Bangor University at WojoodNER 2024: Advancing Arabic Named Entity Recognition with CAMeLBERT-Mix

Norah Alshammari, William Teahan
Bangor University
{nrl23czs, w.j.teahan}@bangor.ac.uk

Abstract

This paper describes the approach and results of Bangor University's participation in the WojoodNER 2024 shared task, specifically for Subtask-1: Closed-Track Flat Fine-Grain NER. We present a system utilizing a transformer-based model called bert-base-arabic-camelbert-mix, fine-tuned on the Wojood-Fine corpus. A key enhancement to our approach involves adding a linear layer on top of the bert-base-arabic-camelbert-mix to classify each token into one of 51 different entity types and subtypes, as well as the 'O' label for non-entity tokens. This linear layer effectively maps the contextualized embeddings produced by BERT to the desired output labels, addressing the complex challenges of fine-grained Arabic NER. The system achieved competitive results in precision, recall, and F1 scores, thereby contributing significant insights into the application of transformers in Arabic NER tasks.

1 Introduction

Named Entity Recognition (NER) is a crucial task in Natural Language Processing (NLP) aimed at identifying and categorizing named entities in text. The Arabic language, with its rich morphology and variety of dialects, poses significant challenges for NER, necessitating tailored solutions. The WojoodNER 2024 shared task, as established by Jarrar et al. (2024), seeks to advance Arabic NER by emphasizing fine-grained entity recognition. This year's challenge features the newly introduced Wojood-Fine corpus, which offers more detailed annotations and an expanded range of entity types compared to previous versions.

Bangor University's response to Subtask-1 of this challenge involves a strategic use of the bert-base-arabic-camelbert-mix, a transformer-based model specifically optimized for Arabic. This model is fine-tuned on the Wojood-Fine corpus to handle the intricacies of fine-grained NER. Unique to our approach is the integration of a linear classification layer atop the pre-trained model, enhancing its ability to classify each token into one of 51 different entity types and subtypes, as well as the non-entity 'O' label.

Our methodological advancements are driven by the need to address specific challenges posed by the Arabic script's absence of capitalization, its morphological diversity, and contextual ambiguities. The approach adopted for the Wojood-Fine corpus is meticulously designed to improve model generalization across varied Arabic texts, thereby enhancing performance metrics such as precision, recall, and F1 scores.

The paper further discusses specific challenges encountered during the task, such as managing the complex nature of fine-grained entities and optimizing the model to achieve robust performance across diverse dialects and textual genres. Through this participation, Bangor University not only contributes to advancing the state of Arabic NER but also offers insights that are valuable to the broader NLP community, showcasing effective strategies to tackle the linguistic complexities of Arabic NER.

2 Related Work

NER has significantly advanced from identifying simple flat entities to complex nested entities, particularly in Arabic, a language rich with script and dialect variations. Historically limited by the scope of available corpora, early Arabic NER datasets like ANERCorp (Benajiba et al., 2007) and ACE2005 (Walker et al., 2005) were focused on Modern Standard Arabic from news and political domains, offering primarily flat annotations. The introduction of the CANER Corpus (Salah and Zakaria, 2018) expanded to 14 entity types in

Proceedings of The Second Arabic Natural Language Processing Conference, pages 880–884
August 16, 2024 ©2024 Association for Computational Linguistics

Classical Arabic, enriching the context with religious annotations.

Recent progress includes the Wojood corpus by Jarrar et al. (2022), which supports both flat and nested entities across various Arabic forms, marking a significant enhancement in Arabic NER datasets with its diverse linguistic coverage of about 550K tokens. The evolution from rule-based to machine learning and deep learning models has facilitated broader generalization capabilities without intensive manual input.

Shared tasks such as MultiCoNER and HIPE-2022 have been crucial in driving NER forward by setting benchmarks and fostering innovation. The WojoodNER series, especially the 2023 edition outlined by Jarrar et al. (2023), introduced a platform for comprehensive exploration of both flat and nested NER in Arabic, establishing a new paradigm in Arabic linguistic analysis.

3 Data

The WojoodNER 2024 shared task utilizes the Wojood-Fine corpus introduced by Liqreina et al. (2023), a robust dataset tailored for Arabic NER challenges. The dataset is structured to enhance the granularity and scope of entity recognition within Arabic text, catering to a fine-grained classification scheme that extends beyond traditional NER datasets.

3.1 Dataset Composition

Wojood-Fine is meticulously structured into 70% training, 10% validation, and 20% testing splits, aligning with standard practices for machine learning dataset preparation. This division ensures that models are trained on a substantial portion of the data, validated for generalization on a smaller segment, and finally tested to evaluate model performance under controlled, unbiased conditions.

