WojoodNER 2024: The Second Arabic Named Entity Recognition Shared Task

Mustafa Jarrar[σ] **Nagham Hamad**[σ] **Mohammed Khalilia**[σ] **Bashar Talafha**[λ]
AbdelRahim Elmadany[λ] **Muhammad Abdul-Mageed**[λ,ξ]
[σ]Birzeit University, Palestine
[λ]The University of British Columbia
[ξ]MBZUAI
{mjarrar,nhamad,mkhalilia}@birzeit.edu {btalafha,a.elmadany,muhammad.mageed}@ubc.ca

Abstract

We present WojoodNER-2024, the second Arabic Named Entity Recognition (NER) Shared Task. In WojoodNER-2024, we focus on fine-grained Arabic NER. We provided participants with a new Arabic fine-grained NER dataset called $Wojood_{Fine}$, annotated with subtypes of entities. WojoodNER-2024 encompassed three subtasks: (i) Closed-Track Flat Fine-Grained NER, (ii) Closed-Track Nested Fine-Grained NER, and (iii) an Open-Track NER for the Israeli War on Gaza. A total of 43 unique teams registered for this shared task. Five teams participated in the Flat Fine-Grained Subtask, among which two teams tackled the Nested Fine-Grained Subtask and one team participated in the Open-Track NER Subtask. The winning teams achieved F_1 scores of 91% and 92% in the Flat Fine-Grained and Nested Fine-Grained Subtasks, respectively. The sole team in the Open-Track Subtask achieved an F_1 score of 73.7%.

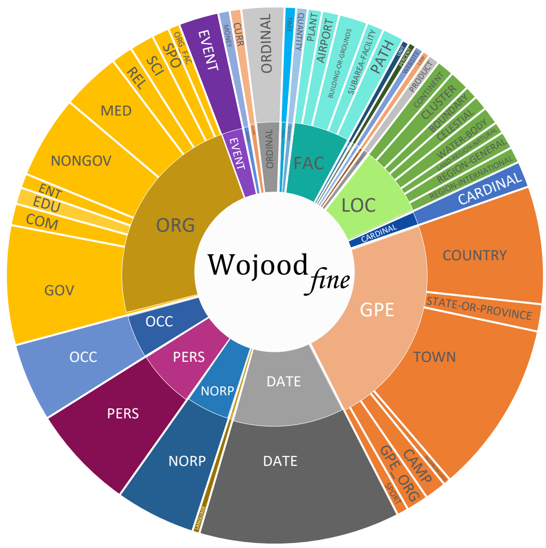

Figure 1: Visualization of the fine-grained entity types in $Wojood_{Fine}$

1 Introduction

NER plays a crucial role in various Natural Language Processing (NLP) applications, such as question-answering systems (Shaheen and Ezzeldin, 2014), knowledge graphs (James, 1991), and semantic search (Guha et al., 2003), information extraction and retrieval (Jiang et al., 2016), word sense disambiguation (Jarrar et al., 2023b; Al-Hajj and Jarrar, 2021), machine translation (Jain et al., 2019; Khurana et al., 2022), automatic summarization (Summerscales et al., 2011; Khurana et al., 2022), interoperability (Jarrar et al., 2011) and cybersecurity (Tikhomirov et al., 2020).

NER involves identifying mentions of named entities in unstructured text and categorizing them into predefined classes, such as PERS, ORG, GPE, LOC, EVENT, and DATE. Given the relative scarcity of resources for Arabic NLP, research in Arabic NER has predominantly concentrated on "flat" entities and has been limited to a few "coarse-grained" entity types, namely PERS, ORG, and LOC. To address this limitation, the WojoodNER shared task series was initiated (Jarrar et al., 2023a). It aims to enrich Arabic NER research by introducing Wojood and $Wojood_{Fine}$ (Liqreina et al., 2023), nested and fine-grained Arabic NER corpora.

In WojoodNER-2024 we provide a new version of Wojood, called $Wojood_{Fine}$. $Wojood_{Fine}$ enhances the original Wojood corpus by offering fine-grained entity types that are more granular than the data provided in WojoodNER-2023. For instance, GPE is now divided into seven subtypes: COUNTRY, STATE-OR-PROVINCE, TOWN, NEIGHBORHOOD, CAMP, GPE_ORG, and SPORT. LOC, ORG, and FAC are also divided into subtypes as shown in Figure 1. $Wojood_{Fine}$ contains approximately 550k tokens and annotated with 51 entity types and subtypes, covering 47k subtype entity mentions. It is worth mentioning that SinaTools supports Wojood and can be accessed via Application Programming Interface (API) (Hammouda et al., 2024).

