Alex-U 2023 NLP at WojoodNER shared task: AraBINDER (Bi-Encoder for Arabic Named Entity Recognition)

Mariam Hussein[*], **Sarah Khaled**[*], **Marwan Torki** and **Nagwa El-Makky**
Computer and Systems Engineering Department
Alexandria University
es-mariam99.mf, es-sara.khaled2019, mtorki, nagwamakky@alexu.edu.eg

Abstract

Named Entity Recognition (NER) is a crucial task in natural language processing that facilitates the extraction of vital information from text. However, NER for Arabic presents a significant challenge due to the language's unique characteristics. In this paper, we introduce AraBINDER, our submission to the Wojood NER Shared Task 2023 (ArabicNLP 2023). The shared task comprises two sub-tasks: sub-task 1 focuses on Flat NER, while sub-task 2 centers on Nested NER. We have participated in both sub-tasks. The Bi-Encoder has proven its efficiency for NER in English. We employ AraBINDER (Arabic Bi-Encoder for Named Entity Recognition), which uses the power of two transformer encoders and employs contrastive learning to map candidate text spans and entity types into the same vector representation space. This approach frames NER as a representation learning problem that maximizes the similarity between the vector representations of an entity mention and its type. Our experiments reveal that AraBINDER achieves a micro F-1 score of 0.918 for Flat NER and 0.9 for Nested NER on the Wojood dataset.

1 Introduction

Named Entity Recognition (NER) is a fundamental task in natural language processing that involves identifying and classifying named entities, such as person names, locations, organizations, and temporal expressions, within text. In recent years, deep learning models, particularly transformer-based architectures (Hanslo, 2022), have revolutionized NER by capturing contextual information effectively. However, applying these models to Arabic NER presents several difficulties. One of the major challenges is the lack of comprehensive and annotated Arabic NER data, which hinders the fair evaluation of Arabic NER models (Qu et al., 2023). Other previous work addressed Named entity Recognition as a Sequence labeling problem (Affi and Latiri, 2022), Span-based classification (Yu et al., 2020) or Seq-to-seq generation (Wang et al., 2019). There have been some approaches that dealt with the problem as a machine reading comprehension problem (MRC) (Li et al., 2020)(Elkordi et al., 2023). Meanwhile BINDER (Zhang et al., 2022) deals with NER as a representation learning problem that maximizes the similarity between the vector representations of an entity mention and its type. This makes it easy to handle Nested and Flat NER alike, and can better leverage noisy self-supervision signals. Moreover, it demonstrates superiority over past approaches in terms of speed and efficiency.

The use of dual networks dates back to (Bromley et al., 1993) for signature verification and (Chopra et al., 2005) for face verification. Moreover, (Humeau et al., 2019) conducted a comparison of three distinct architectures Bi-Encoder, Poly-Encoder, and Cross-Encoder all employing deep pre-trained transformers as encoders. In our solution, we use the Bi-Encoder architecture that has also been used in various tasks, such as information retrieval (Gillick et al., 2018), open-domain question answering (Karpukhin et al., 2020), and entity linking (Wu et al., 2020) and proved to achieve state of the art results.

Furthermore, in recent work, all tokens or spans that do not represent entities (non-entities) were categorized under a single class called "Outside" (O). Notably, our solution diverges from this conventional method since we use the proposed dynamic thresholding loss within the context of contrastive learning. This approach involves learning dynamic thresholds specific to candidates, aiding in the differentiation of entity spans from non-entity ones. While contrastive learning (CL) has considerably advanced numerous natural language processing (NLP) tasks, its application within the Arabic con-

[*] Equal contribution

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 797–802

text has been somewhat limited(Qu et al., 2023). Recently, There has been a focus on this area as (Shapiro et al., 2022) demonstrated the efficacy of CL for Arabic hate speech detection, resulting in significant improvements over baselines. In a similar vein, (Abdul-Mageed and Lakshmanan, 2022) conducted experiments applying CL to diverse Arabic NLP tasks including dialect identification, emotion classification, sarcasm detection, and the identification of abusive and adult content.

In this paper, we bridge the gap by introducing AraBINDER, a novel approach to address these challenges. BINDER (Zhang et al., 2022), learns to differentiate between entities and non-entities, even when confronted with limited annotated data. This capability enhances its generalization potential, rendering it applicable to both Nested and Flat NER paradigms. For Sub-task 1 and Sub-task 2, we apply AraBINDER using our best model achieving micro F1 scores of 0.918 and 0.90 respectively.

