LIPN at WojoodNER shared task: A Span-Based Approach for Flat and Nested Arabic Named Entity Recognition

Niama Elkhbir[†⋆], **Urchade Zaratiana**[†×⋆], **Nadi Tomeh**[†], **Thierry Charnois**[†]
[×] FI Group, [†] LIPN, CNRS UMR 7030, France
{elkhbir,zaratiana,tomeh,charnois}@lipn.fr
[⋆]: Equal contribution

Abstract

The Wojood Named Entity Recognition (NER) shared task introduces a comprehensive Arabic NER dataset encompassing both flat and nested entity tasks, addressing the challenge of limited Arabic resources. In this paper, we present our team **LIPN** approach to addressing the two subtasks of WojoodNER SharedTask. We frame NER as a span classification problem. We employ a pretrained language model for token representations and neural network classifiers. We use global decoding for flat NER and a greedy strategy for nested NER. Our model secured the first position in flat NER and the fourth position in nested NER during the competition, with an F-score of 91.96 and 92.45 respectively. Our code is publicly available (https://github.com/niamaelkhbir/LIPN-at-WojoodSharedTask).

1 Introduction

Named Entity Recognition (NER) plays a crucial role in various Natural Language Processing (NLP) applications, enabling the extraction and classification of entities from unstructured text. These entities span a wide range of categories, including individuals, organizations, locations, and dates, among others. While NER has witnessed significant progress, challenges persist, particularly in contexts marked by resource scarcity and linguistic complexity, such as the Arabic language.

In this context, the focus of Arabic NLP has predominantly revolved around flat entities (Liu et al., 2019; Helwe et al., 2020; Al-Qurishi and Souissi, 2021; El Khbir et al., 2022; Affi and Latiri, 2022), and the exploration of nested entity recognition in Arabic NLP has been relatively limited, primarily due to the scarcity of suitable nested Arabic datasets.

To address these limitations, the WojoodNER SharedTask 2023 (Jarrar et al., 2023) initiative was launched with the goal of overcoming these challenges. This initiative introduces the Wojood corpus (Jarrar et al., 2022), an extensively annotated Arabic NER dataset comprising approximately 550,000 tokens. It includes annotations for 21 distinct entity types, covering both Modern Standard Arabic (MSA) and dialectal variations, as well as flat and nested entity annotations.

The shared task objective is twofold: firstly, to encourage innovative solutions in flat NER, and secondly, to tackle nested NER. For both tasks, the aim is to develop models that can effectively identify and classify entities while accounting for complexities.

This paper outlines our strategy for tackling these subtasks. Our approach relies on a span-based methodology, employing token encoding, span enumeration, and subsequent classification. During inference, we employ global decoding for flat NER and a greedy decoding strategy for nested NER. Our contributions led us to achieve the top position in flat NER and the fourth position in nested NER during the WojoodNER SharedTask 2023.

In the following sections, we provide detailed insights into our methodology, experimentation, and the results achieved, highlighting the efficacy of our approach within the WojoodNER SharedTask 2023.

2 Related Work

Evolution of NER Approaches Early efforts in NER relied on handcrafted rules and lexicons for both flat (Zhou and Su, 2002) and nested entities (Shen et al., 2003; Zhang et al., 2004). Then, machine learning techniques gained prominence. Many studies focused on statistical models, such as Conditional Random Fields (CRFs) and Support Vector Machines (SVMs). These models demonstrated improved performance in identifying entities by capturing contextual dependencies and patterns within the data (McCallum and Li, 2003; Takeuchi and Collier, 2002). Deep learning

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 789–796

techniques, particularly recurrent neural networks (RNNs) and recently, transformer-based architectures like BERT (Devlin et al., 2019) and GPT (Radford et al., 2019), revolutionized NER. These models leverage contextual embeddings to capture intricate relationships and dependencies, achieving state-of-the-art results in various languages and domains for both flat (Xia et al., 2019; Zheng et al., 2019; Arkhipov et al., 2019; Lothritz et al., 2020; Yu et al., 2020; Yang et al., 2021) and nested (Sohrab and Miwa, 2018; Katiyar and Cardie, 2018; Dadas and Protasiewicz, 2020; Wang et al., 2020) entities.

Approaches for NER Traditionally, NER tasks have been framed as sequence labeling (Lample et al., 2016; Akbik et al., 2018), i.e., token-level classification. Recently, innovative approaches have extended beyond token-level prediction. Some methods have treated NER as a question-answering problem (Li et al., 2020), while others have employed sequence-to-sequence models (Yan et al., 2021; Yang and Tu, 2022). In this work, we focus on span-based methods (Liu et al., 2016; Sohrab and Miwa, 2018; Fu et al., 2021; Zaratiana et al., 2022b), which involve enumerating all possible spans and then classifying them into specific entity types.

