ELYADATA at WojoodNER Shared Task: Data and Model-centric Approaches for Arabic Flat and Nested NER

Imen Laouirine[†] and **Haroun Elleuch**[†] and **Fethi Bougares**
firstname.lastname@elyadata.com

Abstract

This paper describes our submissions to the WojoodNER shared task organized during the first ArabicNLP conference. We participated in the two proposed sub-tasks of flat and nested Named Entity Recognition (NER). Our systems were ranked first over eight and third over eleven in the Nested NER and Flat NER, respectively. All our primary submissions are based on DiffusionNER models, where the NER task is formulated as a boundary-denoising diffusion process. Experiments on nested WojoodNER achieves the best results with a micro F1-score of **93.73%**. For the flat sub-task, our primary system was the third-best system, with a micro F1-score of **91.92%**[1].

Keywords: nested NER, flat NER, DiffusionNER, PIQN, data re-sampling.

1 Introduction

Named Entity Recognition is the task of locating a word or a phrase that references a particular entity within a given text. It is among the most prominent challenges in Natural Language Processing (NLP). NER has been an active research area with growing interest and significant development over the past twenty years. This increased interest has led to the organization of multiple NER shared tasks and evaluation campaigns, especially for English (Tjong Kim Sang, 2002) and some other languages (Tsygankova et al., 2019) (Benikova et al., 2014) (Nguyen et al., 2019).

As for the used techniques and methods applied to NER, there are mainly four approaches according to Jehangir et al. (2023): rule-based algorithms, supervised and unsupervised machine learning algorithms, and deep-learning algorithms. The rule based approach relies on predefined grammatical rules and dictionaries to identify named entities. These systems are precise, but do not generalize well. Supervised learning approaches can achieve high accuracy, yet demand labelled data and may give poor results for unseen domains. In contrast, unsupervised learning techniques retrieve named entities without requiring labelled data. However, the absence of labels makes it challenging to assess the performance of unsupervised models effectively (Jehangir et al., 2023). The advent of deep learning models like Transformers (Vaswani et al., 2017) which excel across domains and languages, enabled researches to make bigger strides towards better performance.

Multiple studies have addressed the challenges of identifying and classifying named entities in Arabic text. Notable initiatives include the work by Benajiba et al. (2007a), which introduced a NER system relying only on n-grams and maximum entropy, and it achieved an F1-score of 54.11% on ANERcorp dataset (Benajiba et al., 2007b). Additionally, Gridach (2016) has used deep learning methodologies to enhance the performance of Arabic NER with an F1-score 88.64% on the same dataset (Qu et al., 2023). Furthermore, multiple pretrained models such as AraBERT, MARBERT, and JABER were fine-tuned to achieve respectively 90.51%, 80.5% and 84.20% F1-score on ANERcorp (Qu et al., 2023).

Despite these previous efforts and progress made for Arabic NER, nested NER still constitutes a challenging task not well studied in Arabic. Nested NER refers to the particular case of NER where entities are nested within each other, possibly with different tags. Another common challenge in Arabic NER lies the specific nature of the language itself: Arabic is a morphologically rich and highly ambiguous language with numerous dialectal variants and a significant amount of code-switching (Jarrar et al., 2022; Darwish et al., 2021).

These factors, combined, make NER in its nested or flat variants, a particularly challenging task when

[0†] Equal contribution

[1] Our code is available at https://github.com/elyadata/NER_shared_task_2023

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 759–764

performed on Arabic and dialectal Arabic. Hence, the motivation for the 2023 NER Shared Task (Jarrar et al., 2023) which aims to produce models that can overcome the aforementioned challenges. The main contribution of this work can be summarized as follows:

- Experimenting with re-sampling techniques for dataset unbalance alleviation.
- Fine-tuning state-of-the-art-models for both the nested and flat NER sub-tasks.
- Achieving best results for the nested NER model submission.

This paper is organized as follows. Section 2 describes the Wojood dataset. Section 3 presents the implemented flat and nested tasks, and Section 4 details the experiments and the obtained results. The overall results are discussed in Section 5 before concluding the paper in section 6.

2 Dataset

The NER Shared Task dataset "Wojood" (Jarrar et al., 2022) consists of 16817 sentences with an average sentence length of 23.45 words and a vocabulary size of 44881 for training, 3133 sentences with an average length of 17.8 and a 13134 vocabulary size for validation. The test set has 5990 sentences with an average length of 18.68 and a vocabulary size of 20920. It comprises both Modern Standard Arabic (MSA) and dialectal Arabic sentences. However, as detailed in Table 3 of Jarrar et al. (2022), the Wojood dataset is unbalanced, with the most common class being GPE (Geopolitical Entity) in the nested train set and ORG (Organization) in the flat train set, and the least frequent class being PRODUCT for Nested and UNIT for Flat.