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 43 teams registered to participate in the shared task. Among these, five teams successfully submitted their models for evaluation on the blind test set during the final evaluation phase.

The rest of the paper is organized as follows: Section 2 provides a brief overview of Arabic NER. We describe the three subtasks and the shared task restrictions in Section 3. Section 4 introduces shared task datasets and evaluation. We present the participating teams, submitted systems and shared task results in Section 5. We conclude in Section 6.

2 Literature Review

NER has been an area of active research for many years, witnessing notable progress recently. This section will cover the evolution from initial efforts in recognizing flat-named entities to the current focus on nested NER, with a particular emphasis on Arabic NER, including discussions on corpora, methodologies, and shared tasks.

Corpora. The majority of Arabic NER corpora are designed for flat NER annotation. ANERCorp (Benajiba et al., 2007), derived from news sources, contains approximately $150k$ tokens and focuses on four specific entity types. CANERCorpus (Salah and Zakaria, 2018) targets Classical Arabic (CA) and includes a dataset of $258k$ tokens annotated for 14 types of entities related to religious contexts. The ACE2005 (Walker et al., 2005) corpus is multilingual and includes Arabic texts annotated with five distinct entity types. The Ontonotes5 (Weischedel et al., 2013) dataset features around 300k tokens annotated with 18 different entity types. However, these corpora are somewhat dated and primarily cover media and political domains, which may not accurately reflect contemporary Arabic usage, particularly as language models are sensitive to changes over time and across domains. Recently, (Jarrar et al., 2022) introduced Wojood, the largest Arabic NER corpus to date, notable for supporting both flat and nested entity annotations. This corpus, essential for this shared task, includes about 550k tokens and covers 21 unique entity types across Modern Standard Arabic (MSA) and two Arabic dialects (Palestinian Curras2 and Lebanese Baladi corpora (Haff et al., 2022)). $Wojood_{Fine}$ (Liqreina et al., 2023), an extension of Wojood adds support for entity sub-types, with a total of 51 entities organized in two-level hierarchy. It is important to note that Wojood has been recently extended to include relationships (Aljabari et al., 2024).

Methodologies. Research in Arabic NER employs a variety of approaches, ranging from rule-based systems (Shaalan and Raza, 2007; Jaber and Zaraket, 2017) to machine learning techniques (Settles, 2004; Abdul-Hamid and Darwish, 2010; Zirikly and Diab, 2014; Dahan et al., 2015; Darwish et al., 2021). Recent studies have adopted deep learning strategies, utilizing character and word embeddings in conjunction with Long-Short Term Memory (LSTM) (Ali et al., 2018), and BiLSTM architectures paired with Conditional Random Field (CRF) layer (El Bazi and Laachfoubi, 2019; Khalifa and Shaalan, 2019). Deep Neural Networks (DNN) are explored in (Gridach, 2018), alongside pretrained Language Models (LM) (Jarrar et al., 2022; Liqreina et al., 2023). Wang et al. (2022) conducted a comprehensive review of various approaches to nested entity recognition, including rule-based, layered-based, region-based, hypergraph-based, and transition-based methods. Fei et al. (2020) introduced a multi-task learning framework for nested NER using a dispatched attention mechanism. Ouchi et al. (2020) developed a method for nested NER that calculates all region representations from the contextual encoding sequence and assigns a category label to each. Readers can also refer to the WojoodNER-2023 shared task for DNN techniques used for flat and nested ArabicNER (Jarrar et al., 2023a).

Shared tasks. While numerous shared tasks exist for NER across different languages and domains, such as MultiCoNER for multilingual complex NER (Malmasi et al., 2022) the HIPE-2022 for NER and linking in multilingual historical documents (Ehrmann et al., 2022), RuNNE-2022 for nested NER in Russian (Artemova et al., 2022), and NLPCC2022 for entity extraction in the material science domain (Cai et al., 2022). WojoodNER-2023 for flat and nested Arabic NER (Jarrar et al., 2023a), upon which WojoodNER-2024 builds on to offer support for entity subtypes.

There are several related shared tasks for understanding Arabic MSA and dialects, such as the ArabicNLU for word-sense disambiguation (Khalilia et al., 2024; Jarrar et al., 2023b), NADI for dialect identification (Abdul-Mageed et al., 2023), AraFinNLP for Cross-dialect Intent detection (Malaysha et al., 2024), among others.

