AlexU-AIC at WojoodNER shared task: Sequence Labeling vs MRC and SWA for Arabic Named Entity Recognition

Shereen Elkordi
Alexandria University
es-shereen.elkordi2018@alexu.edu.eg
Applied Innovation Center
selkordi@aic.gov.eg

Noha Adly
Alexandria University
nadly@alexu.edu.eg
Applied Innovation Center
nadly@aic.gov.eg

Marwan Torki
Alexandria University
mtorki@alexu.edu.eg

Abstract

Named entity recognition (NER) is one of many challenging tasks in Arabic Natural Language Processing. It is also the base of many critical downstream tasks to help understand the source of major trends and public opinions. In this paper, we will describe our submission in the Wojood NER Shared Task of Arabic-NLP 2023. We used a simple machine reading comprehension-based technique in the Flat NER Subtask ranking eighth on the leaderboard with a 91.13% F1-score. For the Nested NER Subtask, we fine-tuned a pre-trained language model and got a 92.61% F1 score ranking third on the leaderboard.

1 Introduction

Arabic internet content has witnessed a leap in the past years which encourages the community to explore a large spectrum of tasks. Named Entity Recognition (NER) is one of the fundamental tasks that can be included in many applications. It uses semantic text features to identify names, organizations, locations, and many other mentions in a given text. This information can be used to identify social media trends (Li et al., 2022), summarize articles (Nan et al., 2021) or as a component in question answering (Mollá et al., 2006) and machine translation (Nowakowski et al., 2022).

Many techniques to solve the NER problem have emerged and can be classified into three categories: sequence labeling, span-based classification, and sequence-to-sequence generation. Sequence labeling mainly classifies the entity type of each word or token. This category has been investigated widely in high and low-resource languages (Yang et al., 2018; Katiyar and Cardie, 2018).

For the span-based models, They depend on generating all possible spans in the input and classifying each span (Yu et al., 2020). For the sequence-to-sequence models, a decoder is required to start generating the tag for each token (Zhu et al., 2020; Straková et al., 2019), or generate all found tags with their span indices (Yan et al., 2021). Apart from the mentioned categories, other methods have been proposed that include contrastive learning as in (Huang et al., 2022; Das et al., 2022; Zhang et al., 2022; Hussein et al., 2023).

Lots of challenges exist for the Arabic NER problem, i.e. the lack of a large well-annotated dataset or language-dependent problems (Shaalan, 2014). These issues may have restricted the exploration of all the mentioned techniques. The sequence labeling technique has been the most investigated (Qu et al., 2023). Many encoders have been deployed starting from recurrent neural networks till the transfer learning from pre-trained language models like AraBERT (Antoun et al., 2020) and its variants. Recently, there were attempts to explore the multitasking track as in (Jarrar et al., 2022).

In this paper, we are trying to explore the machine reading comprehension method (MRC) (Li et al., 2020) and compare it to the sequence labeling technique with a pre-trained language model as a baseline. MRC injects a prompt alongside the input text to help the model better exploit the features that will aid it in answering the prompt. The model is guided not just to perform sequence labeling but to understand the meaning behind it and maybe better generalize to uncommon cases.

We describe our submission, to the Wojood NER Shared Task (Jarrar et al., 2023), which covers using the pre-trained model JABER (Ghaddar et al., 2021) in a sequence labeling technique, and formulating the Arabic NER task as a machine reading comprehension task following (Li et al., 2020). Further, we followed (Izmailov et al., 2018) on averaging the best checkpoints of the Flat NER model producing our best result.

2 Data

For our experiments, we used the Wojood dataset (Jarrar et al., 2022) which contains 21 entity labels.

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 771–776

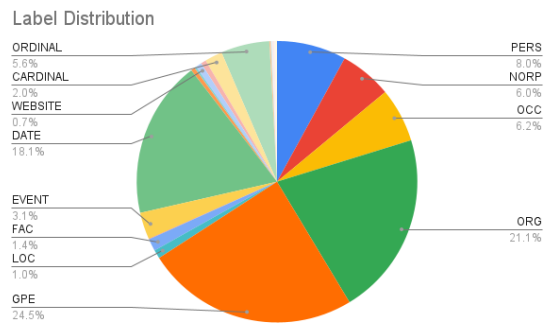

Figure 1: Label Distribution For the Nested NER Train Split

It contains three splits: train, validation, and test with 16,817, 3,133, and 5,990 sentences respectively. The train entities distribution can be found in Figure 1. To use this dataset in an MRC model, It needed some preparation. We created a new dataset where each sample includes the following fields: the context (the input text), the query (the entity type), and the start and the end positions of the answer to the queried type. These positions are indicated using the index of the start and end entity word in the context respectively.

