

4 Shared Task Datasets and Evaluation

This section presents the dataset, evaluation metrics, and the submission process.

Datasets. WojoodNER-2023 shared task employs the Wojood corpus as its primary dataset (Jarrar et al., 2022). The Wojood corpus encompasses approximately 550K tokens, spanning both MSA and two Arabic dialects, annotated using 21 entity

[3] https://dlnlp.ai/st/wojood/

Entity Name	NER Tag	FlatNER				NestedNER			
		TRAIN	DEV	TEST	Total	TRAIN	DEV	TEST	Total
Person	PERS	4,496	650	1,409	**6,555**	4,994	730	1,562	**7,286**
Group of people	NORP	3,505	488	948	**4,941**	3747	520	1006	**5273**
Occupation	OCC	3,774	544	1,058	**5,376**	3,887	551	1,95	**5,533**
Organization	ORG	10,731	1,566	3,047	**15,344**	13,174	1,869	3,738	**18,781**
GeoPolitical Entity	GPE	8,133	1,132	2,281	**11,546**	15,300	2,163	4,315	**21,778**
Geographical location	LOC	510	63	168	**741**	619	76	204	**899**
Facility (e.g., landmarks)	FAC	689	85	165	**939**	880	111	224	**1,215**
Product	PRODUCT	36	5	13	**54**	36	5	14	**55**
Event	EVENT	1,863	253	556	**2,672**	1,934	267	577	**2,778**
Date	DATE	10,667	1,567	3,091	**15,325**	11,290	1,656	3,288	**1,6234**
Time	TIME	286	55	84	**425**	288	55	84	**427**
Language	LANGUAGE	131	15	51	**197**	132	15	51	**198**
Website	WEBSITE	434	45	128	**607**	434	45	128	**607**
Law	LAW	374	44	78	**496**	374	44	78	**496**
Cardinal	CARDINAL	1,245	182	360	**1,787**	1,263	183	363	**1,809**
Ordinal	ORDINAL	2,805	410	858	**4,073**	3,488	504	1,070	**5,062**
Percent	PERCENT	105	13	19	**137**	105	13	19	**137**
Quantity	QUANTITY	44	3	7	**54**	46	3	8	**57**
Unit	UNIT	7	0	2	**9**	48	3	9	**60**
Money	MONEY	171	20	36	**227**	171	20	36	**227**
Currency	CURR	19	1	5	**25**	179	21	41	**241**
	Total	**50,025**	**7,141**	**14,364**	**71,530**	**62,389**	**8,854**	**17,910**	**89,153**

Table 1: Distribution of NER tags in WojoodNER-2023 Subtask1 (i.e., FlatNER) and Subtask2 (i.e., NestedNER) across the training (i.e., TRAIN) , development (i.e., DEV), and test (i.e., TEST) splits for the WojoodNER-2023.

types. Wojood annotation guidelines are optimized for nested Arabic NER annotations. However, for the purposes of the shared task, we generate a flat NER dataset by reducing the nested NER annotation to the top level only as demonstrated in Figure 2 and 3. For both subtasks, we split the data 70/10/20 for training, development, and test dataset respectively at the domain level. This split ensures similar data distribution across the three datasets. Table 1 present the statistics and characteristics of WojoodNER-2023's subtask1 and subtask2 training, development, and test datasets.

Evaluation metrics. The official evaluation metric for subtask1 and subtask2 is the macro-averaged F_1 score. In addition to this metric, we also report system performance in terms of `Precision`, `Recall`, and `Accuracy` for submissions to both subtasks.

Submission roles. We allowed participant teams to submit up to *four* runs for each test set, for both subtasks. In each one, we strictly retain only the submission with the highest score from each participating team. Although the official results were solely derived from the blind test set. To streamline the evaluation of participant systems, we have set up two separate CodaLab (Pavao et al., 2023) competitions for scoring each subtask.[4] We are keeping the CodaLab (Pavao et al., 2023) for each subtask active even after the official competition has concluded. This is aimed at facilitating researchers who wish to continue training models and evaluating systems with the shared task's blind test sets. As a result, we will not disclose the labels for the test sets in any of the subtasks.

