DRU at WojoodNER 2024: A Multi-level Method Approach

Hadi Hamoud and **Chadi Abou Chakra** and **Nancy Hamdan**
and **Osama Rakan Al Mraikhat** and **Doha Albared** and **Fadi A. Zaraket**
Arab Center for Research and Policy Studies, Doha
{hhamoud,cabouchakr,nhamdan,oalmraikhat,dal007,fzaraket}@dohainstitute.edu.qa

Abstract

In this paper, we present our submission for the WojoodNER 2024 Shared Tasks addressing flat and nested sub-tasks (1, 2). We experiment with three different approaches. We train (i) an Arabic fine-tuned version of BLOOMZ-7b-mt, GEMMA-7b, and AraBERTv2 on multi-label token classifications task; (ii) two AraBERTv2 models, on main types and sub-types respectively; and (iii) one model for main types and four for the four sub-types. Based on the Wojood NER 2024 test set results, the three fine-tuned models performed similarly with AraBERTv2 favored (F1: Flat=.8780 Nested=.9040). The five model approach performed slightly better (F1: Flat=.8782 Nested=.9043).

1 Introduction

Named Entity Recognition (NER) is a crucial task in natural language processing for detecting and categorizing named entities like personal names, locations, and organizations in text. It has diverse applications, such as extracting genes and diseases in biomedical texts (Perera et al., 2020), identifying financial entities in finance and legal documents (Au et al., 2022), and recognizing names and locations in social media posts (Moon et al., 2018).

Named entities are classified as flat or nested. Flat entities are contiguous words with no overlap, while nested entities have a layered structure where one entity can contain another. This complexity has led to the development of specialized tools and models for both types. In the development of Arabic NER, the Wojood corpus (Jarrar et al., 2022) was devised which provides a robust foundation for Arabic flat and nested NER. It includes 550K manually annotated tokens across 21 entity types. The Wojood corpus was further transformed into Wojood-Fine (Liqreina et al., 2023), which introduces 31 new fine-grained sub-types for four key entity categories: Geopolitical Entity (GPE), Location (LOC), Organization (ORG), and Facility (FAC).

To further accelerate Arabic NER research, the WojoodNER shared task series (Jarrar et al., 2023) was introduced. The 2024 shared task (Jarrar et al., 2024) utilizes the Wojood-Fine corpus through two sub-tasks:

(i) *WojoodFine-Flat*: Each token is assigned to the first high-level tag, in addition to an assigned sub-type when applicable

(ii) *WojoodFine-Nested*: Each token is assigned to all identified entity types along with sub-types when applicable.

Both sub-tasks are confined to the use of the given Wojood-Fine training and development sets. This paper presents our submissions to the WojoodNER 2024 Shared Task and offers the following contributions:

- Training three models: BLOOMZ-7b-mt fine-tuned on a custom Arabic Dataset (Muennighoff et al., 2023) (BigScience Workshop et al., 2023), Google's GEMMA (Gemma Team et al., 2024), and AraBERTv2(Antoun et al., 2020) on multi-label token classification tasks using the given training set.
- Training two separate AraBERTv2-based models: one model to detect the main 21 types, second model to detect the given sub-types only.
- Training five separate AraBERTv2-based: one model to detect the main 21 types, and one model for each of the four sub-type categories specifically.

2 Related Work

The Arabic NLP research community has made notable advancements in Arabic NER, particularly in the WojoodNER 2023 Shared Task, with different teams showcasing various effective approaches.

Proceedings of The Second Arabic Natural Language Processing Conference, pages 874–879

	#Sentences	#Tokens	#Ent Flat	#Ent Nested
Train	23,125	390,999	191,821	221,216
Dev	3,304	57,555	27,317	31,591

Table 1: Dataset Statistics

The ELYADATA team used both data and model-centric approaches, focusing on data cleaning and re-sampling for dataset imbalance. Their best model, which did not use re-sampling, treated NER as a boundary-denoising diffusion process using the DiffusionNER model, fine-tuned with the AraBERT encoder (Laouirine et al., 2023).