- Training Dataset: This comprises 70% of the total dataset, used for training the models. This portion is critical for learning the varied entity representations within the data.
- Validation Dataset: This accounts for 10% of the data. This subset is crucial for tuning model parameters and preventing overfitting by providing a platform for periodic evaluation during the training process.
- Testing Dataset: This makes up 20% of the data, used solely for final model evaluation. This part does not include ground truth labels as it is used in a blind test scenario to ensure the impartial assessment of the model's performance.

3.2 Entity Classes

The dataset defines a comprehensive set of entity classes, which include both singular (B- beginning) and continuation (I- inside) tags for entities that span multiple tokens. The classes encompass a wide range of entity types such as cardinal numbers, organizations, dates, languages, nationalities, persons, and more. The distribution of entities in each data split is described in detail in the shared task paper by Jarrar et al. (2024).

3.3 Challenges and Innovations

The introduction of the Wojood-Fine corpus presents several challenges and opportunities:

- Granularity: The fine-grained nature of the entity types demands precise model tuning and sophisticated feature extraction techniques.
- Data Sparsity: Given the detailed classification, certain classes may have fewer examples, posing challenges in model training due to imbalanced data.
- Contextual Ambiguity: The presence of nested and overlapping entities adds complexity, requiring advanced parsing strategies to discern boundaries and relationships between entities.

The dataset's structure and the task's setup aim to push the boundaries of Arabic NLP by addressing these challenges, promoting the development of models that are robust, accurate, and capable of understanding the nuanced aspects of Arabic syntax and semantics.

4 Our Model

In our approach to the NER Shared Task 2024 Subtask-1 (Closed-Track Flat Fine-Grain NER), we utilized a transformer-based model, specifically the CAMeL-Lab's bert-base-arabic-camelbert-mix, which was introduced by Inoue et al. (2021) due to its robust pre-training on a diverse Arabic language corpus. This choice was motivated by the model's proven capability in understanding the complexities and nuances of the Arabic language, which is crucial for the fine-grained named entity recognition required in this task.

4.1 Model Architecture

The core of our NER system is the transformer model bert-base-arabic-camelbert-mix, which employs the BERT (Bidirectional Encoder Representations from Transformers) architecture. This architecture is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers. The bidirectional nature of BERT allows it to capture context from both directions, enhancing its understanding of the syntactic and semantic nuances of the Arabic language.

The model was fine-tuned on the Wojood-Fine dataset, which provides fine-grained entity annotations across a spectrum of categories. The fine-tuning process involved modifying the top layer of the model to predict these 51 different entity types and subtypes based on the dataset's labeling scheme. Each token fed into the model could be associated with one of these labels or a special 'O' label indicating no entity. The training process leveraged the pre-trained weights of the BERT model, allowing it to learn task-specific features effectively while maintaining the general language understanding it acquired during pre-training.

4.2 Preprocessing and Tokenization

We employed the accompanying tokenizer from bert-base-arabic-camelbert-mix, which effectively handles the intricacies of Arabic scripts. This tokenizer divides text into subwords, maintaining essential syntactic cues for accurate entity recognition. During preprocessing, we aligned tokenized inputs with their respective labels, incorporating necessary markers like [CLS] for sentence initiation and [SEP] for termination.

4.3 Procedure

Training was conducted using the following key parameters:

- Batch Size: 32, to balance memory constraints and gradient update efficiency.
- Epochs: 50, ensuring comprehensive learning without overfitting.
- Optimizer: AdamW, adjusting learning rates per-parameter for better model refinement.
- Learning Rate: 1e-4, for controlled adjustments during backpropagation.
- Learning Rate Scheduler: The StepLR scheduler adjusts the learning rate based on a step function, which helps in maintaining stable training dynamics over epochs.
- Loss Function: The Cross-Entropy Loss was used to handle the multi-class classification nature of the NER task.

The training loop included evaluation on the development set at the end of each epoch, which allowed us to monitor the model’s performance and adjust parameters if necessary. To prevent overfitting, we also implemented early stopping based on the validation loss. If the validation loss did not improve for 10 consecutive epochs, training would terminate.

4.4 Model Evaluation and Selection

During training, we maintained a focus on the validation loss, ensuring that our model improvements were not achieved at the expense of overfitting. The model with the lowest validation loss on the development set was selected as the final model. This approach ensured that our system was tuned not only for accuracy but also for robustness in practical applications, aligning with the goals of the NER Shared Task 2024.