2 Data

We conducted our work on the Wojood (Jarrar et al., 2022) dataset provided by the shared task (Jarrar et al., 2023). The shared task focuses on identifying named entity mentions in unstructured text and classifying them into predefined classes this is divided into two sub-tasks, sub-task 1 focuses on Flat NER while sub-task 2 centers on Nested NER. The data for sub-task 2 differed in the manner that in the Nested scheme some tokens had more than one entity type assigned to it.

The corpus of Wojood consists of about 27K sentences and 550K tokens and is manually annotated covering both Modern Standard Arabic (MSA) and Dialect Arabic (DA) in multiple domains. It contains about 75K entities, out of which 22.5% are Nested. The data was annotated for 21 entity types with IOB tags. The dataset introduced four new tags which are occupation, website, unit, and currency.

We follow the data split provided by the shared task: 70% of the data for training, 10% for development, and 20% for testing. Table 1 shows the label distribution of both Flat and Nested entities of the training and development sets. Since the provided data was IOB tagged, we have modified it by removing the tags and labeling each sentence with a unique ID. Also, the model uses the start and end of each span to modify the loss objective which is explained further on the paper for this purpose we extract the word's start and end characters for each sentence along with the start and end characters for entities in that sentence.

Tags	%Train	%Validation
PERS	8	8.24
NORP	6.01	5.87
OCC	6.23	6.22
ORG	21.12	21.11
GPE	24.52	24.43
LOC	0.99	0.86
FAC	1.41	1.25
PRODUCT	0.06	0.06
EVENT	3.1	3.02
DATE	18.1	18.7
TIME	0.46	0.62
LANGUAGE	0.21	0.17
WEBSITE	0.7	0.51
LAW	0.6	0.5
CARDINAL	2.02	2.07
ORDINAL	5.59	5.69
PERCENT	0.17	0.15
QUANTITY	0.07	0.03
UNIT	0.08	0.03
MONEY	0.27	0.23
CURR	0.29	0.24

Table 1: The distribution for Entity types in the train and validation sets of Wojood.

3 Method

In this section, we introduce the methodology of AraBINDER, which utilizes the Bi-Encoder architecture first introduced in (Zhang et al., 2022) for Arabic-named entity recognition (NER). The foundation of our model is the Bi-Encoder framework, which involves encoding both entity types and text using the Transformer-based architecture. To provide a comprehensive understanding, we begin by explaining the background of this Bi-Encoder framework. By leveraging the Bi-Encoder architecture and incorporating contrastive learning objectives, AraBINDER presents a robust and effective approach for Arabic NER. In the following sections, we will elaborate on the implementation details and experimental results to validate the performance of our proposed methodology.

3.1 Bi-Encoder for NER

The architecture of AraBINDER, as depicted in Figure 1, is based on a Bi-Encoder framework that has primarily been explored in the context of dense retrieval(Karpukhin et al., 2020). It has been put to the test in the case of NER for English and Chinese languages and demonstrated superior performance so we employ it to the Arabic language. The Bi-Encoder comprises two Transformer models, namely the entity type encoder and the text encoder, which are isomorphic and fully decoupled. For the task of NER, our model takes two types of

inputs: entity-type descriptions and text containing potential named entities. At a high level, the entity type encoder generates representations for each entity of interest (e.g., "person" in Figure 1), while the text encoder produces representations for each input token in the given text where named entities may appear (e.g., "ميركل" in Figure 1). Based on these representations, we generate a set of span candidates and match them with each entity type in the vector space. As illustrated in Figure 1, the model aims to maximize the similarity between the entity type and positive spans while minimizing the similarity with negative spans. The introduction of Bidirectional Encoder Representation from Transformers (BERT) (Kenton and Toutanova, 2019) led to a revolution in the NLP world, as BERT-based models achieved state-of-the-art results in many tasks such as Machine Translation (Ghazvininejad et al., 2019), Question Answering (Yang et al., 2019), Text summarization (Zhang et al., 2019) and many more tasks. We utilize a pre-trained language model and fine-tune it for our NER task. In our experiments, we experimented with several pre-trained BERT-based models on Arabic such as CAMeLBERT (Inoue et al., 2021) and AraBERT with both versions (Antoun et al., 2020), but AraBERTv2 produced better results so we continued our experiments using it. AraBERTv2 has two different models that differ in the training data whereas the second one contains the same training data but in addition to 60M Multi-Dialect Tweets, most of its training data is MSA instead of DA. They also use Farasa (Darwish and Mubarak, 2016) Arabic morphological segmentation in the text preprocessing and we believe that this is beneficial to our task at hand based on the nature of the provided data, which contained some MSA.