3 Data

	#Sentences	#Tokens	#F-Ent	#N-Ent
Train	16817	394500	50032	62403
Valid	3133	55827	7141	8854

Table 1: Statistics on Train and Validation Splits of Wojood Corpus.

The Wojood corpus is annotated for 21 different entity types, and it offers two versions: Wojood Flat and Wojood Nested. Both versions share identical training, validation, and test splits, differing only in the way entities are labeled. In Wojood Flat, each token receives a label corresponding to the first high-level label assigned to that token in Wojood Nested. Table 1 presents an overview of the statistics for the train and validation splits, including the number of sentences, tokens, flat entities (#F-Ent), and nested entities (#N-Ent).

Furthermore, Table 2 provides a breakdown of entity label counts for both flat and nested versions within the train and validation splits.

To offer insights into the entity distribution based on span lengths, Figure 1 displays the entity count distribution concerning span lengths within the Flat Wojood training data. Note that for the sake of clarity in visualization, we have excluded entity counts for span lengths of 27, 29, 39, 43, and 124, each of which occurs either once or twice. We have established a maximum entity span length of 10 for our span-based model. Any entities surpassing this threshold are automatically excluded. Specifically, the training set includes 140 such entities, predominantly categorized as Website, Date, and Event. Similarly, in the validation set, 19 entities exceed the 10-span limit.

	Flat		Nested	
Label	**Train**	**Val**	**Train**	**Val**
CARDINAL	1245	182	1263	183
CURR	19	1	179	21
DATE	10667	1567	11291	1656
EVENT	1864	253	1935	267
FAC	689	85	882	111
GPE	8133	1132	15300	2163
LANGUAGE	131	15	132	15
LAW	374	44	374	44
LOC	510	63	619	76
MONEY	171	20	171	20
NORP	3505	488	3748	520
OCC	3774	544	3887	551
ORDINAL	2805	410	3488	504
ORG	10737	1566	10737	1566
PERCENT	105	13	105	13
PERS	4496	650	4996	730
PRODUCT	36	5	36	5
QUANTITY	44	3	46	3
TIME	286	55	288	55
UNIT	7	-	48	3
WEBSITE	434	45	434	45

Table 2: Entity Label Statistics in Wojood Corpus.

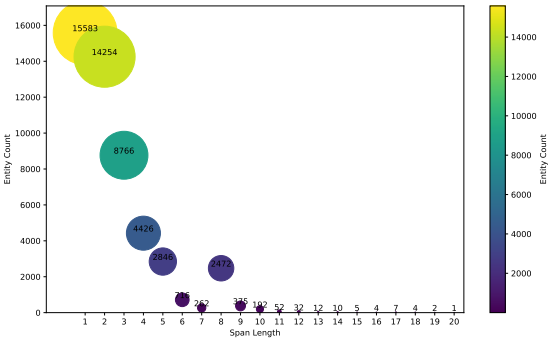

Figure 1: Entity count distribution by span length in the Flat Wojood training data.

4 System

In this paper, we approach the named entity recognition task as a span classification problem. Given

an input sequence $\boldsymbol{x} = \{x_i\}_{i=1}^{L}$, our goal is to classify all possible spans within the sequence, which can be defined as:

$$\boldsymbol{y} = \bigcup_{i=1}^{L} \bigcup_{j=i}^{L} s_{ijc} \tag{1}$$

where i, j, and c represent the start position, end position, and span type respectively. The probability of a specific span classification $\boldsymbol{y}$ given the input sequence $\boldsymbol{x}$ can be expressed as:

$$p_\theta(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp \sum_{s_{ijc} \in \boldsymbol{y}} \phi_\theta(s_{ijc}|\boldsymbol{x})}{\mathcal{Z}_\theta(\boldsymbol{x})} \tag{2}$$

where $\phi_\theta(.)$ is the span scoring function and $\mathcal{Z}_\theta(\boldsymbol{x})$ is the partition function. During training, our objective is to minimize the negative log-likelihood of the gold span classifications.

Decoding During inference, our aim is to determine:

$$\boldsymbol{y}^* = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} \sum_{s_{ijc} \in \boldsymbol{y}} \phi_\theta(s_{ijc}|\boldsymbol{x}) \tag{3}$$

In other words, we seek to identify the span labeling configuration ($\boldsymbol{y} \in \mathcal{Y}$) that achieves the highest score (sum of individual span ($s_{ijc} \in \boldsymbol{y}$)). For unconstrained span classification, a straightforward approach is to assign the label with the highest score to each individual span, as follows:

$$s_{ijc^*} = \arg\max_{c} \phi_\theta(s_{ijc}|\boldsymbol{x}) \tag{4}$$

However, for both flat and nested NER, such a decoding strategy is suboptimal as it can lead to violations of structural constraints. For **flat NER**, where overlapping entity spans are not allowed, an efficient solution has been proposed in our previous works (Zaratiana et al., 2022c,a)[1]. This approach involves a two-stage decoding process: first, spans predicted as non-entities are filtered out, and then a maximum independent set algorithm is applied to the remaining spans to obtain the optimal set of entity spans. In contrast, for **nested NER**, where nesting is permitted but conflicting boundaries are prohibited, we employ a greedy algorithm to achieve a valid span classification. This algorithm iteratively selects the highest-scoring span that does not conflict with already selected spans.