The UM6P and UL team developed a BERT-based multi-task learning model for flat and nested Arabic NER. They used Arabic Pretrained Language Models (PLMs) to encode sentences and combined several training objectives to boost performance. Among the PLMs evaluated were, QARiB (Abdelali et al., 2021), CAMeLBERT-Mix (Inoue et al., 2021), and ARBERTv2 which yielded the best results (El Mahdaouy et al., 2023).

The AlexU-AIC team used sequence labeling and machine reading comprehension (MRC) techniques for Arabic NER. They applied an MRC-based approach for flat NER and fine-tuned a PLM for nested NER. Using the JABER (Ghaddar et al., 2022) pre-trained model for sequence labeling, they created queries for each entity category in the MRC method and enhanced results with Stochastic Weighted Averaging (SWA) (Elkordi et al., 2023).

The LIPN team used a span-based approach with a PLM for token representation and neural network classifiers for both flat and nested Arabic NER. They applied global decoding for flat NER and a greedy strategy for nested NER, involving token encoding, span enumeration, and classification (El Elkhbir et al., 2023).

3 Data

We provide here a description of the released datasets by the shared task organizers. The WojoodFine corpus is offered in two versions, Flat, and Nested. Training, development, and test splits are available in both versions. Table 1 showcases statistics on each of the provided splits. Tokens are labeled using the Inside, Outside, Beginning (IOB) tagging format. In the flat dataset, a given token is assigned the first appearing high-level main entity type label assigned to that token in the sentence. In addition, sub-types are assigned to the token when applicable. For the nested dataset, all identified entities and sub-entities are included. Table 2 presents entity sub-types statistics in flat and nested datasets. Two special cases worth noting, are sub-types GPE_ORG and ORG_FAC. In the occurrence of these two sub-types, two additional sub-types are assigned, one from the first main entity, and the other from the second main entity. Figure 1 showcases an example. ORG is assigned as main entity, with ORG_FAC as a sub-type. Therefore we should assign two "subsub" entities matching ORG and FAC respectively.

		Flat		Nested	
Tag	**Sub-type Tag**	**Train**	**Dev**	**Train**	**Dev**
GPE	COUNTRY	3,445	532	6,320	936
	STATE-OR-PROVINCE	2,503	365	2,753	401
	TOWN	8,685	1,224	13,084	1,892
	NEIGHBORHOOD	214	15	228	15
	CAMP	1,379	164	1,402	167
	GPE_ORG	1,274	209	1,324	217
	SPORT	6	2	6	2
LOC	CONTINENT	68	10	136	23
	CLUSTER	335	41	476	59
	BOUNDARY	46	12	46	12
	CELESTIAL	2	0	2	0
	WATER-BODY	166	35	186	35
	LAND-REGION-NATURAL	305	41	329	45
	REGION-GENERAL	703	97	709	97
	REGION-INTERNATIONAL	143	25	149	25
ORG	GOV	12,410	1,859	12,543	1,875
	COM	1,245	107	1,248	108
	EDU	1,178	130	1,944	249
	ENT	1	2	1	2
	NONGOV	11,654	1,585	11,753	1,599
	MED	6,260	914	6,260	914
	REL	202	33	202	33
	SCI	349	43	354	45
	SPO	22	7	22	7
	ORG_FAC	286	26	286	26
FAC	PLANT	3	0	3	0
	AIRPORT	12	0	13	0
	BUILDING-OR-GROUNDS	1,566	226	1,654	244
	SUBAREA-FACILITY	253	39	260	39
	PATH	155	10	156	10
	Total	**54,870**	**7,753**	**63,849**	**9,077**

Table 2: Training and Development datasets sub-entity count

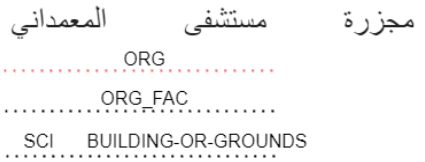

Figure 1: Example of sub-types and subsub-types

4 Multi-label Token Classification

We applied the same training process for all our model iterations. The only change was in input classification. We one-hot encoded B- (Beginning) and I- (Inside) type and sub-type token tags. We utilized binary cross-entropy with logits loss as the primary loss function, handling multiple overlap-

ping entity types per token. This function is particularly useful for binary classification. Literature suggests this works with multi label NER(Richie et al., 2023). The function combines a sigmoid and binary cross-entropy under one layer.