3.2 Contrastive Learning

The primary goal of NER contrastive learning, illustrated in Figure 1, is to bring the representations of entity mention spans near their corresponding entity type embeddings (positive instances) and distant from irrelevant types (negative) in vector space. For instance, we aim to position the entity type "Person" closer to the mentioned span "ميركل " while maintaining a notable distance from any other word.

To accomplish this, we applied the multi-objective formulation in (Zhang et al., 2022) that comprises two distinct objectives based on the span and token

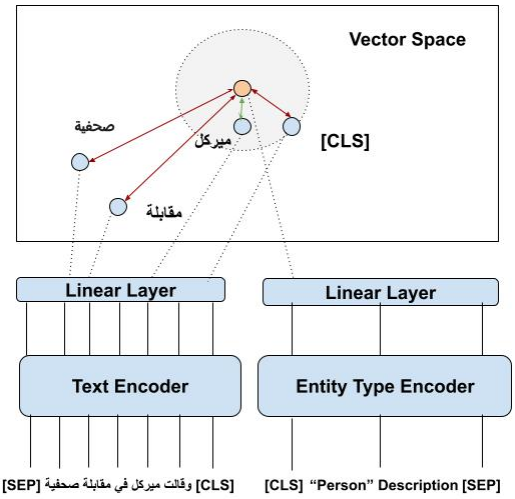

Figure 1: AraBINDER Architecture.

embedding spaces, respectively. These objectives work together to guide the model in learning meaningful representations that capture the relationships between entity types and their associated mentions, enabling accurate and effective NER. Recognizing that the span-based objective in isolation might fall short, we enhance it with a position-based objective. The latter addresses a limitation where all negative spans receive equal penalties, even if they partially correspond to correct spans, for example, spans that share a common start or end token with the gold entity span. To address the challenge of predicting partially accurate spans, we introduce supplementary position-based contrastive learning objectives, which have the potential to enhance the model's ability to predict start and end positions more accurately.

In the case of handling non-entities, the model, using the previously mentioned objectives may be able to distinguish between entities of different types, but it may fail to push away from non-entities to address this issue, we use the similarity between the special token [CLS] and the entity type as a dynamic threshold, as shown in Figure 1. Intuitively, the representation of [CLS] reads the entire input text and summarizes the contextual information, which could make it a good choice to estimate the threshold to separate entity spans from non-entity spans. In simpler terms, the final equation for the loss in Eq. (1) consists of three main parts, start loss, end loss, and span loss following the overall training objective in (Zhang et al., 2022). The equations of the three loss functions are given in (Zhang et al., 2022) and are not included here due to space

limitations.

$$L = \alpha \ell_{\text{start}} + \gamma \ell_{\text{end}} + \lambda \ell_{\text{span}} \quad (1)$$

4 Experiments

4.1 Experiment Setting

All experiments were conducted using a single v100 GPU. We utilized the given training dataset for training our model and exploited the validation dataset to choose the hyper-parameters. A maximum input sequence length is set to 128, sequences greater than this length would be truncated and sequences less than this length would be padded to obtain the same length. For all experiments, we ignore sentence boundaries and tokenize and split text into sequences with a stride of 16. All base models are trained for 20 epochs with a learning rate of 3e-5 and a batch size of 8 sequences with a maximum token length of 128. For evaluation We follow the standard evaluation protocol and use micro F1, which indicates that a predicted entity span is considered correct if its span boundaries and the predicted entity type are both correct, we also include precision and recall in our results.

5 Results

In all our experiments, we exploit the AraBERTv2-Twitter base that is trained on MSA in addition to Multi-Dialect Arabic Tweets, since our data contain both MSA and DA in multiple domains. We demonstrate our results on the development set for Flat NER and Nested NER in Table 2 and Table 3 respectively, while Table 4 and Table 5 show Flat NER results and Nested NER results on the test set respectively.

