El-Kawaref at WojoodNER shared task: StagedNER for Arabic Named Entity Recognition

Nehal Elkaref
German University in Cairo
Cairo, Egypt
nehal.elkaref@student.guc.edu.eg

Mohab Elkaref
IBM Research Europe
Daresbury, United Kingdom
mohab.elkaref@ibm.com

Abstract

Named Entity Recognition (NER) is the task of identifying word-units that correspond to mentions as location, organization, person, or currency. In this shared task (Jarrar et al., 2023) we tackle flat-entity classification for Arabic, where for each word-unit a single entity should be identified. To resolve the classification problem we apply *StagedNER* as proposed by (Elkaref et al., 2023), which involves fine-tuning NER downstream tasks that divides the learning process of a transformer-model into two phases, where a model is tasked to learn sequence tags and then entity tags rather than learn both together simultaneously for an input sequence. We create an ensemble of two base models using this method that yield a score of F1 performance of 90.03% on the validation set and 91.95% on the test. The submitted model has ranked second for its F1 score, fourth in precision and ranked first scoring the highest recall.

1 Introduction

Named Entity Recognition (NER) is a vital sub-task for a plethora of NLP applications, those of which include machine translation (Ugawa et al., 2018), co-reference resolution(Clark and Manning, 2016) and information extraction (Cheng et al., 2021). The sub-task exhibits challenges when addressed from the lens of Arabic data, this comes back to the fact that the language is one of the richest in morphological inflections. To add more, attributes that typically help in locating entities such as capitalisation is not featured in the language. Arabic is also agglutinative in nature where one word could be combination of lemma, prefixes and suffixes (AbdelRahman et al., 2010) (Qu et al., 2023).

Arabic NER (ANER) has been approached using a wide spectrum of methods through the years however, more recently development of pre-trained language models (PLMs) specifically transformer-based models that learn context-aware representations has elevated the performance on ANER datasets. These models include MARBERT and ARBERT (Abdul-Mageed et al., 2021), AraBERT(Antoun et al., 2020a).

The architecture of these PLMs has been extended and equipped with different networks. To exemplify, (Al-Qurishi and Souissi, 2021) utilized a range of transformer based models namely AraBERT, XLM-Roberta (Conneau et al., 2019) and AraElectra (Antoun et al., 2020b) coupled with Conditional Random Field (CRF) to fine-tune an ANER downstream task revealing that AraBERT exhibited the highest scores. BiLSTM and BiGRU-CRF models have also been fine-tuned on Arabic BERT in an attempt to classify entities based on classical Arabic. (Alsaaran and Alrabiah, 2021). In similar vein, we leverage transformer based models to classify flat entities. However, we employ an alternative technique to fine-tuning PLMs on NER tasks where the learning regiment for a model is distributed over two stages for better learning (Elkaref et al., 2023).

In the next sections we begin by describing the data purposed for this shared task (Jarrar et al., 2023) in section 2 and highlight how we re-purposed train, validation and test sets to perform a two-staged fine-tuning process. Next, we give an extensive explanation of our adopted fine-tuning method in section 3. In sections 4 and 5 we present results of the submitted system, and discuss and analyse system performance on the validation set. Finally, we summarize and recap the proposed system and re-highlight performance scores and findings in section 6.

2 Data

Data utilised was from Wojood corpus (Jarrar et al., 2022), a rich and substantial corpus for Arabic NER that encompasses a wide range of entity types.

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 803–808

The corpus is also further extended to include annotation for nested entities, however for the scope of this shared task paper only annotations purposed for flat entities are used. The total number of tokens amounts to over 550K of Modern Standard Arabic (MSA) and dialectical Arabic tokens. To add more, MSA tokens are more frequent, where about 86% of tokens are MSA and the rest come in the Levant dialect. The corpus covers a different domains for each Arabic class; MSA tokens were acquired from two resources, the Birzeit University digital Palestinian archive, "Awraq", and online articles[1]. The former covers cultural heritage and modern history of Palestine while the latter includes web articles of health, law, finance, politics, migration, terrorism, ICT and elections. Meanwhile, dialectical tokens were obtained from supplementary Lebanese and Palestinian corpora (Haff et al., 2022) (Jarrar et al., 2014) (Jarrar et al., 2017) and other additional Levant resources collectively discussing general topics. Train, development and test splits were provided; as depicted in figure 1 Out of word vocabulary (O)