5 Shared Task Teams & Results

5.1 Participating Teams

In total, we received 45 unique team registrations. At the testing phase, a total of 57 valid entries were submitted by 12 unique teams. We received 35 submissions for FlatNER from *eleven* teams and 22 submissions for NestedNER from *eight* teams. Table 2 lists the teams, their affiliation, and the tasks they participated in (Subtask1 – FlatNER and Subtask2 – NestedNER). From 12 teams we received 11 description papers from which we accepted 8 for publication and 3 were rejected (for quality or not adhering to the shared task guidelines).

[4]The different CodaLab competitions are available at the following links: `subtask-1` and `subtask-2`.

Team	Affiliation	Task
Alex-U 2023 NLP (Hussein et al., 2023)	Alexandria University	1,2
AlexU-AIC (Elkordi et al., 2023)	Alexandria University	1,2
AlphaBrains (Ehsan et al., 2023)	University of Gujrat, Pakistan	1,2
ARATAL	IPSA	1
El-Kawaref (Elkaref and Elkaref, 2023)	German University in Cairo	1
ELYADATA (Laouirine et al., 2023)	ELYADATA	1,2
Fraunhofer IAIS	Fraunhofer IAIS	1
LIPN (El Khbir et al., 2023)	LIPN, Université Paris 13	1,2
Lotus (Li et al., 2023)	MBZUAI	1,2
R00	Jordan University of Science and Technology	1,2
Think NER	Ulm University	1,2
UM6P & UL (El Mahdaouy et al., 2023)	Mohammed VI Polytechnic University	1,2

Table 2: List of teams that participated in either one or both subtasks. Teams with accepted papers are cited.

5.2 Baselines

For both subtasks, we fine-tune the AraBERT$_{v2}$ (Antoun et al., 2020) and ARBERT$_{v2}$ (Abdul-Mageed et al., 2021) pre-trained models using the training data that is specific to each subtask for 20 epochs and employed a learning rate of $1e-5$, along with a batch size of 16. To ensure model optimization, we incorporate early stopping with a patience setting of 5. After each epoch, we evaluated the model's performance and selected the best-performing checkpoints based on their performance on the respective development set. Subsequently, we present the performance metrics of the best-performing model on the test datasets.

Rank	Team	F1	Pre.	Rec.
1	LIPN	**91.96**	92.56	91.36
2	El-Kawaref	91.95	91.43	92.48
3	ELYADATA	91.92	91.88	91.96
4	Alex-U 2023 NLP	91.80	91.61	92.00
5	Think NER	91.25	90.76	91.73
6	ARATAL	91.13	90.49	91.77
7	UM6P & UL	91.13	90.70	91.57
8	AlexU-AIC	91.13	91.33	90.92
	Baseline-I (ARBERT$_{v2}$)	89.20	88.32	90.09
	Baseline-II (AraBERT$_{v2}$)	87.33	86.00	88.00
9	AlphaBrains	87.15	87.45	87.58
10	Lotus	83.39	80.90	86.04
11	R00	76.99	76.67	77.31
12	Fraunhofer IAIS	64.45	65.53	63.40

Table 3: Results of Subtask1 – FlatNER.

5.3 Results

Table 3 and Table 4 present the leaderboards of Subtask1 – FlatNER and Subtask2 – NestedNER, respectively, sorted by macro-F_1 in descending order. The macro-F_1 score for each team represents the highest score among the four allowed submissions for each task.

Rank	Team	F1	Pre.	Rec.
1	Elyadata	93.73	93.99	93.48
2	UM6P & UL	93.03	92.46	93.61
3	AlexU-AIC	92.61	92.10	93.13
4	LIPN	92.45	92.31	92.59
	Baseline-I (ArBERT$_{v2}$)	91.68	91.01	92.35
5	Think NER	91.4	90.03	92.82
	Baseline-II (AraBERT$_{v2}$)	91.06	90.74	91.38
6	Alex-U 2023 NLP	90.01	89.39	90.63
7	AlphaBrains	88.84	88.45	89.23
8	Lotus	76.02	82.19	70.72

Table 4: Results of Subtask2 – NestedNER.