The sizes of the dataset splits went up to 353,157, 65,793, and 125,790 since there are 21 new data samples for each sentence. We tried two different types of queries: the Arabic translation of the labels (keywords) using Google Translate [1] and the annotation guidelines of each entity as mentioned in (Jarrar et al., 2022). Examples of the dataset using the annotation guidelines queries can be shown in Table 1

For the flat NER task, we found 39,724 and 5,799 answered queries in the train and validation sets respectively. These numbers increased to 47,457 and 6,973 in the nested NER task. We can notice that the Geopolitical entity (GPE), Date, and Organization categories comprise most of the dataset with more than 11K occurrences each. In contrast, Percentage, Quantity, and Unit categories have less than 50 occurrences each.

3 System

This section will describe our two approaches: the sequence labeling technique and the problem formulation as a machine reading comprehension problem.

[1] https://translate.google.com/?sl=ar&tl=en&op=translate

3.1 Sequence labeling Models

We conducted several experiments using the same codebase as (Jarrar et al., 2022). The flat NER model is composed of a pre-trained language model with a classifier layer of 43 classes following the IOB2 scheme (I-tag and B-tag for each category + O-class). The nested NER model uses 21 parallel classification layers for each category, where the output number of classes for each layer is 3 (B, I, and O). Several backbones (PLM) were explored using the sample data provided here [2]. For training the model, we used the cross entropy loss between the predicted index and the ground truth class index.

We used AraBERTv2, AraBERTv2-Twitter (Antoun et al., 2020), MARBERT, MARBERTv2 (Abdul-Mageed et al., 2021), AraElectra (Antoun et al., 2021), CamelBERT (Inoue et al., 2021) and JABER(Ghaddar et al., 2021). The best-performing encoders were JABER followed by AraBERTv2. Therefore we used JABER for training our sequence labeling models on both flat and nested tasks. JABER (Ghaddar et al., 2021) is a pre-trained language model that uses a byte-level byte pair encoding (BBPE) with data cleaning tricks, leveraging better representation of the input text.

3.2 Machine Reading Comprehension (MRC)

We decided to explore the effect of MRC by applying the method mentioned in (Li et al., 2020). It starts by creating a query for each category in the dataset. We created 21 queries for each data sample. The model's role is to extract the answer span to the query from the context (the data sample). The input to the model is the concatenation of the query and the context.

The model consists of a pre-trained encoder followed by two binary classifiers for which a token embedding is an input. The first binary classifier detects whether the provided context token represents the start of the query answer span. The second classifier predicts if the token is the end of an answer span.

There is another binary classifier whose role is to predict whether a token i and a token j from the same sentence can represent an answer span (start and end respectively). This is to match the end index with its start in case multiple start and end indices are found for the same query. This classifier

[2] https://github.com/SinaLab/ArabicNER

Query	Start Position	End Position
Geopolitical like countries, cities, and states الجيوسياسية مثل البلدان والمدن والدول	[6, 13]	[7, 13]
Legal or social bodies like institutions, companies, agencies, teams, parties, armies, and governments. الهيئات القانونية أو الاجتماعية مثل المؤسسات والشركات والوكالات والفرق والأحزاب والجيوش والحكومات	[1]	[4]

Table 1: Example of data samples for the context: Message from the Makassed Islamic Charity Association in Jerusalem to the Acting Prime Minister in Jerusalem.
رسالة جمعية المقاصد الخيرية الإسلامية في مدينة القدس إلى رئيس الوزراء بالوكالة في القدس.
13 12 11 10 9 8 7 6 5 4 3 2 1 0

works with the two binary classifiers to filter the spans and produce the answer.

The ground truth labels consist of two lists of length N and a matrix of size NxN, where N is the number of tokens. The first list indicates if the token is the start of an answer span, while the other indicates the end. The matrix entry indicates if the token i and a token j is an answer span. The model is trained using the binary cross entropy loss.

3.3 Stochastic Weighted Average (SWA)

To improve the results, we adopted the technique mentioned in (Izmailov et al., 2018). They show that averaging multiple checkpoints of the model can improve the performance. Due to the large size of the created dataset, this choice is more convenient than an ensemble. It leads to a better usage of the computational power and decreases the inference time. Hence, we averaged the weights of the best four checkpoints of the MRC model in the flat NER Subtask.