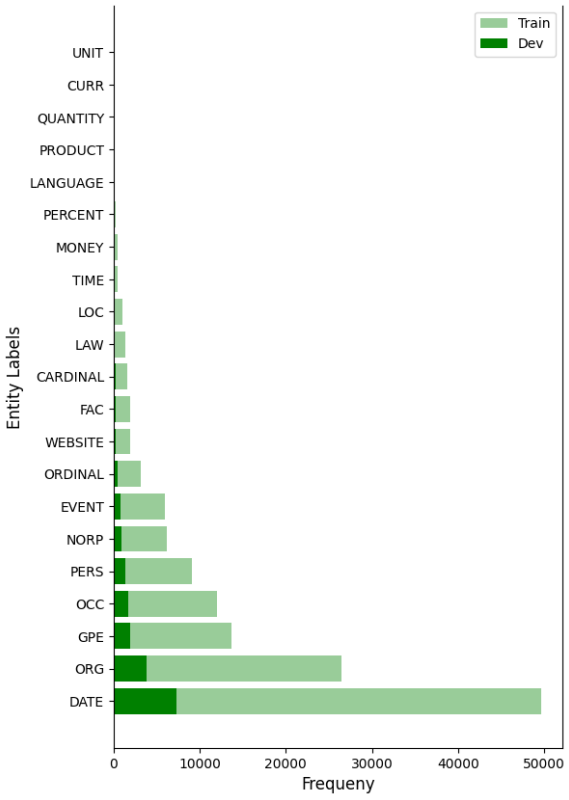

Figure 1: Entity distribution of train and validation sets of Wojood NER corpus without out-of-word vocabulary tag

[1] un.org,hrw.org,msf.org,who.org,mipa.institute,elections.ps,sa.usembassy.gov,diplo-matie.ma,quora.com

instances exceed other entity classes, where there were about 258K and 36K in the train and validation sets of O instances, for that reason figure 1 brings focus to other less dominant meaningful entities; whereby Date, Organization, Geopolitical, Occupation and Person are recurring throughout the data in comparison to Language, Product, Quantity, Currency and Unit which are rarely present.
As briefly mentioned before, the core idea of the proposed learning technique relies on separation of learning of sequence labels (BIO) tags and entity classes, hence the data goes through a separation of sequence labels and entities. Moreover, we rely on AraBERT's pre-processor[2] whereby diacritics and elongations are removed by default.

3 System Description

The backbone of the submitted system relies on fine-tuning a language model based on BERT's transformer architecture (Devlin et al., 2018). Typically, data utilised in the fine-tuning process for NER tasks follows BIO format, whereby at word-level, each entity is accompanied by an appropriate B (beginning) or I (inside), or O (outside) tag, hence the model is tasked to learn a position of an entity and the entity itself altogether. Meanwhile, we adopt the *StagedNER* approach (Elkaref et al., 2023) whereby the learning process is split into two sub-tasks, the first mimics a sequence-labelling problem where the model learns to assign appropriate BIO tags for each input, and the second sub-task is the original entity classification task. We note that this method is not analogous to sequential learning, as two separate instances of a transformer are leveraged in this method, thus each instance is assigned to exclusively learn either a BIO tag or an entity class.

Classifying BIO tags The first stage entails fine-tuning transformer on simply BIO tags of input sequences. To strengthen the transformer's learning at this stage we supply it with Part-of-Speech (POS) tags as an additional feature to help identify class spaces better, where representations from the model are pooled and summed with its appropriate POS tag.

Classifying Entity types In the second stage, a second untrained instance of the same transformer is utilized and is fine-tuned to predict

[2] https://github.com/aub-mind/arabert

entity classes. Additionally, BIO labels predicted from the first stage are passed to the model, in doing so, we ensure during entity prediction time the transformer is aware of the boundaries of entity.

Overall Framework In figure 2 we illustrate StagedNER's framework bottom to top, the input sequence is passed on to the first transformer instance, where resulting representations for sub-word tokens are summed then fed to the classification layer to predict output BIO tags. Additionally, to incorporate POS tags, they are firstly added to the tokenizer as special tokens and then inserted between input token sequences. Next, the original input sequence is given to the second transformer where once again sub-word tokens are summed. When summing sub-word tokens, BIO tags from the first stage are taken and leveraged in-order to pool vectors representing the beginning and end of an entity. The pooled vectors are finally passed onto the classification layer to predict entity types. We note that during training and validation BIO tags utilised were the GOLD BIO tags, while for the test set we relied on predicted BIO tags from the first stage transformer.