For FlatNER, `LIPN` team (El Khbir et al., 2023) achieved the highest F_1 score of 91.96, while `El-Kawaref` (Elkaref and Elkaref, 2023) came in second place with 91.95 and `Elyadata` in third place with 91.92. Notably, on FlatNER, *eight* teams surpass our two baselines performance, as seen in Table 3. Moreover, the winning team (i.e, `LIPN` (El Khbir et al., 2023)) outperforms the Baseline-I by 2.76%. *Three* teams underperform Baseline-I and Baseline-II. However, the gap between the baseline-I and the worst-performing model is about 24.75%. We also notice that the difference in the F_1 score among the top *eight* teams is marginal ($\sigma = 0.41$).

We also analyzed the performance at the entity-type level in FlatNER and we noticed that certain entity types are more challenging to learn by all submitted models, including the baseline. The main reason for their low performance is the rarity of those entities in the dataset, with frequency reaching as low as 9 for `UNIT` and 54 for both `PRODUCT`

Team Name	F_1	Preprocessing	Features			Techniques						
			TF-IDF	Word Embeds	Resampling	Neural Nets	Contrast. L	Ensemble	Adapter	Multitask	PLM	Hie. Cls
FlatNER												
LIPN	91.96					✓		✓			✓	
El-Kawaref	91.95	✓									✓	
Elyadata	91.92	✓			✓	✓					✓	
Alex-U 2023 NLP	91.80						✓				✓	
ThinkNER	91.25											
UM6P & UL	91.13									✓	✓	
AlexU-AIC	91.13	✓									✓	✓
ARATAL	91.13							✓			✓	
AlphaBrains	87.51			✓		✓				✓		
Lotus	83.39	✓	✓						✓		✓	
Fraunhofer IAIS	64.45										✓	
NestedNER												
Elyadata	93.73	✓			✓	✓					✓	
UM6P & UL	93.03									✓	✓	
AlexU-AIC	92.61	✓									✓	✓
LIPN	92.45					✓		✓			✓	
ThinkNER	91.40											
Alex-U 2023 NLP	76.02						✓				✓	
AlphaBrains	88.84			✓		✓				✓		
Lotus	76.02	✓	✓						✓		✓	

Table 5: Summary of approaches used by participating teams in subtask1 (i.e., FlatNER) and subtask2 (i.e., NestedNER). Teams are sorted by their performance on the official metric, Macro-F_1 score. The term "Neural Nets" refers to any model based on neural networks (e.g., FFNN, RNN, CNN, and Transformer) trained from scratch. PLM refers to neural networks pretrained with unlabeled data such as ARBERT$_{v2}$. (Hie. Cls, hierarchical classification approach); (Contrast. L, contrastive learning).

and QUANTITY. The highest F_1 for PRODUCT is 61.54 (Hussein et al., 2023), for QUANTITY 50.00 (Elkaref and Elkaref, 2023) and for UNIT 50.00 (Elkaref and Elkaref, 2023; Hussein et al., 2023; Laouirine et al., 2023). CURR also achieved low performance among all participants ($F_1 \leq 66.67$) with exception to (Elkaref and Elkaref, 2023), which reported an $F_1 = 88.89$, despite its low frequency in the data of 25 occurrences. Our Baseline-II achieved low performance on the three entities mentioned above, but outperformed all submitted models on QUANTITY with an $F_1 = 75.00$.

For NestedNER, the ELYADATA team (Laouirine et al., 2023) ranks in the first position with an F_1 score of 93.73, followed by UM6P & UL team (El Mahdaouy et al., 2023) with a score of 93.09 and in third place AlexU-AIC with a score of 92.61. Notably, there are *four* teams that outperform baseline-I with F_1 score gap between the baseline and the best model of 2.05%. Whereas, the gap between baseline-I and the worst-performing model is about 15.66%. The difference in the F_1 score among the top four teams is $\sigma = 0.57$.