$$\ell_n := \begin{cases} w_n \cdot [(1-y_n) \cdot x_n + \ln(1+e^{-x_n})] & \text{if } x_n > 0 \\ w_n \cdot [-x_n \cdot y_n + \ln(e^{x_n}+1)] & \text{otherwise} \end{cases}$$

It is more numerically stable than using a plain sigmoid followed by a binary cross-entropy loss due to expecting logits exposed to inputs. Therefore, in order to obtain the output probabilities for each class, we directly obtain output probabilities and subject them to thresholds. To accommodate for variations, we apply a pre-processing step before training for classification, and ignore tokens pertinent to specific model architectures. We use -100 label across the one-hot encoded vector, and ignore them in prediction.

We trained all iterations on the nested dataset, and extracted the highest level entity type and corresponding sub-types to infer the flat prediction.

5 Type/Sub-type Classification Method

We first attempted to train different models, with different architectures, on predicting all types and sub-types in one shot. The given model would take as input a 103 length one-hot encoded vector, for each of the B- and I- types and sub-types, and the missing tag notation "O". We experimented with BLOOM, GEMMA, and BERT.

BLOOMZ-7b-mt (Muennighoff et al., 2023) model has 7 billion parameters. It augments BLOOMZ with multilingual mT0 capabilities (Muennighoff et al., 2023). We further fine-tuned this model on a high-quality Arabic dataset, including Arabic high school curricula books, and specialized books in social sciences and humanities.

Our second option, GEMMA, is a 7 billion parameter model, pre-trained on a wide variety of languages including Arabic.

AraBERTv2 is the third and lightweight option. It follows BERT architecture and is trained on Arabic Wikipedia, and quality newspapers such as Assafeer.

6 One$\times$1 Classifiers Method

We separated type and sub-type classification and dedicated a model for each. In our first iteration of this design, we trained one instance of AraBERTv2 to exclusively predict main types. We trained another instance to predict sub-types. We then evaluated the data across both models and aggregated their outputs, compensating for missing tags and rooting to improve recall.

7 One$\times$4 Classifier Method

Instead of only one model for sub-types, we trained four instances, each specialized in the sub-types of a specific group: GPE, ORG, FAC, LOC. The other main types have no subtypes. **Experiment 1** predicts the 21 main types with no changes. The four specialized encompass all the sub-types. **Experiment 2** allows predicting GPE with ORG and ORG with FAC. At the subtype level, the GPE-specialized model does not predict GPE_ORG and the ORG-specialized model does not predict ORG_FAC. The intuition is that this reduces class intersection confusion for the main types model. For example, if an entity was initially tagged as GPE and had sub-entity tags: GPE_ORG, COUNTRY, and GOV, we transformed GPE_ORG to ORG. The final output aggregates the results from both stages.

8 System Description

Larger models, namely BLOOMZ and GEMMA. were fine-tuned with a sequence length configuration of 512 tokens, for 12 epochs on a batch size of 8 for training and evaluation. Model fine-tuning was executed on 1x NVIDIA A100 80 GB VRAM GPU. **BERT-based Models** were trained for 10 epochs on a batch size of 16, on 1x NVIDIA A4000 16 GB VRAM GPU.

9 Results

We experimented with BLOOMZ, GEMMA, and AraBERTv2. All models performed similarly, with AraBERTv2 favored slightly (F1 0.8780 and 0.9040 for flat and nested respectively).

We experimented with the One$\times$1 setup, where one stage was dedicated to predicting main types and another for sub-types. Our first iteration performed adequately, but was still outscored by a single-layer approach in both tasks (0.8780 and 0.9040 compared to 0.8746 and 0.9033 in F1).

One$\times$4 experiment 2, where we resolved subclass interactions, proved to be the most effective, featuring the best performance on both tasks with 0.8782 and 0.9043.