3 Task Description

WojoodNER-2024 confronts the intricacies of Arabic NER with three distinct subtasks: Flat Fine-Grained NER, Nested Fine-Grained NER, and Open-Track NER. These subtasks provide an evaluation environment, allowing researchers to experiment with diverse approaches for identifying and classifying named entities, along with their subtypes, under controlled (closed) and flexible (open) settings.

Remark: the Wojood dataset used in WojoodNER-2023 (Jarrar et al., 2023a) cannot be used in this Shared Task because the two datasets follow different annotation guidelines.

3.1 Closed-Track Flat Fine-Grained NER

In this subtask, we provide the $Wojood_{Fine}$ Flat train (70%) and development (10%) datasets. The final evaluation of the submitted contributions from participants is conducted on the test set (20%). The flat NER dataset follows the same split as the nested NER dataset. The key difference in flat NER is that each token is assigned a single tag, corresponding to the first high-level tag assigned in the nested NER dataset, and followed by a single tag in the second level (subtype). This subtask is a closed track, thus participants can only use the provided datasets to train their systems, with no external datasets permitted.

3.2 Closed-Track Nested Fine-Grained NER

This subtask is similar to Subtask 1. We provide the Wojood-Fine Nested train (70%) and development (10%) datasets, with the final evaluation conducted on the test set (20%). This subtask is a closed track, which means participants can only use the provided datasets to train their systems.

3.3 Open-Track NER - Israeli War on Gaza

This subtask aims to enable participants to explore the benefits of NER in real-world scenarios. Participants can use external resources and are encouraged to experiment with generative models in various ways, such as fine-tuning, zero-shot learning, and in-context learning. The emphasis on generative models in this subtask is intended to help the Arabic NLP research community gain a better understanding of the capabilities and performance gaps of Large Language Models (LLMs) in information extraction, which is currently a less explored area.

We have curated NER dataset called $Wojood^{Gaza}$ pertaining to the ongoing Israeli War on Gaza, based on the assumption that discourse about recent global events will involve mentions from different data distributions. For this subtask, we have collected data from five news domains related to the War, while keeping the identities of these domains confidential. Participants have been provided with a development dataset (10k tokens, 2k from each of the five domains) and a testing dataset (50k tokens, 10k from each domain). Both datasets have been manually annotated with fine-grained named entities, following the same annotation guidelines as in Subtask 1 and Subtask 2, as outlined in (Liqreina et al., 2023). This subtask is divided into two subtasks: 3A-flat and 3B-nested.

3.4 Restrictions

This section outlines the guidelines for participating in the WojoodNER-2024 Shared Task. These rules have been put in place to ensure fairness and transparency for all participants. They also aim to uphold the credibility of the task's assessment process, which is further elaborated on the official shared task FAQ page.

External data. For Subtask 1 and 2, participants are strictly forbidden from using external data from previously labeled datasets or employing taggers previously trained to predict named entities. The use of any resources with prior knowledge of NER is not permitted. On the contrary, Subtask 3 allows the use external resources.

Data format constraints. Submissions for the task must be in a single file containing the model's predictions in CoNLL format. This format includes multiple space-separated columns: the first column for tokens and the subsequent columns for tags. For both flat and nested NER, the tag columns follow a predefined order specified on the shared task webpage. The IOB2 scheme (Sang and Veenstra, 1999) is used for submissions, consistent with the Wojood dataset. Additionally, text segments are separated by a blank line.

4 Datasets and Evaluation Metrics

In this section, we will describe the dataset, evaluation metrics, and the submission procedure.

Datasets The WojoodNER-2024 shared task utilizes the $Wojood_{Fine}$ corpus as a dataset for Subtasks 1 and 2 (Liqreina et al., 2023). For Subtask 3, a different dataset called $Wojood^{Gaza}$ is utilized that is related to the War on Gaza. The $Wojood_{Fine}$ corpus comprises approximately 550k tokens, annotated with nest named entities, using 51 entity types. For the purposes of the shared task, we created a flat NER dataset based on the nest NER dataset. That is, the flat NER dataset is created by simplifying the nested NER and reducing subtypes to the top level only as illustrated in Figure 2 and 3. For both Subtask 1 and Subtask 2, we partitioned the data at the domain level into training, development, and test datasets with a split of 70:10:20, respectively.

Table 1 presents the details of the datasets used in Subtask 1 (FlatNER) and Subtask 2 (NestNER).

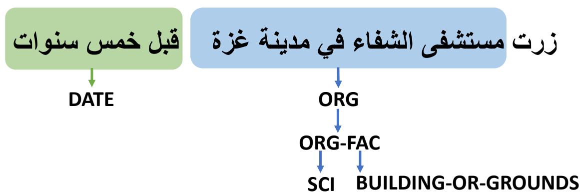

Figure 2: Flat NER example.