3.4 Model Evaluation and Post processing

For the flat model inference, each sentence will be queried for every tag. The answer is returned as a list of start and a list of end positions. The answers for all 21 queries are gathered so that each word is given only one tag with the IOB2 scheme. We face a challenge here where there could be words that are included in many answer spans i.e. given two or more different labels. This can be summarized in three cases:

1. A word given B-tag1 and B-tag2
2. A word given I-tag1 and B-tag2
3. A word given I-tag1 and I-tag2

We solve this problem for the flat NER by assigning priorities to labels. These priorities are based on the frequency of the label in the training set. The more the label exists in the train set, the higher priority it gets (we are counting the B-tags only). We also make sure that the label of the word matches that of its previous in the case of I-tags. In this way, the longest named entity streak is preserved and the priority selection happens mainly in case of conflicting B-tags only.

3.5 Training Details

All models were trained on a V100 GPU. For the submitted nested model we used JABER encoder in the sequence labeling technique with a batch size of 8, a learning rate of 1e-5, and a maximum sequence length of 512. The model achieves its best result at epoch 40 and is trained for 24 hours. As For MRC models in the tasks, several experiments were done while varying the learning rate between 3e-5, 3e-6, and 2e-5. We also tried using a maximum sequence length of 200 and 256.

For the submitted flat model, we used an AraBERT-based MRC model that is trained with a batch size of 10, a learning rate of 3e-5, and a 256 maximum sequence length. The model stabilizes at epoch 10 and is trained for 48 hours. Our implementation is based on the MRC official code[3].

4 Results

We started with the sequence labeling technique in both tasks. The results with JABER on the validation set are higher in both flat and nested tasks hence we used them as our first test submission.

We tried to enhance the results by employing the MRC technique. We tried the two backbones AraBERT and JABER for both tasks. In the flat NER task, the results improved, unlike the nested

[3] https://github.com/ShannonAI/mrc-for-flat-nested-ner/

task. To further improve the results we tried performing the SWA technique which gave us the best results on the flat NER task. A table of the conducted experiments and results can be shown in Table 2

	F1-Score	Precision	Recall
	Flat NER Subtask		
Seq. Lab. (AraBERT)	0.8688	0.8558	0.8822
Seq. Lab. (JABER)	0.9052	0.90	0.9106
MRC (AraBERT)	0.9065	0.9192	0.8942
MRC (JABER)	0.9086	0.9207	0.8969
MRC (AraBERT) + keywords	0.9038	0.9208	0.8875
MRC (JABER) + keywords	0.9037	0.9249	0.8836
MRC (AraBERT) + SWA	**0.9113**	0.9133	0.9092
MRC (JABER) + SWA	0.9095	0.9152	0.9039
	Nested NER Subtask		
Seq. Lab. (AraBERT)	0.8929	0.8832	0.9028
Seq. Lab. (JABER)	**0.9261**	0.921	0.9313
MRC (AraBERT)	0.9124	0.9214	0.9036
MRC (JABER)	0.9203	0.926	0.9146
MRC (AraBERT) + keywords	0.9177	0.9188	0.9167
MRC (JABER) + keywords	0.9138	0.9241	0.9039
MRC (JABER) + SWA	0.9219	0.9226	0.9212

Table 2: Results on the test set using Sequence labeling and MRC techniques Associated with SWA.

5 Discussion

By inspecting the model performance on the validation set. We found that the flat and nested models perform poorly in the quantity, website and product classes. This is due to the insufficient number of data samples as well as the inconsistency in the annotations. An example for the inconsistency: 'Vodafone Cash and Orange Cash', these are two equivalent entities but the ground truth label for 'Vodafone Cash' is Organisation while the label for 'Orange' is Product.

For the flat NER task, the two best-performing models are MRC (AraBERT) and MRC(JABER) with stochastic weighted averaging. We analyzed the output to find the cases mentioned in Section 3.4. We found 100 words with different B-tag and I-tag labels amongst them 51 words with different I-tag-only labels and 12 words with different B-tag-only labels in the AraBERT-based model. An example of the B-tag confusion is the word 'Google' where it is assigned the labels B-ORG and B-WEBSITE. The JABER-based model has 163 words with conflicting B-tag and I-tag labels, amongst them 68 with conflicting I-tags only and 38 with conflicting B-tags only.