3 NER systems

In this work, we approached the NER task with two different methods: a data-centric approach encompassing dataset re-sampling and data preprocessing and a model-centric approach using several model architectures.

3.1 Data cleaning

The data cleaning process involved several steps. Firstly, a definition of 974 stop-words was established. Then, stop words were removed from the train set. Additionally, both exclamation marks (!) and question marks (?) were removed from the text, enabling the model to focus only on sentence context. However, full-stops(.) and commas (,) were kept to avoid offsetting numerical values. Moreover, each occurrence of two or more dots were replaced with just one to maintain text clarity.

3.2 Data-centric approach

Based on the key observation of the unbalanced nature of the Wojood data-set, we decided to experiment with a data-centric approach using various re-sampling methods.

NER datasets are typically unbalanced, with an over-representation of the *Outside* <*O*> tag. To mitigate this unbalance, Wang and Wang (2022) proposed 4 different re-sampling methods in order to increase the occurrences of sentences including sequences tagged with an under-represented class. The following is an outline of these methods.

Smoothed Count (sC) re-sampling: This is the re-sampling method upon which the following are built and the most simplistic. It works by re-sampling sentences that contain the most named entities not classified as *Outside* <*O*>.

Smoothed re-sampling incorporating Count and Rareness (sCR): This method iterates on *sC* by incorporating a rarity factor. Using this technique, sentences with rarer tokens are more likely to be re-sampled. This method incorporates a rarity component in addition to the smoothed count.

Smoothed re-sampling incorporating Count, Rareness, and Density (sCRD): In order to focus on sentences that have a higher number of entity tokens per sentence length, the entity density is added in this method, while still considering the rarity of the tokens.

Normalized and Smoothed re-sampling incorporating Count, Rareness, and Density (nsCRD): This method adds a utility factor to the sCRD re-sampling function, to model the usefulness or the pertinence of a token. Thus, the more varied the tags in a sentence, the higher its marginal utility factor, and the higher its chances of re-sampling when compared to a less varied sentence.

3.3 Model-centric approach

Two models were used. A Parallel Instance Query Networks model (PIQN) (Shen et al., 2022) and a DiffusionNER model (Shen et al., 2023). PIQN extracts entities from a sentence concurrently using

a parallel approach. Each individual query instance predicts a single entity. By concurrently processing all of these query instances, multiple entities can be retrieved in parallel. This model contains three main components: the encoder module, the entity prediction module that conducts entity localization and entity classification and the dynamic label assignment module that assigns the ground truth entities to the instance queries.

DiffusionNER is a state-of-the-art model in the field of NER. It is a diffusion model (Ho et al., 2020) that adds noises spans to the ground truth entity boundaries and learns to reverse this process to reconstruct correctly the entity boundaries during training. During inference, it randomly selects noisy spans from a standard Gaussian distribution, then it generates named entities by applying a denoising operation using the acquired reverse diffusion process. For each of the two mentioned models above, several BERT (Devlin et al., 2019) encoder derivatives were used, namely MARBERT (Abdul-Mageed et al., 2021), AraBERT (Antoun et al., 2020), and BioBERT (Lee et al., 2019).

4 Experiments and results

In order to reproduce the results obtained in Jarrar et al. (2022), we started by training a multitask-based baseline model for each evaluation condition (flat and nested) using AraBERT (Antoun et al., 2020). Table 1 compare our baseline results to the ones obtained in (Jarrar et al., 2022). The scores are calculated on the *development* set provided by the shared task organizers.

Model	sub-task	Micro F1-score
(Jarrar et al., 2022)	flat	86.81
Our baseline	flat	**87.55**
(Jarrar et al., 2022)	nested	90.47
Our baseline	nested	**90.53**

Table 1: Reproducing results of Jarrar et al. (2022). Results are measured on the development set related to each version of the models (flat or nested).

As stated in section 3, we have opted for two main axes of experimentation: data-centric and model-centric approaches. These are covered in the following sections.