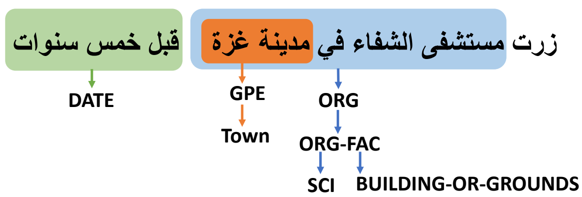

Figure 3: Nested NER example.

The dataset for Subtask 3 is called $Wojood^{Gaza}$. It includes 60k tokens that we collected and annotated specifically for this shared task. The corpus was collected from online news articles published at these outlets: Institute for Palestine Studies, World Health Organization, Palestinian Ministry of Health, Palestine Monetary Authority, Aljazeera, Palestine Economy Portal, Wafa, BNews, AlAraby, Law for Palestine, United Nations, CNN Business, Al Arabiya, Sky News, CNBC Arabia, RT Arabic, Euro News, BBC Arabic.

The articles that were collected from the period of January-March 2024, covering one of these five domains (politics, law, economy, finance, health) and were directly related to the War on Gaza. For each domain, we collected about 12k tokens. Participants are provided with the development dataset (10k tokens, 2k from each of the five domains), and a testing dataset (50k tokens, 10k from each domain). Domain names are not provided to the participants. $Wojood^{Gaza}$ was annotated following the same guidelines as $Wojood_{Fine}$ (Liqreina et al., 2023).

Evaluation metrics. The official and primary evaluation metric for Subtask 1, Subtask 2, and Subtask 3 is the micro-averaged F_1 score. In addition to this metric, we also report system performance in terms of Precision, Recall, and Accuracy.

Submission rules. Participating teams were allowed to submit up to four runs for each test set across the three subtasks. For each team's submissions, we retained only the highest score per task. Although the official results were derived exclusively from the blind test set, we streamlined the evaluation process by establishing four separate CodaLab competitions, one for each subtask[1]. We are keeping the CodaLab 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 ground truth labels for the test sets for any of the subtasks.

5 Shared Task Teams & Results

5.1 Participating Teams

Overall, we received 43 unique team registrations, 26 of them registered in the CodaLab, and only seven teams have submitted their results. These seven teams have submitted 263 valid entries during the testing phase. Specifically, 76 submissions for FlatNER were received from six teams, 168 submissions for NestedNER came from four teams, eight submissions for Gaza-Flat from one team, and 11 submissions for Gaza-Nested from 1

[1] The different CodaLab competitions are available at the following links: `Subtask 1`, `Subtask 2` and `Subtask 3A`, `Subtask 3B`.