The performance at the entity level for NestedNER is analyzed to explain the challenge for all submitted models. As previously mentioned, the scarcity of some entities in the dataset influences the performance of some entity types in FlatNER. This scarcity influences the results on NestedNER, too. The product, quantity, and website obtained the lowest performance in all models. The highest performance for the product is 66.67% which is obtained by ThinkNER team. For the quantity, the 63.16% F1-score is obtained by (El Mahdaouy et al., 2023). For website, the best performance is 69.26% F1-score. The unit entity also achieved a low performance among all teams except (Elkordi et al., 2023) which obtained 80% F1-score.

The final observation we will highlight is the pattern of scores across the two subtasks, where all scores (micro-F1, precision, and recall) are higher in NestedtNER compared to FlatNER. This was also observed in the baseline (Jarrar et al., 2022).

It may seem counter-intuitive, but in fact, FlatNER is harder than NestedtNER. Recall that the Wojood annotation guideline was optimized for nested NER and the flat annotations are simply the top-level tags found in the nested annotations. This conversion from nested to flat annotations caused some tokens to have conflicting tags in the dataset, which breaks the high annotation consistency found in the nested dataset. Another reason for this pattern is the co-occurrence among nested tags. For instance, an entity mention tagged with OCC is more likely to have nested entity mentions tagged as ORG or PERS, rather than entity mentions tagged with PRODUCT, EVENT or DATE.

5.4 General Description of Submitted Systems

All the models submitted to the shared task adopt the transfer learning approach, leveraging pre-trained models trained on various data sources. Generally, we observe that the top-performing models addressed the challenge of identifying nested entities of the same type, a limitation described by Jarrar et al. (2022).

Table 5 summarizes the techniques employed by the participating teams in the WojoodNER-2023 shared task. The common theme is the use of pre-trained models by all participants. The choice of models include AraBERT (Antoun et al., 2020), MARBERT (Abdul-Mageed et al., 2021), ARBERT (Abdul-Mageed et al., 2021), XLM-R (Conneau et al., 2019), and CAMelBERT (Inoue et al., 2021). AraBART$_{v2}$ is the pre-trained language model used the most in the shared task, where it was utilized by seven teams in FlatNER and five teams in NestedNER. MARBERT comes in second place in terms of usage, where six teams used it in both subtasks (Figure 4).

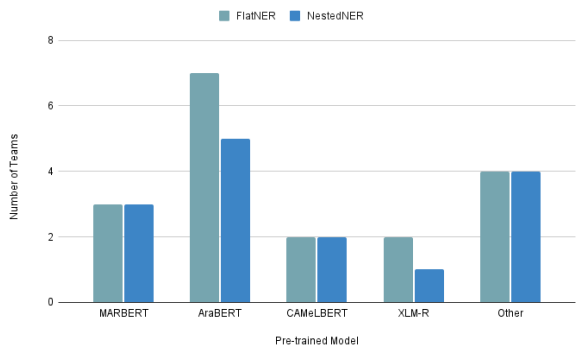

Figure 4: Distribution of pre-trained models across teams.

It was observed in the submissions that compare AraBERT with MARBERT and CAMeLBERT that the AraBERT transformer consistently outperformed the others. This is noteworthy, especially considering that AraBERT is pre-trained solely on MSA data and has a smaller size than both MARBERT and CAMeLBERT.

Other transformer-based pre-trained models were also utilized. For instance, `Elyadata` fine-tuned BioBERT (Lee et al., 2020), but the results were much worse than the baseline, which is expected since BioBERT is trained on English biomedical corpus. In a comparative study, the `UM6P & UL` (El Mahdaouy et al., 2023) team explored the capabilities of QARiB (Abdelali et al., 2021), a model pre-trained specifically on Arabic tweets, against ARBERT$_{v2}$ (Abdul-Mageed et al., 2021), which is trained on an expansive and diverse Arabic datasets. Their finding shows ARBERT$_{v2}$'s superiority over other models. The rest of this section will discuss the systems submitted by each team in more details.