5 Results

The results obtained from the NER system developed for the WojoodNER 2024 competition exhibit strong performance across training, validation, and testing phases, as highlighted in the table 1 below:

Metric	Training	Validation	Testing
Precision	95.19%	87.86%	88.06%
Recall	96.57%	90.49%	84.90%
F1-Score	95.87%	89.16%	86.45%

Table 1: Model performance results.

Table 2 shows how our model performs compared to other submissions during the shared task:

Subtask 1- Flat				
Team Name	**F1-score**	**Precision**	**Recall**	**Rank**
mucAI	91	91	90	1
muNERa	90	91	89	2
Addax	90	89	91	2
baseline	89	89	90	
DRU - Arab Center	87	86	88	4
Bangor	86	88	85	5
Subtask 2- Nested				
Team Name	**F1-score**	**Precision**	**Recall**	**Rank**
baseline	92	92	92	
muNERa	91	92	90	1
DRU - Arab Center	90	90	90	2

Table 2: Performance results in Subtask 1 - Flat and Subtask 2 - Nested

6 Discussion of Results

6.1 Training Phase

During the training phase, the model achieved exceptionally high metrics, demonstrating a strong understanding of the dataset. The F1-score of 95.87% indicates that the model was very effective in detecting and classifying entities according to the training data. Such high performance can suggest good generalization but also raises concerns about potential overfitting, particularly if validation and test scores are significantly lower.

6.2 Validation Phase

In the validation phase, there was a noticeable decrease in all metrics compared to the training phase, which is a common scenario in machine learning models where the model faces data it has not seen during training. The F1-score of 89.16% is robust, albeit lower than the training phase, suggesting that while the model generalizes well, there's room for improvement, especially in managing the model's complexity and handling features it hasn't learned during training.

6.3 Testing Phase

The testing phase results are critical as they represent the model's ability to perform under completely unseen conditions, similar to how it would perform when deployed in a real-world scenario. An F1-score of 86.45% is commendable and aligns closely with the validation scores, indicating that the model's performance is consistent even on completely unseen data. The precision and recall values are balanced, suggesting that the model neither favors precision over recall nor vice versa.

6.4 Implications and limitations

The high performance on the training dataset and the robust results on validation and testing datasets imply that the model architecture and training regimen were well-suited for this task. However, the drop from training to validation performance highlights the need for strategies to reduce overfitting. Techniques such as dropout, regularization, and possibly more data augmentation could be explored to bridge the gap between training and validation performances.

Conclusion

Bangor University's involvement in the WojoodNER 2024 has demonstrated the effectiveness of transformer-based models, particularly the bert-base-arabic-camelbert-mix, in addressing the complexities of fine-grained Arabic NER. Our results across the training, validation, and testing phases attest to the model's ability to achieve high accuracy and generalize well across unseen data. This participation not only pushes the envelope in Arabic NER research by tackling intricate classification challenges but also sets a benchmark for future work in deep learning applications within complex linguistic tasks. Moving forward, we plan to refine our training strategies and preprocessing methods to enhance model performance consistently across all phases.

References

Benajiba, Y., Rosso, P. and Benedíruiz, J.M., 2007. 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 (pp. 143-153). Springer Berlin Heidelberg.

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H. and Habash, N., 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. arXiv preprint arXiv:2103.06678.

Jarrar, M., Abdul-Mageed, M., Khalilia, M., Talafha, B., Elmadany, A., Hamad, N. and Omar, A., 2023. WojoodNER 2023: The First Arabic Named Entity Recognition Shared Task. arXiv preprint arXiv:2310.16153.

Jarrar, M., Khalilia, M. and Ghanem, S., 2022. Wojood: Nested arabic named entity corpus and recognition using bert. arXiv preprint arXiv:2205.09651.

Jarrar, M., Hamad, N., Khalilia, M., Talafha, B., Elmadany, A. and Abdul-Mageed, M., 2024. WojoodNER 2024: The Second Arabic Named Entity Recognition Shared Task. In: Proceedings of the 2nd Arabic Natural Language Processing Conference (Arabic-NLP), Part of the ACL 2024. Association for Computational Linguistics.

Liqreina, H., Jarrar, M., Khalilia, M., El-Shangiti, A.O. and AbdulMageed, M., 2023. Arabic fine-grained entity recognition. arXiv preprint arXiv:2310.17333.

Salah, R.E. and Zakaria, L.Q.B., 2018, March. Building the classical Arabic named entity recognition corpus (CANERCorpus). In 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP) (pp. 1-8). IEEE.

Walker, C., Strassel, S., Medero, J. and Maeda, K., 2005. Ace 2005 multilingual training corpus-linguistic data consortium. URL: https://catalog. ldc. upenn. edu/LDC2006T06.