[1] https://github.com/urchade/Filtered-Semi-Markov-CRF

	Flat NER		
TEAM	**P**	**R**	**F1**
LIPN (*Ours*)	92.56	91.36	91.96
El-Kawaref	91.43	92.48	91.95
ELYADATA	91.88	91.96	91.92
Alex-U 2023 NLP	91.61	92.00	91.80
tdink NER	90.76	91.73	91.25
	Nested NER		
TEAM	**P**	**R**	**F1**
ELYADATA	93.99	93.48	93.73
UM6P	92.46	93.61	93.03
AlexU-AIC	92.10	93.13	92.61
LIPN (*Ours*)	92.31	92.59	92.45
tdink NER	90.03	92.82	91.40

Table 3: Top 5 results for the Wojood flat/nested ner shared task.

Token and Span Representations In our approach, the span score $\phi_\theta(s_{ijc}|\boldsymbol{x})$ is computed as a linear projection of the span representation, obtained through a $1D$ convolution of token representations from a BERT-based model:

$$\boldsymbol{s}_{ijc} := w_c^T \texttt{Conv1D}_k([\boldsymbol{h}_i; \boldsymbol{h}_{i+1}; \ldots; \boldsymbol{h}_j]) \tag{5}$$

where $h_i \in \mathbb{R}^D$ is the token representation at position i, k is the size of the convolutional filter (j-i), and $w_c \in \mathbb{R}^D$ is a learned weight matrix for the span label c.

5 Results

Evaluation Metrics Following the shared task guidelines, we assess the performance of our model using precision, recall, and F1-score.

Settings and Hyperparameters For token representation, we use `bert-base-arabert` (Antoun et al., 2020) as a pretrained language model. Subsequently, we process the encoded tokens through a bidirectional Long Short-Term Memory (bi-LSTM) encoder to obtain the final representations. We set a maximum span length of 10 for enumerating all possible spans, which is a good balance between recall and training speed (Refer to Limitations Section).

Our model is trained with a batch size of 12 and evaluated with a batch size of 32. We set a learning rate of 5e-6 for BERT and 1e-3 for other model

parameters. We use the Adam optimizer and train our model for 50,000 steps, conducting evaluations every 250 steps.

We ran our experiments on a server equipped with v100 GPUs, and we estimated the needed computational budget for training to be 50 GPU hours.

Main results Eleven teams took part in the shared task, but due to space limitations, we present the results of the top 5 teams from the official leaderboard, which includes our own, in Table 3. The main results highlight the performance of our model in both the flat and nested Named Entity Recognition (NER) tasks. Our model achieved a good balance between precision and recall in both tasks, with a higher F-score in nested NER compared to flat NER.

Results by Class Table 4 presents the F1-scores associated with each label for both flat NER and nested NER on the validation set. Our model demonstrates high performance across both tasks for various entity types, including CURR, DATE, GPE, LAW, MONEY, ORDINAL, ORG, PERCENT, and PERS, all of which achieve an F-score exceeding 92.00.

The worst performance is observed for PRODUCT and WEBSITE, with F1-scores of 60.00 and 63.77, respectively. We provide further insights into this performance in section 6.2.

6 Discussion

6.1 Class Imbalance

One of the problems encountered in the Wojood dataset is class imbalance, where certain classes are significantly underrepresented in the training set. For example, the classes CURR, PRODUCT, QUANTITY, and UNIT constitute only 0.04%, 0.07%, 0.8%, and 0.01% of the training data, respectively. In contrast to dominant classes like DATE (21.23%), GPE (16.25%), and ORG (21.46%).

Such class imbalance can potentially skew evaluation results, especially when based solely on F-scores for these minority classes. Further work may involve sampling or data augmentation techniques to rebalance the dataset and provide more equitable representation and accurate assessment of the performance on these underrepresented classes.

6.2 Analysis of Model Errors

In this section, we analyze the remaining errors of our model in the validation set for flat NER.