3.1 Experimental Setup

We utilise AraBERTV02 (Antoun et al., 2020a) a transformer-based language model pre-trained on a collection of Arabic corpora[3] majority of which is in MSA. POS tags are generated for train, development and test sets using CAMel Tools Part-of-Speech tagger (Obeid et al., 2020) for MSA and Levant (LEV) each exclusively. Leveraging POS tags of different classes of Arabic was motivated by the nature of that data being a mix of dialectical (Levant) and MSA. (Jarrar et al., 2022)
Two instances of AraBERT are fine-tuned according to hyperparameters mentioned in table 1. We experimented with the same range of learning rates for both AraBERT transformers but we found 5e-5 to work best along with a batch size of 8 while the same dropout rate was used consistently in all experiments. Additional information about infrastructure are given in table 2. A span of experiments is conducted to yield three models, each of which utilises either Levant or MSA POS tags or no tags at all. In doing so, we produce four ensembles using different combinations of these three models.

[3] https://huggingface.co/aubmindlab/bert-large-arabertv02#dataset

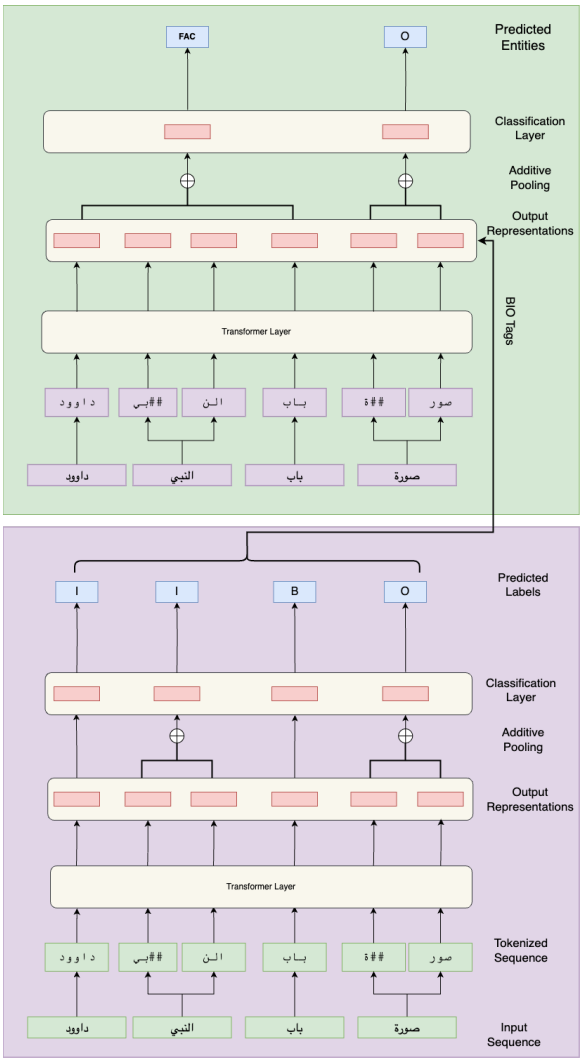

Figure 2: Bottom-up illustration of StagedNER framework, starting with BIO tag identification stage and up to entity classification stage

Hyperparameter	Value
learning rate	5e-6, 2e-6, 5e-5
dropout	0.1
epochs	8, 16

Table 1: Hyperparameter experimented with

Infrastructure	
GPU	A100 80GB
training time (mins/epoch)	11

Table 2: Infrastructure utilized

We submit four ensembles, three of which are comprised of two models that use MSA or Levant

POS tags or none at all, and a final variant that ensembles systems that uses MSA, LEV and no tags at all.

4 Results

We report micro F1, precision and recall scores for development and test sets in tables 4 and 5 for every ensemble and their unique POS combination. Additionally we showcase our performance compared to other teams in table 3 Scores show that the

Team	F1	P	R	Rank
LIPN	**91.96%**	**92.56%**	91.36%	1
ELYADATA	91.92%	91.88%	91.96%	3
Alex-U 2023 NLP	91.80%	91.61%	92.00%	4
tdink NER	91.25%	90.76%	91.73%	5
Our System	91.95%	91.43%	**92.48%**	2

Table 3: Shared task leaderboard and F1, precision and recall scores on the test set

Ensemble	DEV		
	F1	P	R
Baseline	86.81%	-	-
LEV + MSA + No POS	89.94%	88.92%	90.98%
LEV + MSA	89.95%	89.08%	90.84%
LEV + No POS	89.16%	89.90%	90.8%
MSA + No POS	**90.03%**	**88.92%**	**91.12%**

Table 4: F1, precision and recall scores for the validation set

best performing model in-terms of F1 and recall is the forth ensemble for both validation and test sets that combined two base models, the first utilised MSA POS tags while the second relied only on representations learned during training. However for precision scores, the ensemble falls behind by 0.01% to an ensemble that leverages Levant and MSA POS tags.