Entity Name	NER Tag	FlatNER				NestedNER			
		TRAIN	DEV	TEST	Total	TRAIN	DEV	TEST	Total
Cardinal	CARDINAL	1291	170	341	1802	1312	170	342	1824
Organization	ORG	10590	1488	3006	15084	13143	1863	3741	18747
Government	GOV	5689	848	1673	8210	5764	860	1695	8319
Date	DATE	10705	1592	3028	15325	11346	1691	3206	16243
Language	LANGUAGE	139	16	43	198	140	16	43	199
Group of people	NORP	3586	508	1008	5102	3952	551	1094	5597
Person	PERS	4519	611	1408	6538	5044	677	1565	7286
Occupation	OCC	3717	514	1090	5321	3822	532	1124	5478
GeoPolitical Entity	GPE	8052	1116	2395	11563	16113	2310	4676	23099
Country	COUNTRY	2911	436	834	4181	5744	835	1622	8201
Event	EVENT	1850	282	549	2681	1929	292	569	2790
Facility	FAC	560	86	179	825	777	116	227	1120
Building or ground	BUILDING-OR-GROUNDS	646	92	193	931	706	102	204	1012
Town	TOWN	4970	690	1460	7120	8374	1216	2431	12021
Loction	LOC	747	108	234	1089	985	141	317	1443
Continent	CONTINENT	65	10	23	98	133	23	57	213
Money	MONEY	172	22	33	227	172	22	33	227
Currency	CURR	15	2	8	25	176	24	41	241
Ordinal	ORDINAL	2739	445	889	4073	3444	544	1083	5071
Educational	EDU	440	49	134	623	821	109	229	1159
Time	TIME	309	33	84	426	311	33	84	428
Sports	SPO	11	2	8	21	11	2	8	21
Sport	SPORT	5	2	0	7	5	2	1	8
Land Region Natural	LAND-REGION-NATURAL	158	22	52	232	179	26	59	264
Cluster	CLUSTER	138	18	55	211	222	28	78	328
Quantity	QUANTITY	43	3	9	55	46	3	9	58
Unit	UNIT	6	1	2	9	46	4	11	61
State-or-Province	STATE-OR-PROVINCE	1146	159	372	1677	1292	179	421	1892
Non-Governmental	NONGOV	4030	566	1143	5739	4071	573	1158	5802
Neighborhood	NEIGHBORHOOD	78	5	29	112	87	5	30	122
Water-Body	WATER-BODY	76	14	18	108	88	14	21	123
Percent	PERCENT	92	12	33	137	92	12	33	137
Camp	CAMP	595	69	167	831	605	71	168	844
Path	PATH	52	6	18	76	52	6	18	76
Media	MED	2886	419	807	4112	2886	419	807	4112
Region-General	REGION-GENERAL	275	37	67	379	278	37	69	384
GPE_ORG	GPE_ORG	1000	161	316	1477	1036	167	325	1528
Website	WEBSITE	412	80	116	608	412	80	116	608
Commercial	COM	458	39	111	608	459	40	111	610
Celectial	CELESTIAL	2	0	2	4	2	0	2	4
Subarea - Facility	SUBAREA-FACILITY	91	16	23	130	96	16	23	135
Medical-Science	SCI	102	12	29	143	104	13	30	147
Religious	REL	61	10	24	95	61	10	25	96
ORG_FAC	ORG_FAC	87	7	19	113	87	7	19	113
Region-International	REGION-INTERNATIONAL	67	12	29	108	70	12	29	111
Entertainment	ENT	1	1	1	3	1	1	1	3
Boundary	BOUNDARY	15	4	3	22	15	4	3	22
Plant	PLANT	1	0	0	1	1	0	0	1
Law	LAW	368	47	90	505	368	47	90	505
Product	PRODUCT	61	8	17	86	62	8	19	89
Airport	AIRPORT	5	0	1	6	5	0	1	6
	Total	76034	10850	22173	109057	96947	13913	28068	138928

Table 1: Distribution of NER tags in WojoodNER-2024 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-2024.

team. Table 2 provides details about the teams, their affiliations, and their tasks (1– FlatNER, 2– NestedNER, 3A– Gaza-Flat, and 3B– Gaza-Nested). Out of the seven teams, we received six description papers, which are all accepted for publication.

5.2 Baselines

For Subtask 1 and Subtask 2, we fine-tuned the AraBERT$_{\text{v2}}$ (Antoun et al., 2020) pre-trained model using subtask-specific training data for 20 epochs, with a learning rate of $1e^{-5}$ and a batch size of 8. To ensure optimal model performance, we incorporated 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 sets. We then present the performance metrics of the best-performing model on the test datasets.

Team	Affiliation(s)	Task
Addax (Issam AIT YAHIA, 2024)	Um6p College Of Computing, Morocco	1
Bangor University (Alshammari and Teahan, 2024)	Bangor University, UK	1
DRU (Hamoud et al., 2024; Hamdan et al., 2024)	Arab Center for Research and Policy Studies, Qatar	1,2,3
mucAI (Abdou and Mohsen, 2024)	Technical University of Munich, Germany Helwan University of Cairo, Egypt	1
muNERa (Alotaibi et al., 2024)	King Abdulaziz City for Science and Technology (KACST), Saudi Data and Artificial Intelligence Authority (SDAIA), and King Salman Global Academy for Arabic Language (KSGAAL), Saudi Arabia	1,2

Table 2: List of teams that participated in the WojoodNER-2024 subtasks.

5.3 Results

Table 3, Table 4, and Table 5 presents the leaderboards for Subtask 1–FlatNER, Subtask 2–NestedNER, and Subtask 3A–Gaza respectively, organized in descending order based on the micro-F_1 scores. The micro-F_1 score listed for each team reflects their highest score out of the four allowed submissions for each task.

Rank	Team	F_1	Pre.	Rec.
1	mucAI	91	91	90
2	muNERa	90	91	89
2	Addax	90	89	91
	Baseline-I (ARBERT$_{v2}$)	89	89	90
3	DRU - Arab Center	87	86	86
4	Bangor	86	88	85

Table 3: Results of Subtask 1–FlatNER.