We wanted to analyze the efficiency of our priority-based selection scheme. We compared it with choosing randomly the B-tag label amongst the conflicting ones. We conduct 5 runs, calculate the validation F1-score at each time, and average them. For the AraBERT-based model, we find the priority scheme to score 0.90642 and the random scheme to score 0.90675. For the JABER-based model, the priority scheme produces 0.90173 while the random scheme scores 0.90155.

We notice that the more confusion in the model output, the more the random scheme fails. The first model had 12 conflicting B-tag words while the second had 38. Hence, to ensure determinism and reproducibility, we decided to follow the priority scheme. As a plan, we can choose a better scheme that would keep the model confidence scores for all 21 inferences for the sentence and compare conflicting ones to choose the B-tag with the highest score.

The Flat NER results show that the effect of adding SWA to the AraBERT-based MRC model is greater than adding it to the JABER-based model. We investigated the F1 score of each class for all the checkpoints involved in SWA. For the JABER-based models, no checkpoint could have enhanced greatly the scores of the best checkpoint.

On the other hand, other checkpoints included in the AraBERT-SWA model perform better in the cardinal, GPE, money, time, and website classes which corrected the labels on 32 samples. Meanwhile, there was a slight degradation in language, law, location, occupation, product, and quantity classes which yielded the mislabeling of 9 samples. The degradation is not effective though due to the sparsity of these classes in the dataset. In total, there was an improvement in the performance over the best checkpoint.

6 Conclusion

Arabic NER has been an underexplored problem, the lack of a large dataset can be one of the reasons. In this work, we investigate the effect of applying the machine reading comprehension technique to the Arabic NER problem. We tried two different types of prompts and concluded that the label description is more beneficial than inserting keywords as queries. We compared MRC and the sequence labeling technique. We also investigated the effectiveness of applying the stochastic weighted averaging technique. We found that the results are comparable between the sequence labeling and MRC and either of them can be used in NER. Many other methods still exist and can be tackled and finetuned for Arabic usage.

7 Limitations

MRC suffers from low scalability and long inference time. For every sentence, the required number of inferences is equal to the number of categories in the dataset. Also, the created training dataset is very sparse, many queries have no answer. Future trials can include training with a balanced set of answered and unanswered queries.

Moreover, another limitation that would affect the model performance is the absence of a considerable amount of samples for some of the classes in the dataset, i.e. the Unit class. There is no occurrence of this class in the Flat validation set which makes us unable to judge the model performance.

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. *arXiv preprint arXiv:2003.00104*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

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.

Abbas Ghaddar, Yimeng Wu, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao Xing, Yasheng Wang, Duan Xinyu, Zhefeng Wang, Baoxing Huai, et al. 2021. Jaber and saber: Junior and senior arabic bert. *arXiv preprint arXiv:2112.04329*.

Yucheng Huang, Kai He, Yige Wang, Xianli Zhang, Tieliang Gong, Rui Mao, and Chen Li. 2022. COPNER: Contrastive learning with prompt guiding for few-shot named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2515–2527, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

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)*.

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.

Arzoo Katiyar and Claire Cardie. 2018. Nested named entity recognition revisited. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 861–871, New Orleans, Louisiana. Association for Computational Linguistics.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Diego Mollá, Menno Van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Proceedings of the Australasian language technology workshop 2006*, pages 51–58.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. *arXiv preprint arXiv:2102.09130*.

Artur Nowakowski, Gabriela Pałka, Kamil Guttmann, and Mikołaj Pokrywka. 2022. Adam mickiewicz university at wmt 2022: Ner-assisted and quality-aware neural machine translation. *arXiv preprint arXiv:2209.02962*.

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*.

Khaled Shaalan. 2014. A survey of arabic named entity recognition and classification. *Computational Linguistics*, 40(2):469–510.

Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural architectures for nested NER through linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Florence, Italy. Association for Computational Linguistics.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various NER subtasks. 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 5808–5822, Online. Association for Computational Linguistics.

Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3879–3889, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. Association for Computational Linguistics.

Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. 2022. Optimizing bi-encoder for named entity recognition via contrastive learning. *arXiv preprint arXiv:2208.14565*.

Huiming Zhu, Chunhui He, Yang Fang, and Weidong Xiao. 2020. Fine grained named entity recognition via seq2seq framework. *IEEE Access*, 8:53953–53961.