5 Discussion

By inspecting tables 4 and 5, we can see that ensembles that incorporated MSA POS tags has had the highest F1 scores, this is analogous with the

Ensemble	TEST		
	F1	P	R
LEV + MSA + No POS	91.88%	91.33%	92.44%
LEV + MSA	91.92%	**91.44%**	92.40%
LEV + No POS	91.78%	91.11%	92.45%
MSA + No POS	**91.95%**	91.43%	**92.48%**

Table 5: F1, precision and recall scores for test set

Arabic class distribution within the dataset, where majority of the data is curated and collected from MSA resources. We report additional F1, percision and recall below.

Entity	P	R	F1
CURR	00.00%	00.00%	00.00%
DATE	94.37%	95.21%	94.79%
EVENT	73.78%	77.87%	75.78%
FAC	72.41%	74.12%	73.26%
GPE	90.01%	91.52%	90.76%
LANGUAGE	85.71%	80.00%	82.76%
LAW	82.98%	88.64%	85.71%
LOC	71.64%	76.190%	73.85%
MONEY	95.00%	95.00%	95.00%
NORP	73.53%	79.88%	76.58%
OCC	85.89%	89.52%	87.67%
ORDINAL	95.134%	95.60%	95.36%
ORG	91.08%	93.29%	92.17%
PERCENT	00.00%	00.00%	00.00%
PERS	93.15%	96.31%	94.70%
PRODUCT	50.00%	40.00%	44.44%
QUANTITY	50.00%	66.67%	57.14%
TIME	75.00%	65.45%	69.90%
WEBSITE	54.54%	53.33%	53.93%

Table 6: Scores per entity class

By inspecting table 6, we find that ensemble had no problem classifying regularly occurring entities such as Date, GPE, ORG and PERS and managed to perform competitively on less occurring entities such as ORDINAL. The ensemble however falls behind on WEBSITE and PRODUCT. When examining instances belonging to such entities we found them to be either dialectical or even non-Arabic such as **AK** or واورو نج . This in-turn suggests re-

lying on MSA POS tags is not enough, and using a model that was not exposed to non-Arabic data during pre-training might not be the ideal choice when dealing with non MSA data, therefore stronger POS for dialectical data is required that has been trained on a diverse range of topics. Moreover, we hypothesise that a model pre-trained on dialectical data such as MARBERT if it was part of the ensemble we would have witnessed stronger results.

6 Conclusion

In this shared task, we tackled flat entity classification on Wojood corpus, a predominantly MSA dataset, where we applied an alternative fine-tuning method, where one model is used to learn BIO tags and another separate model is used to learn entity classes, instead of a single model that learns to perform both tasks jointly. The motivation behind this was to lessen the number of classes a model had to learn; where instead of learning sequence variations of one entity such as I-ORG, B-ORG, the model simply learns to identify ORG and another model is tasked to learn BIO sequence tags. To strengthen the learning of BIO tags we equip the model with MSA and Levant POS tags and created four ensembles based on different combinations of them. Results show that having MSA POS tags made a difference in performance where the best performing ensemble that include MSA POS tags scored 90.03% and 91.95% on the development and test sets respectively. Our best performing model can be demonstrated on HuggingFace Spaces[4].

References

Samir AbdelRahman, Mohamed Elarnaoty, Marwa Magdy, and Aly Fahmy. 2010. Integrated machine learning techniques for arabic named entity recognition. *IJCSI*, 7(4):27–36.

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.

Muhammad Saleh Al-Qurishi and Riad Souissi. 2021. Arabic named entity recognition using transformer-based-crf model. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 262–271.

Norah Alsaaran and Maha Alrabiah. 2021. Classical arabic named entity recognition using variant deep neural network architectures and bert. *IEEE Access*, 9:91537–91547.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020a. 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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020b. Araelectra: Pre-training text discriminators for arabic language understanding. *arXiv preprint arXiv:2012.15516*.

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

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

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

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

Mohab Elkaref, Nathan Herr, Shinnosuke Tanaka, and Geeth De Mel. 2023. NLPeople at SemEval-2023 task 2: A staged approach for multilingual named entity recognition. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1148–1153, Toronto, Canada. Association for Computational Linguistics.

Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras+ baladi: Towards a levantine corpus. *arXiv preprint arXiv:2205.09692*.

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

[4] https://huggingface.co/spaces/nehalelkaref/flat-arabic-entity-classification

Mustafa Jarrar, Nizar Habash, Diyam Akra, and Nasser Zalmout. 2014. Building a corpus for palestinian arabic: a preliminary study. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 18–27.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. Curras: an annotated corpus for the palestinian arabic dialect. *Language Resources and Evaluation*, 51:745–775.

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.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

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*.

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