For FlatNER, the mucAI team (Abdou and Mohsen, 2024) achieved the highest F_1 score of 91, with muNERa (Alotaibi et al., 2024) and Addax (Issam AIT YAHIA, 2024) securing second place with 90, DRU taking third place with 87, and Bangor taking fourth place with 86. Notably, three teams outperformed our baseline, as shown in Table 3. The winning team mucAI(Abdou and Mohsen, 2024) surpassed the baseline by 2%. The performance gap between our baseline and the lowest-performing model is approximately 3%. Furthermore, the difference in F_1 scores among the teams is minimal, with a standard deviation of $\sigma = 1.94$.

Rank	Team	F1	Pre.	Rec.
	Baseline-I (ARBERT$_{v2}$)	92	92	93
2	muNERa	91	92	90
3	DRU - Arab Center	90	90	90

Table 4: Results of Subtask 2 – NestedNER.

For NestedNER, none of the teams outperformed the baseline. The muNERa team (Alotaibi et al., 2024) achieved the highest F_1 score of 91, but still 1% below the baseline, followed by DRU team (Hamoud et al., 2024) with a score of 90.

Rank	Team	F_1	Pre.	Rec.
1	DRU - Arab Center	73.7	71.9	75.6

Table 5: Results of Subtask 3 – Gaza-FlatNER.

For the open-track Gaza-FlatNER, only DRU team (Hamoud et al., 2024) reported their results with a recall of 75.9 and F_1 score of 73.5.

5.4 General Description of Submitted Systems

For Subtask 1 and Subtask 2, all models submitted to the shared task employed the transfer learning approach, utilizing pre-trained models trained on diverse data sources. For Subtask 3, LLMs with in-context learning techniques were utilized.

Addax (Issam AIT YAHIA, 2024) proposed a combined tagging approach that merges the main entity type and its subtypes into a single category (e.g., "B-GPE+B-COUNTRY" for "Palestine"). This method follows the IOB2 scheme for entity boundaries and simplifies training by focusing on a single combined tag per entity, integrating both main and subtype information. The model architecture utilizes a two-channel parallel hybrid neural network with an attention mechanism. It employs BERT-based model (AraBERTv0.2-Twitter) embeddings for contextualized word representations and consists of two distinct channels: one using Conv1D layers for local feature extraction and another with Bi-GRU layers to capture long-range dependencies. Additionally, an attention layer focusing on the most relevant input features has been added in each channel.

Bangor (Alshammari and Teahan, 2024) added a linear layer on top of a BERT-based model (bert-base-arabic-camelbert-mix) to classify each token into one of 51 different entity types and subtypes, as well as the "O" label for non-entity tokens. This linear layer maps the contextualized embeddings produced by BERT to the desired output labels.

muNERa (Alotaibi et al., 2024) team adapted Wojood dataset to fit the input requirements of the Translation between Augmented Natural Languages (TANL) framework (Paolini et al., 2021). The preprocessing steps included extracting hierarchical tags (parent, subtype, sub-subtype) and their spans using the IOB2 scheme. Each token and its corresponding labels were reformatted to align with the TANL framework's specifications. TANL was used for Subtask 1 and Subtask 2. In this framework, both input and output are structured in augmented natural languages and enclosed in square brackets (e.g., [token | entity type]). For nested entities, TANL can represent entity hierarchies, such as [token [token | entity type1] | entity type2]. They utilized two distinct TANL models for handling flat and nested entities. A decoder-encoder model (AraT5v2) is used as base for the TANL model. Additionally, they used a FastText (FT) classifier as a secondary tagger, first using TANL to detect spans and assign level-1 (parent) tags, and then applying the FT classifier to tag the detected spans with level-2 and level-3 tags. The best-performing TANL architecture was achieved without using FT.

mucAI (Abdou and Mohsen, 2024) team proposed a two-step methodology: joint vanilla fine-tuning followed by k-Neared Neighbor (KNN) at inference time. BERT (AraBERTv02) is used as the backbone for generating word embeddings. These embeddings are then fed into two multi-layer perceptrons (MLP) that are trained jointly. The first head predicts one of the predefined 21 main entity tags. The second head predicts one of the predefined 31 sub-entities. A “Datastore” is constructed as a database that has a contextualized representation for each token alongside the label in each sentence in the training set. The "Datastore" was queried during inference to retrieve the k nearest neighbors based on a similarity score, derive the distribution of labels from these neighbors, and then interpolate this distribution with the main MLP model's distribution using an interpolation factor to obtain the final label probabilities.