Based on the development set provided, we calculated evaluation metrics per entity. The results presented in Table 3 reveal a disparity in the model's performance across different entity types.

Label	Precision	Recall	F1-Score	Support
MED	0.9976	0.9952	0.9964	419
COUNTRY	0.9833	0.9892	0.9863	835
TOWN	0.9825	0.9671	0.9747	1217
CAMP	0.9459	0.9859	0.9655	71
STATE-OR-PROVINCE	0.9392	0.9497	0.9444	179
LANGUAGE	0.7692	0.625	0.6897	16
GPE_ORG	0.4364	0.9042	0.5887	167
WEBSITE	0.5455	0.45	0.4932	80
TIME	0.2	0.0303	0.0526	33
CURR	0.0	0.0	0.0	24

Table 3: Evaluation Metrics for best and worst performing entities on the five-model approach

	Model	Precision	Recall	F1
Flat	M1	**0.8668**	0.8712	0.8690
	M2	0.8584	0.8626	0.8605
	M3	0.8659	**0.8905**	**0.8780**
Nested	M1	**0.9067**	0.8926	0.8996
	M2	0.8913	0.8769	0.8840
	M3	0.9025	**0.9056**	**0.9040**

Table 4: Results of Fine-tuning Different Architectures. **M1**: Finetuned Bloomz-7b1-mt, **M2**: gemma-7b, **M3**: AraBERTv2

10 Discussion

The results in Table 4 highlight that while all models performed similarly, AraBERTv2 had a slight edge, particularly notable in both flat and nested F1 scores. As showcased in Table 5, the One$\times$1 setup, although logical in its approach to separate main types and sub-types prediction, did not surpass the single-layer approach, which maintained higher F1 scores.

The One$\times$4 experiment 1's lower F1-score on the nested task compared to the single-layer approach suggests that increasing complexity in model stages does not necessarily lead to better performance. However, One$\times$4 experiment 2's superior performance underscores the benefit of resolving subclass interactions within the same framework.

The disparity in model performance across different entity types, as shown in Table 3, indicates that some entities are easier to predict with high precision and recall, while others present more challenges. This suggests areas for targeted improvements, such as the need for additional training data tailored to the harder-to-predict entities, which was a limitation in this study due to the nature of the challenge.

	Model	Precision	Recall	F1
Flat	M1	0.8659	**0.8905**	0.8780
	M2	0.8646	0.8848	0.8746
	M3	**0.8673**	0.8894	**0.8782**
Nested	M1	0.9025	0.9056	0.9040
	M2	0.8940	**0.9127**	0.9033
	M3	**0.9031**	0.9054	**0.9043**

Table 5: Results of Fine-tuning Different Approaches. **M1**: AraBERTv2 on all types and sub-types, **M2**: One$\times$1 approach, **M3**: One$\times$4 approach

	Team	Precision	Recall	F1
Flat	mucAI	91	90	91
	muNERa	91	89	90
	Addax	89	91	90
	baseline	89	89	89
	DRU AC	86	88	87
	Bangor	88	85	86
Nested	baseline	92	93	92
	muNERa	92	90	91
	DRU AC	90	90	90

Table 6: WojoodFine 2024 Shared Task results for flat and nested subtasks compared with the given baselines.

11 Conclusion & Future Work

This study explored three attempts to enhance the performance of WojoodFine: fine-tuning with different architectures, specializing a model for predicting types and another for sub-types, and further differentiation by developing individual models for each sub-type category. Among these strategies, training a model for the main tag and models on each sub-type category scored the highest accuracy on sub-tasks 1 and 2 with 0.8782 and 0.9043 F1 scores respectively. Table 6 showcases the shared task results. Our approach ranked second on codalab (Pavao et al., 2023) for the nested sub-task, outperforming the reported results in (Liqreina et al., 2023) (0.885 F1).

Looking forward, the two-stage approach employed in this study shows substantial promise and opens the door for further exploration. The promising results observed from predicting sub-type tags independently suggest that focusing on these tags could significantly refine the overall process.