4.1 Dataset re-sampling

To assess the benefits of data re-sampling, we started from the baseline of the flat sub-task and re-trained a new model for each re-sampling method presented above. This results in four different new models, each trained on a resampled version of the data. As is shown in table 2, compared to the baseline, an improvement of 0.37 was obtained with *sC* re-sampling whereas a slight improvement when using *sCR* method. However, a drop in F1-score has been observed with entity density-based re-sampling methods *sCRD* and *nsCRD*.

Model	Micro F1-score
baseline	87.55
baseline with sC	**87.92**
baseline with sCR	87.62
baseline with sCRD	87.37
baseline with nsCRD	87.42

Table 2: Comparing re-sampling methods on flat Wojood. All scores are obtained using the development set of flat Wojood.

In order to assess the effectiveness of the data cleaning process, we applied the pre-processing steps described in section 3.1 and re-trained the best system from the table above (*baseline with sC*). Unfortunately, data pre-processing affects the system improvement and reduces the F1-score of the *baseline with sC* model to 80.71%. Given this, we have decided to proceed without data pre-processing.

4.2 Model fine-tuning

Regarding the model-centric approaches, we trained several PIQN and DiffusionNER models using multiple BERT encoders.

We started by applying the default PIQN configuration with *biobert-large* encoder on the nested Wojood sub-task. The obtained results were worse than the baseline, which could be explained by the language and domain mismatch. In fact, BioBERT was trained using English Wikipedia, BooksCorpus and several large-scale biomedical corpora.

To remedy this mismatch situation, we replaced the BioBERT encoder by AraBERT (Antoun et al., 2020) which is trained on Arabic and hence more suitable for the Wojood shared task. As shown in table 3, we observed that AraBERT is quite effective. Compared to the baseline, an F1 score improvement of 3.07 and 2.01 are obtained for flat and nested sub-tasks respectively.

As dataset re-sampling is shown to improve the model performance (see Table 2), we tried by training the PIQN model on top of each re-sampling method. This experiment was performed

using the flat Wojood data-set. As we can see in Table 3, all the implemented re-sampling method did not yield the desired results when used with PIQN models trained using AraBERT encoder. Given the results obtained with PIQN, we trained DiffusionNER model without any re-sampling. As listed in Table 3, fine-tuning DiffusionNER with AraBERT encoder module resulted in the highest obtained F1-scores of **91.50%** and **93.19%** for both flat and nested NER respectively.

As Wojood data is a mix of MSA and dialectal Arabic, we have also tried to train DiffusionNER using MARBERT (Abdul-Mageed et al., 2021) encoder. Since MARBERT was trained using a mixture of MSA and dialectal Arabic, we hypothesize that it could improve the NER results. Unfortunately, that was not the case since the usage of MARBERT resulted in lower F1 score compared to the DiffusionNER model trained with AraBERT encoder (Using MARBERT reduces the F1 score of the DiffusionNER model from 93.19 to 90.39).

Model	Flat micro F1-score	Nested micro F1-score
PIQN_AraBERT	90.59	92.54
PIQN_AraBERT_sC	88.98	–
PIQN_AraBERT_sCR	88.63	–
PIQN_AraBERT_sCRD	80.98	–
PIQN_AraBERT_nsCRD	90.04	–
DiffusionNER_AraBERT	**91.50**	**93.19**

Table 3: Results of fine-tuning PIQN and DiffusionNER on flat and nested Wojood development set. Best results are in bold.

All our models were implemented using PyTorch (Paszke et al., 2019) and trained on one Nvidia Quadro RTX 6000 GPU using Adam optimizer (Kingma and Ba, 2017) for a number of epochs ranging between 100 and 150 with a batch size ranging from 8 to 32.

4.3 System submissions

For both flat and nested sub-tasks, we submitted the top two performing systems on the development set.

Table 4 shows the scores obtained on the official test set. Our primary submission for nested NER was ranked first among all participants with an F1-score of **93.73%**. As for our flat NER primary submission, it was ranked third with an F1-score of 91.92%.

Model	sub-task	micro F1-score
DiffusionNER_AraBERT	nested	**93.73**
PIQN_AraBERT	nested	91.86
DiffusionNER_AraBERT	flat	**91.92**
PIQN_AraBERT	flat	90.87

Table 4: Official evaluation results. DiffusionNER and PIQN are respectively the primary and auxiliary submissions.

5 Discussion

We started our experiments by re-sampling the data set in order to alleviate the data imbalance problem. Our experiments show that not all the tested re-sampling improves the F1 score of the Wojood data-set. We also tested model-centric methods using two encoder-only models, namely PIQN and DiffusionNER. Both models have been tested using various pretrained encoders: BioBERT, AraBERT and MARBERT.