DRU-Arab Center (Hamoud et al., 2024) proposed three strategies to deal with the Flat and Nested subtasks. (1) A single-layer approach, where they fine-tuned different BERT-based models to predict all types and subtypes in one shot, using a 103-length one-hot encoded vector for each type and subtype, including the "O" tag. They experimented with GEMMA (Team et al., 2024), and AraBERTv2 (Antoun et al., 2020), and fine-tuned BLOOMZ-7b-mt on a high-quality Arabic dataset (Muennighoff et al., 2023). (2) Another attempt was the One×1 classifier method, which separated type and subtype classification by dedicating a model for each, training one instance of (AraBERTv2) exclusively for predicting main types and another instance for predicting subtypes. (3) In the One×4 Classifier Method, instead of only one model for subtypes, they trained four instances, each specialized in the sub-types of a specific group: GPE, ORG, FAC, LOC, as the other main types have no subtypes. Among these strategies, the One×1 approach achieved the highest performance on both Subtask 1 and Subtask 2.

For the open track Subtask 3, (Hamdan et al., 2024), DRU-Arab Center utilized LLMs (*Cohere's Command R model* (Command R Team)) and in-context learning to solve this task. In the prompt design, they wrote a detailed system prompt that outlines the steps for tagging tokens according to the $Wojood_{Fine}$ annotation guidelines. The prompt instructs the LLM to perform NER for Arabic text by predicting up to three levels of tags—high-level tags, subtypes, and specific subtypes for certain entities—while simplifying the task to two tag levels for practical purposes, and outputting predictions in CSV format; illustrative examples are provided to guide the model, and specific instructions ensure the correct application of the IOB2 schema and handle complex subtypes during post-processing. Command R's output quality issues included producing extra or missing tokens. To solve that, they post-processed the generated output to match the expected format by assigning the tag "O" to ground truth tokens without corresponding predicted tokens or hallucinated tags, and by converting the remaining format issues to the expected output.

6 Conclusion

In this paper, we present the outcomes of the second edition of WojoodNER shared task. The results from the participating teams highlight the ongoing difficulties in NER, yet it is encouraging to see that various innovative approaches, particularly those leveraging the power of language models, have proven effective in tackling this complex task. As we progress, we are dedicated to advancing research in this field. Our vision includes continuous efforts to improve Arabic NER, drawing on the valuable insights from WojoodNER-2024 and exploring new solutions. Additionally, we plan to expand the $Wojood_{Fine}$ corpus to encompass more dialects.

Limitations

Similar to WojoodNER-2023, WojoodNER-2024 aimed for the broadest possible coverage, primarily focusing on MSA data. This dataset used this year, $Wojood_{Fine}$, includes limited data from dialects. It only includes text from Palestinian and Lebanese Arabic. We plan to include the other dialects, especially the Syrian *Nabra* dialects (Nayouf et al., 2023) as well as the four dialects in the *Lisan* (Jarrar et al., 2023c) corpus. Additionally, the $Wojood^{Gaza}$ dataset used in Subtask 3 covers only the initial phase of the Israeli War on Gaza, excluding the subsequent genocidal and starvation events.

Ethics Statement

The datasets provided for this shared task are derived from public sources, eliminating specific privacy concerns. The results of the shared task will be made publicly available to enable the research community to build upon them for the public good and peaceful purposes. Our datasets and research are strictly intended for non-malicious, peaceful, and non-military purposes.

Acknowledgements

This research is partially funded by the Palestinian Higher Council for Innovation and Excellence and by the research committee at Birzeit University.

Muhammad Abdul-Mageed acknowledges support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[2] and UBC ARC-Sockeye.

We extend our gratitude to Taymaa Hammouda for the technical support and to the students who helped and supported us during the annotation process, especially Haneen Liqreina, Lina Duaibes, Shimaa Hamayel, Rwaa Assi, Hiba Zayed, and Sana Ghanim.

[2] https://alliancecan.ca

References

Ahmed Abdou and Tasneem Mohsen. 2024. mucai at wojoodner shared task: Arabic named entity recognition with nearest neighbor search. In *Proceedings of the 2nd Arabic Natural Language Processing Conference (Arabic-NLP), Part of the ACL 2024.* Association for Computational Linguistics.

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, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.

Moustafa Al-Hajj and Mustafa Jarrar. 2021. ArabGlossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 40–48, Online. INCOMA Ltd.

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

Alaa Aljabari, Lina Duaibes, Mustafa Jarrar, and Mohammed Khalilia. 2024. Event-Arguments Extraction Corpus and Modeling using BERT for Arabic. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok, Thailand. Association for Computational Linguistics.