We start by `LIPN` (El Khbir et al., 2023) team, who relies on converting the task from sequence labeling to span classification task. Their approach classifies all possible spans within a sequence. For FlatNER, they employ a two-step decoding process: 1) non-entity spans are filtered out, and 2) for the remaining spans, a maximum independent set algorithm is employed to get the optimal set of entity spans. This fusion of algorithmic techniques with machine learning, coupled with the task's reformation, achieved state-of-the-art results for FlatNER and enabled the `LIPN` (El Khbir et al., 2023) team to secure first place in FlatNER and fourth place in the NestedNER.

`UM6P & UL` (El Mahdaouy et al., 2023) utilized multi-task learning similar to (Jarrar et al., 2022). The sequence is encoded using a transformer encoder and each entity type has one multi-class classification head to predict the IOB2 tag for each token. The model is trained with multiple objectives including cross-entropy loss, dice loss to handle class imbalance, Tversky loss to balance false positives and false negatives, and focal loss to down-weight easy examples. All four objectives are combined as a weighted sum, the authors refer to the unified loss. Additionally, the authors used variance penalty loss that computes the variance across all task losses. The authors experimented with different loss configurations and pre-trained

models, using the unified loss and variance loss with ARBERT$_{v2}$ provided the best performance, ranking the team seventh in FlastNER and second in NestedNER.

ELYADATA (Laouirine et al., 2023) team developed the best-performing NestedNER system. They reformulated the task as a denoising problem. DiffusionNER model architecture (Shen et al., 2023) is used with AraBERT, which introduces noise spans to the gold entity boundaries and is trained to reconstruct the entity boundaries. During the inference phase, it picks noisy spans from a standard Gaussian distribution and then produces named entities by leveraging the learned reverse diffusion process. This novel approach enabled the ELYADATA (Laouirine et al., 2023) team to get first place and achieve state-of-the-art outcomes in NestedNER.

AlexU-AIC (Elkordi et al., 2023) technique relies on machine reading comprehension. In their approach, they formulate a query for each entity type, totaling 21 queries, one for each entity type. Based on the query, the model extracts the answer span from the sequence. Their architecture consists of a transformer encoder followed by two binary classifiers, one classifies if the token is the start of the answer span and another classifies if the token is the end of the answer span. The authors also adopted the stochastic weight averaging technique, in which they average the weights of the four best-performing checkpoints. The team is ranked eighth in FlatNER and third in NestedNER.

AlphaBrains (Ehsan et al., 2023) developed a multi-task learning technique that is similar to (Jarrar et al., 2022), but it employes BiLSTM encoder instead of a transformer. The input to the BiLSTM is a concatenation of learned word embeddings and ELMo representations. The team is ranked ninth in FlatNER and seventh in NestedNER.

El-Kawaref (Elkaref and Elkaref, 2023) proposes StagedNER for FlatNER. In the first stage, the transformer encoder is fine-tuned based IOB2 classification task. In that stage, the authors also used part-of-speech (POS) tagging to improve model performance. The second stage also fine-tunes the transformer encoder on entity type classification task and it takes IOB2 tags as an additional input. During training the authors use the ground truth IOB2 tags and in inference, they use the predicted tags. The team is ranked second in FlatNER.

Alex-U 2023 NLP (Hussein et al., 2023) developed AraBINDER. The approach relies on a contrastive learning objective, where the goal is to maximize the similarity between the entity mention span and its entity type and minimize the similarity with the negative classes. To do that, the authors use a bi-encoder, one for encoding the named entity type and another for encoding the named entity mention. The team is ranked fourth in FlatNER and sixth in NestedNER.

Lotus (Li et al., 2023) proposes a model also inspired by (Jarrar et al., 2022). Their model is based on XLM-R with 21 classification heads, one classifier for each entity type and each classifier is a multi-class that outputs one of the IOB2 tags. The team is ranked tenth in the FlatNER and eighth in the NestedNER.