12 Limitations

Our work has several limitations: the generalizability of our models was restricted to specific entity types and sub-types given in the used dataset, the use of external data was restricted, and additional fine-tuning with more epochs could have improved performance.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwishet al. and Younes Samih. 2021. Pretraining bert on arabic tweets: Practical considerations. *Preprint*, arXiv:2102.10684.

Wissam Antoun, Fady Balyet al. 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.

Ting Wai Terence Au, Ingemar J. Coxet al. and Vasileios Lampos. 2022. E-ner – an annotated named entity recognition corpus of legal text. *Preprint*, arXiv:2212.09306.

BigScience Workshop et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Niama El Elkhbir, Urchade Zaratiana, Nadi Tomehet al. and Thierry Charnois. 2023. LIPN at WojoodNER shared task: A span-based approach for flat and nested Arabic named entity recognition. In *Proceedings of ArabicNLP 2023*, pages 789–796, Singapore (Hybrid). Association for Computational Linguistics.

Abdelkader El Mahdaouy, Salima Lamsiyah, Hamza Alami, Christoph Schommeret al. 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 ArabicNLP 2023*, pages 777–782, Singapore (Hybrid). Association for Computational Linguistics.

Shereen Elkordi, Noha Adlyet al. and Marwan Torki. 2023. AlexU-AIC at WojoodNER shared task: Sequence labeling vs MRC and SWA for Arabic named entity recognition. In *Proceedings of ArabicNLP 2023*, pages 771–776, Singapore (Hybrid). Association for Computational Linguistics.

Gemma Team et al. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Abbas Ghaddar, Yimeng Wu, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao Xing, Yasheng Wang, Duan Xinyu, Zhefeng Wang, Baoxing Huai, Xin Jiang, Qun Liuet al. and Philippe Langlais. 2022. Jaber and saber: Junior and senior arabic bert. *Preprint*, arXiv:2112.04329.

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

Mustafa Jarrar, Muhammad Abdul-Mageed, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, Nagham Hamadet al. and Alaa' Omar. 2023. WojoodNER 2023: The first Arabic named entity recognition shared task. In *Proceedings of Arabic-NLP 2023*, pages 748–758, Singapore (Hybrid). Association for Computational Linguistics.

Mustafa Jarrar, Nagham Hamad, Mohammed Khalilia, Bashar Talafhaet al. and Muhammad Elmadany, AbdelRahim Abdul-Mageed. 2024. WojoodNER 2024: The Second Arabic Named Entity Recognition Shared Task. In *Proceedings of the 2nd Arabic Natural Language Processing Conference (Arabic-NLP), Part of the ACL 2024.* Association for Computational Linguistics.

Mustafa Jarrar, Mohammed Khaliliaet al. and Sana Ghanem. 2022. Wojood: Nested Arabic named entity corpus and recognition using BERT. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3626–3636, Marseille, France. European Language Resources Association.

Imen Laouirine, Haroun Elleuchet al. and Fethi Bougares. 2023. ELYADATA at WojoodNER shared task: Data and model-centric approaches for Arabic flat and nested NER. In *Proceedings of ArabicNLP 2023*, pages 759–764, Singapore (Hybrid). Association for Computational Linguistics.

Haneen Liqreina, Mustafa Jarrar, Mohammed Khalilia, Ahmed El-Shangitiet al. and Muhammad Abdul Mageed. 2023. Arabic fine-grained entity recognition. In *Proceedings of ArabicNLP 2023*, pages 310–323.

Seungwhan Moon, Leonardo Neveset al. and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. *CoRR*, abs/1802.07862.

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 Raffet al. 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.

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

Nadeesha Perera, Matthias Dehmeret al. and Frank Emmert-Streib. 2020. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, 8:673.

Russell Richie, Victor M Ruiz, Sifei Han, Lingyun Shiet al. and Fuchiang (Rich) Tsui. 2023. Extracting social determinants of health events with transformer-based

multitask, multilabel named entity recognition. *Journal of the American Medical Informatics Association*, 30(8):1379–1388.