Using BioBERT has decreased the model performance. This is not surprising, given the fact that BioBERT is mainly trained on biomedical English text, which does not fit our use-case.

We also noted that the results of PIQN and DiffusionNER are better without dataset re-sampling. We observed this when we trained PIQN and DiffusionNER models after re-sampling the flat NER dataset. However, due to time constraints, we didn't run the same experiments for nested NER.

Another observation made during the experiments is that the use of an AraBERT encoder yields better results compared to MARBERT despite the presence of dialectal sentences in the latter. This can be attributed to the pretraining of MARBERT being exclusively on tweets, or to the possibility that the Wojood dataset has more MSA content than dialectal.

6 Conclusion

This paper presents results obtained on two NER sub-tasks of the Wojood shared task, namely flat and nested NER. Our submission relies on the usage of data-centric and model-centric approaches. Data-centric consists of a set of re-sampling methods intended to mitigate the unbalanced nature of Wojood data-set. Various re-sampling method are implemented and led to only limited success. Model-centric approaches, for their part, are designed to train the best model for a given dataset. We experimented with PIQN and DiffusionNER

models trained using various pre-trained encoder. Remarkably, the DiffusionNER fine-tuned with an AraBERT sentence encoder module without any re-sampling or pre-processing, yielded to significant improvements over the baseline results for both nested and flat NER. This allows us to be ranked first out of eight for the nested NER sub-task with a **93.93%** F1-score and third out of eleven for the flat NER sub-task with a **91.92%** F1-score.

Limitations

Despite the high ranking of our submitted system, there are still a number of limitations that should be mentioned and addressed in the future. Those can be summarized into the following categories:

(1) Complexity: These systems are comprised of multiple tightly coupled modules and components, which is relatively complex when compared to the baseline model. Moreover, this complexity results in higher hardware requirements in terms of GPU memory and computing power.

(2) Maximum sequence length: Both models are not able to train or perform inference on sequences longer than 512 characters tokens. A workaround consisting of splitting longer sentences into two smaller ones has been adopted for this work, but that is not an elegant solution for long-term or industrial usage.

(3) Annotated data requirements: Both systems require annotated data, which can be more costly and harder to source. Self-supervised, unsupervised or few-shot learning alternatives like the work of Das et al. (2022) should be explored in order to mitigate the need for high amounts of labelled data.

Ethics Statement

All models used for this work are open-source and publicly available. All results are reproducible, given the Wojood dataset. The authors obtained the dataset by submitting a formal request.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. 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.

Yassine Benajiba, Paolo Rosso, and José Miguel Benedíruiz. 2007a. 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.

Yassine Benajiba, Paolo Rosso, and José Miguel BenedíRuiz. 2007b. Anersys: An arabic named entity recognition system based on maximum entropy. In *Computational Linguistics and Intelligent Text Processing*, pages 143–153, Berlin, Heidelberg. Springer Berlin Heidelberg.

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Padó. 2014. Germeval 2014 named entity recognition shared task.

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 world. *Commun. ACM*, 64(4):72–81.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca Passonneau, and Rui Zhang. 2022. CONTaiNER: Few-shot named entity recognition via contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6338–6353, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mourad Gridach. 2016. 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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

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

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 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.

Basra Jehangir, Saravanan Radhakrishnan, and Rahul Agarwal. 2023. A survey on named entity recognition – datasets, tools, and methodologies. *Natural Language Processing Journal*, 3:100017.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Huyen Thi Minh Nguyen, Quyen The Ngo, Luong X Vu, Vu Mai Tran, and Hien T. Nguyen. 2019. Vlsp shared task: Named entity recognition. *Journal of Computer Science and Cybernetics*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. DiffusionNER: Boundary diffusion for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3875–3890, Toronto, Canada. Association for Computational Linguistics.

Yongliang Shen, Xiaobin Wang, Zeqi Tan, Guangwei Xu, Pengjun Xie, Fei Huang, Weiming Lu, and Yueting Zhuang. 2022. Parallel instance query network for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 947–961, Dublin, Ireland. Association for Computational Linguistics.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Tatiana Tsygankova, Stephen Mayhew, and Dan Roth. 2019. BSNLP2019 shared task submission: Multisource neural NER transfer. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 75–82, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xiaochen Wang and Yue Wang. 2022. Sentence-level resampling for named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2151–2165, Seattle, United States. Association for Computational Linguistics.