6 Discussion

As can be shown from tables 2 and 3, the model performs better for Nested NER than for Flat NER on the development set. We noticed this behavior in several experiments. However, as can be seen from tables 4 and 5, it performs better for Flat NER on the test set, This indicates that the model may have failed to generalize. In the path forward, our focus will revolve around enhancing the performance of underperforming Nested experiments while delving into the exploration of alternative encoders for Arabic, such as JABER (Ghaddar et al., 2022), which could potentially enhance our results. Moreover, we are dedicated to further refine our data pre-processing strategies to tackle the unique challenges posed by Arabic, rectifying annotation errors, and addressing the scarcity of precise data.

Model	Recall	Precision	F1
AraBINDER(ours)	0.918	0.913	0.916

Table 2: Results of Flat NER on the development set.

Model	Recall	Precision	F1
AraBINDER(ours)	0.94	0.918	0.929

Table 3: Results of Nested NER on the development set.

Model	Recall	Precision	F1
(Jarrar et al., 2022)	-	-	87.33
AraBINDER(ours)	0.924	0.914	0.918

Table 4: Results of Flat NER on the test set.

Model	Recall	Precision	F1
(Jarrar et al., 2022)	-	-	0.91
AraBINDER(ours)	0.906	0.893	0.90

Table 5: Results of Nested NER on the test set.

7 Conclusion

In this paper, our approach revolves around the application of AraBINDER to tackle both Flat and Nested Named Entity Recognition (NER) tasks within the shared context. This methodology involves using a Bi-Encoder architecture, proficiently encoding both entity types and textual content. The infusion of contrastive learning into this framework serves to maximize the similarity between individual entity types and their corresponding mention spans.

Our evaluation revolved around BERT-based models trained on Arabic corpora, with a special focus on AraBERT. Through assessment, we observed that the AraBERTv2-Twitter base, pretrained on Arabic data encompassing Modern Standard Arabic (MSA) and Twitter data, performed the best. Notably, it performed better for the Flat NER task, outperforming its Nested NER counterpart.

Limitations

As shown during our experiments, The Nested NER results were not as good as expected and we believe that most of the mistakes were due to the challenge in the nature of Arabic data, and this is a problem for low-resource languages. We notice that some words that use conjunctions as person

names may be confused with team names as in news reporting. For instance, "John and Johns" are two names for separate persons in English, while in Arabic, we find that the "و" is often linked to the following name "محمود واحمد" since there is no clear separation between them. This can be classified, at inference time, as a single-person entity with first and last names, instead of two separate person entities and this may lead to confusion.

References

Muhammad Abdul-Mageed and Laks VS Lakshmanan. 2022. A benchmark study of contrastive learning for arabic social meaning. *WANLP 2022*, page 63.

Manel Affi and Chiraz Latiri. 2022. Arabic named entity recognition using variant deep neural network architectures and combinatorial feature embedding based on cnn, lstm and bert. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 302–312.

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.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE.

Kareem Darwish and Hamdy Mubarak. 2016. Farasa: A new fast and accurate arabic word segmenter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1070–1074.

Shereen Elkordi, Noha Adly, and Marwan Torki. 2023. Alexu-aic at wojoodner shared task: Sequence labeling vs mrc and swa for arabic named entity recognition. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Abbas Ghaddar, Yimeng Wu, Sunyam Bagga, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao Xing, Yasheng Wang, Xinyu Duan, Zhefeng Wang, et al. 2022. Revisiting pre-trained language models and their evaluation for arabic natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3135–3151.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6112–6121.

Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-end retrieval in continuous space. *arXiv preprint arXiv:1811.08008*.

Ridewaan Hanslo. 2022. Deep learning transformer architecture for named-entity recognition on low-resourced languages: State of the art results. In *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 53–60. IEEE.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

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.

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.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

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.

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*.

Ahmad Shapiro, Ayman Khalafallah, and Marwan Torki. 2022. Alexu-aic at arabic hate speech 2022: Contrast to classify. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 200–208.

Yu Wang, Yun Li, Ziye Zhu, Bin Xia, and Zheng Liu. 2019. Sc-ner: A sequence-to-sequence model with sentence classification for named entity recognition. In *Advances in Knowledge Discovery and Data Mining: 23rd Pacific-Asia Conference, PAKDD 2019, Macau, China, April 14-17, 2019, Proceedings, Part I 23*, pages 198–209. Springer.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *NAACL HLT 2019*, page 72.

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.

Haoyu Zhang, Jingjing Cai, Jianjun Xu, and Ji Wang. 2019. Pretraining-based natural language generation for text summarization. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics.

Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. 2022. Optimizing bi-encoder for named entity recognition via contrastive learning. In *The Eleventh International Conference on Learning Representations*.