6 Conclusion and Future Work

In this paper, we present the outcomes of WojoodNER-2023, the inaugural shared task dedicated to both flat and nested NER challenges in the Arabic language. The results obtained from the participating teams underscore the persistent challenges associated with NER. However, it is promising to observe that various innovative approaches, often harnessing the capabilities of language models, have demonstrated their effectiveness in addressing this complex task. As we move forward, we remain committed to further advancing research in this domain. Our vision includes ongoing efforts to enhance the field of Arabic NER, incorporating the valuable insights gained from WojoodNER-2023 and continuing to explore innovative solutions. We plan to extend the Wojood corpus to include more dialects. We plan to include the Syrian Nabra dialects (Nayouf et al., 2023) as well as the four dialects in the Lisan (Jarrar et al., 2023b) corpus.

Acknowledgment

We would like to thank Sana Ghanem for helping us with data annotations, and Tymaa Hammouda for her technical support during the organization of the task.

7 Limitations

While our aim was to achieve the broadest possible coverage, it is essential to acknowledge that WojoodNER-2023 primarily concentrated on MSA data, with only a limited representation of dialects,

specifically covering two dialects, Palestinian and Lebanese.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations. *arXiv preprint arXiv:2102.10684*.

Ahmed Abdul-Hamid and Kareem Darwish. 2010. Simplified feature set for Arabic named entity recognition. In *Proceedings of the 2010 Named Entities Workshop*, Uppsala, Sweden. Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. Arbert & marbert: Deep bidirectional transformers for arabic.

Mohammed NA Ali, Guanzheng Tan, and Aamir Hussain. 2018. Bidirectional recurrent neural network approach for arabic named entity recognition. *Future Internet*, 10(12):123.

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

Ekaterina Artemova, Maxim Zmeev, Natalia Loukachevitch, Igor Rozhkov, Tatiana Batura, Vladimir Ivanov, and Elena Tutubalina. 2022. Runne-2022 shared task: Recognizing nested named entities.

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.

Borui Cai, He Zhang, Fenghong Liu, Ming Liu, Tianrui Zong, Zhe Chen, and Yunfeng Li. 2022. Overview of nlpcc2022 shared task 5 track 2: Named entity recognition. In *Natural Language Processing and Chinese Computing*, pages 336–341, Cham. Springer Nature Switzerland.

Kenneth Ward Church. 2017. Word2vec. *Natural Language Engineering*, 23(1):155–162.

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*.

Fadl Dahan, Ameur Touir, and Hassan Mathkour. 2015. First order hidden markov model for automatic arabic name entity recognition. *International Journal of Computer Applications*, 123(7).

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab worlds. *Commun. ACM*, 64(4):72–81.

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*.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Overview of hipe-2022: Named entity recognition and linking in multilingual historical documents. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 423–446, Cham. Springer International Publishing.

Toqeer Ehsan, Amjad Ali, and Ala Al-Fuqaha. 2023. Alphabrains at wojoodner shared task: Arabic named entity recognition by using character-based context-sensitive word representations. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Ismail El Bazi and Nabil Laachfoubi. 2019. Arabic named entity recognition using deep learning approach. *International Journal of Electrical & Computer Engineering (2088-8708)*, 9(3).

Niama El Khbir, Urchade Zaratiana, Nadi Tomeh, and Thierry Charnois. 2023. Lipn at wojoodner shared task: A span-based approach for flat and nested arabic named entity recognition. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Abdelkader El Mahdaouy, Salima Lamsiyah, Hamza Alami, Christoph Schommer, and Ismail Berrada. 2023. UM6P & UL at wojoodner shared task: Improving multi-task learning for flat and nested arabic named entity recognition. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Elkaref and Elkaref. 2023. El-kawaref at wojoodner shared task: Stagedner for arabic named entity recognition. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Shereen Elkordi, Noha Adly, and Marwan Torki. 2023. Alexu-aic at wojoodner shared task: Sequence labeling vs mrc and swa for arabic named entity recognition. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Hao Fei, Yafeng Ren, and Donghong Ji. 2020. Dispatched attention with multi-task learning for nested mention recognition. *Information Sciences*, 513:241–251.