Nouf M. Alotaibi, Haneen Alhomoud, Hanan Murayshid, Waad Alshammari, Nouf Alshalawi, and Sakhar Alkhereyf. 2024. muNERa at wojoodner 2024 shared task: Multi-tasking NER Approach. In *Proceedings of the 2nd Arabic Natural Language Processing Conference (Arabic-NLP), Part of the ACL 2024.* Association for Computational Linguistics.

Norah Alshammari and William Teahan. 2024. Bangor university at wojoodner shared task 2024: Advancing arabic named entity recognition with camelbert-mix. In *Proceedings of the 2nd Arabic Natural Language Processing Conference (Arabic-NLP), Part of the ACL 2024.* Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

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.

Command R Team. Command R Documentation. `https://docs.cohere.com/docs/command-r`. Accessed: 2024-07-01.

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.

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.

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).

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.

Nancy Hamdan, Hadi Hamoud, Chadi Abou Chakra, Osama Rakan Al Mraikhat, Doha Albared, and Fadi A. Zaraket. 2024. DRU at Wojood NER Shared Task 2024: ICL LLM for Arabic NER. In *Proceedings of the 2nd Arabic Natural Language Processing Conference (Arabic-NLP), Part of the ACL 2024.* Association for Computational Linguistics.

Tymaa Hammouda, Mustafa Jarrar, and Mohammed Khalilia. 2024. SinaTools: Open Source Toolkit for Arabic Natural Language Understanding. In *Proceedings of the 2024 AI in Computational Linguistics (ACLING 2024)*, Procedia Computer Science, Dubai. ELSEVIER.

Hadi Hamoud, Chadi Abou Chakra, Nancy Hamdan, Osama Rakan Al Mraikhat, Doha Albared, and Fadi A. Zaraket. 2024. DRU at Wojood NER Shared Task 2024: A Multi-level Method Approach. In *Proceedings of the 2nd Arabic Natural Language Processing Conference (Arabic-NLP), Part of the ACL 2024.* Association for Computational Linguistics.

Ismail Berrada Issam AIT YAHIA, Houdaifa Atou. 2024. Addax at wojoodner 2024: Attention-based dual-channel neural network for arabic named entity recognition. In *Proceedings of the 2nd Arabic Natural Language Processing Conference (Arabic-NLP), Part of the ACL 2024.* Association for Computational Linguistics.

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

Alankar Jain, Bhargavi Paranjape, and Zachary C Lipton. 2019. Entity projection via machine translation for cross-lingual ner. *arXiv preprint arXiv:1909.05356*.

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

Mustafa Jarrar, Muhammad Abdul-Mageed, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, Nagham Hamad, and Alaa' Omar. 2023a. WojoodNER 2023: The First Arabic Named Entity Recognition Shared Task. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 748–758. ACL.

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. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, and Mohammed Khalilia. 2023b. SALMA: Arabic Sense-annotated Corpus and WSD Benchmarks. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 359–369. ACL.

Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlisch. 2023c. 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.

Ridong Jiang, Rafael E. Banchs, and Haizhou Li. 2016. Evaluating and combining name entity recognition systems. In *Proceedings of the Sixth Named Entity Workshop*, pages 21–27, Berlin, Germany. Association for Computational Linguistics.

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.

Mohammed Khalilia, Sanad Malaysha, Reem Suwaileh, Mustafa Jarrar, Alaa Aljabari, Tamer Elsayed, and Imed Zitouni. 2024. ArabicNLU 2024: The First Arabic Natural Language Understanding Shared Task. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok, Thailand. Association for Computational Linguistics.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2022. Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, 82.

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*, pages 310–323. ACL.

Sanad Malaysha, Mo El-Haj, Saad Ezzini, Mohammed Khalilia, Mustafa Jarrar, Sultan Nasser, Ismail Berrada, and Houda Bouamor. 2024. AraFinNLP 2024: The First Arabic Financial NLP Shared Task. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok, Thailand. Association for Computational Linguistics.

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.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. 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*, pages 12–23. 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*.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *arXiv preprint arXiv:2101.05779*.

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.

Rodney L Summerscales, Shlomo Argamon, Shangda Bai, Jordan Hupert, and Alan Schwartz. 2011. Automatic summarization of results from clinical trials. In *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pages 372–377. IEEE.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Mikhail Tikhomirov, N. Loukachevitch, Anastasiia Sirotina, and Boris Dobrov. 2020. Using bert and augmentation in named entity recognition for cybersecurity domain. In *Natural Language Processing and Information Systems*, pages 16–24, Cham. Springer International Publishing.

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.