Mourad Gridach. 2018. Deep learning approach for arabic named entity recognition. In *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part I 17*, pages 439–451. Springer.

Ramanathan Guha, Rob McCool, and Eric Miller. 2003. Semantic search. In *Proceedings of the 12th international conference on World Wide Web*, pages 700–709.

Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras + baladi: Towards a levantine corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Mariam Hussein, Sarah Khaled, Marwan Torki, and Nagwa Elmakky. 2023. Alex-u 2023 nlp at wojoodner shared task: Arabinder (bi-encoder for arabic named entity recognition). In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. *arXiv preprint arXiv:2103.06678*.

Amin Jaber and Fadi A Zaraket. 2017. Morphology-based entity and relational entity extraction framework for arabic. *arXiv preprint arXiv:1709.05700*.

P. James. 1991. *Knowledge graphs*. Number 945 in Memorandum Faculty of Applied Mathematics. University of Twente, Faculty of Applied Mathematics.

Mustafa Jarrar, Anton Deik, and Bilal Faraj. 2011. Ontology-based data and process governance framework -the case of e-government interoperability in palestine. In *Proceedings of the IFIP International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA'11)*, pages 83–98.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022. Wojood: Nested arabic named entity corpus and recognition using bert. Marseille, France.

Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, and Mohammad Khalilia. 2023a. Salma: Arabic sense-annotated corpus and wsd benchmarks. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*. ACL.

Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlisch. 2023b. Lisan: Yemeni, irqi, libyan, and sudanese arabic dialect copora with morphological annotations. In *The 20th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*. IEEE.

Muhammad Khalifa and Khaled Shaalan. 2019. Character convolutions for arabic named entity recognition with long short-term memory networks. *Computer Speech & Language*, 58:335–346.

Imen Laouirine, Haroun Elleuch, and Fethi Bougares. 2023. Elyadata at wojoodner shared task: Data and model-centric approaches for arabic flat and nested ner. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jiyong Li, Dilshod Azizov, Hilal AlQuabeh, and Shangsong Liang. 2023. Lotus at wojoodner shared task: Multilingual transformers: Unveiling flat and nested entity recognition. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Haneen Liqreina, Mustafa Jarrar, Mohammed Khalilia, Ahmed Oumar El-Shangiti, and Muhammad Abdul-Mageed. 2023. Arabic fine-grained entity recognition. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*. ACL.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. SemEval-2022 task 11: Multilingual complex named entity recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1412–1437, Seattle, United States. Association for Computational Linguistics.

Amal Nayouf, Mustafa Jarrar, Fadi zaraket, Tymaa Hammouda, and Mohamad-Bassam Kurdy. 2023. Nâbra: Syrian arabic dialects with morphological annotations. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*. ACL.

Hiroki Ouchi, Jun Suzuki, Sosuke Kobayashi, Sho Yokoi, Tatsuki Kuribayashi, Ryuto Konno, and Kentaro Inui. 2020. Instance-based learning of span representations: A case study through named entity recognition. *arXiv preprint arXiv:2004.14514*.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2023. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Erik F Sang and Jorn Veenstra. 1999. Representing text chunks. *arXiv preprint cs/9907006*.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*, pages 107–110.

Khaled Shaalan and Hafsa Raza. 2007. Person name entity recognition for arabic. In *Proceedings of the 2007 workshop on computational approaches to semitic languages: common issues and resources*, pages 17–24.

Mohamed Shaheen and Ahmed Magdy Ezzeldin. 2014. Arabic question answering: systems, resources, tools, and future trends. *Arabian Journal for Science and Engineering*, 39:4541–4564.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Diffusionner: Boundary diffusion for named entity recognition. *arXiv preprint arXiv:2305.13298*.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2005. Ace 2005 multilingual training corpus-linguistic data consortium. *URL: https://catalog. ldc. upenn. edu/LDC2006T06*.

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.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23:170.

Ayah Zirikly and Mona Diab. 2014. Named entity recognition system for dialectal Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 78–86, Doha, Qatar. Association for Computational Linguistics.