Label	Flat	Nested
CARDINAL	89.44	87.98
CURR	100	100
DATE	96.20	96.46
EVENT	85.05	84.98
FAC	78.05	82.73
GPE	92.21	96.93
LANGUAGE	83.87	87.50
LAW	95.35	93.18
LOC	81.60	87.25
MONEY	95.00	91.89
NORP	79.25	79.20
OCC	89.66	89.99
ORDINAL	94.59	96.04
ORG	93.57	94.24
PERCENT	96.30	96.30
PERS	95.35	95.62
PRODUCT	60.00	66.67
QUANTITY	80.00	100
TIME	78.35	74.00
UNIT	-	80.00
WEBSITE	63.77	66.67

Table 4: F1-Scores by Entity Labels

Correct Span Offsets, but Incorrect Label Within this category, our model correctly identifies the span offsets but assigns incorrect labels to these spans. We identified a total of 68 instances where the model demonstrated this behavior.

To gain deeper insights into these errors, we provide in Figure 2 a visual representation of the confusion matrix for entity labels.

Approximately 45% of these errors arise from the ambiguity associated with certain entity labels, notably LOC, ORG and GPE. These errors often concern country or city names, such as الولايات المتحدة or السعودية, which, depending on the context, may belong to any of these categories.

Similarly, ambiguity between CARDINAL and ORDINAL labels accounts for 7% of this error category, while WEBSITE and ORG labels contribute approximately 6%. Also, NORP and ORG labels account for 7%. The remaining errors on labels can be found in Figure 2.

We observe comparable error patterns in the nested NER task. In Figure 3, we provide the confusion matrix for nested NER.

Span Boundary Errors with Correct Label Within this category, our model correctly predicts the entity label but fails to accurately identify the

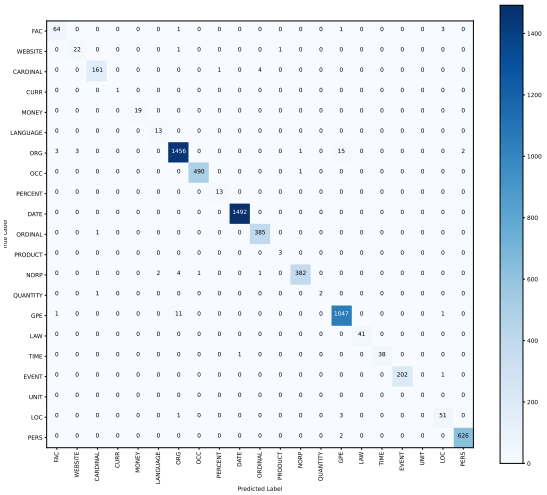

Figure 2: Confusion Matrix of Entity labels for flat NER.

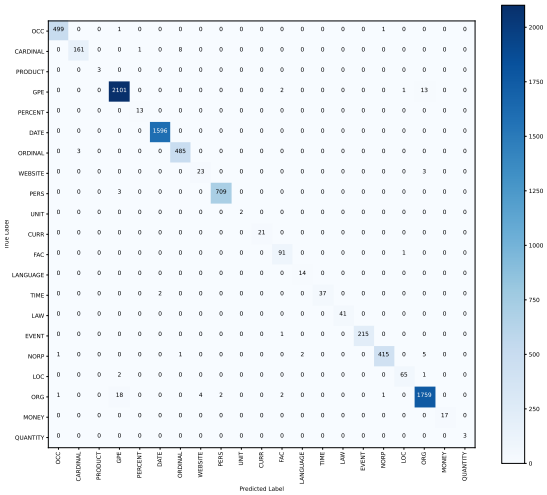

Figure 3: Confusion Matrix of Entity labels for nested NER.

start and end positions (span boundaries) of the entity within the text. We identified 167 instances where the model demonstrated this behavior. This category can be further broken down into two subtypes: (1) Span Start Error: The span start position is correct but the end position is incorrect; and (2) Span End Error: The span end position is correct but the start position is incorrect. Some of these errors seem to be annotation errors. See Table 5 for concrete examples.

False Negatives with Novel Entities Another type of error occurs when our model predicts spans that are not included in the gold annotations. We identified 305 instances where the model demonstrated this behavior. Although we did not conduct a precise quantification, a notable subset of these errors can be categorized as "false negatives". These false negatives are not part of the gold standard annotations, but they may have legitimacy as valid entities, thus the term "Novel Entities". Table 6 in the Appendix provides some illustrative examples of these errors.

7 Conclusion

Our approach to Arabic Named Entity Recognition in the WojoodNER Shared Task 2023 yielded competitive results, securing first place in flat NER and fourth in nested NER. This success highlights the potential of span-based methods and advanced decoding strategies. Moreover, we identified areas for improvement, including addressing class imbalance and refining span boundary predictions.

Limitations

Span Length Limitation Errors: In addition to the errors mentioned in Section 6.2, another type of errors is due to the span length limitation. As mentioned in Section 3, we have set a predefined limit of 10 tokens for span lengths, thus excluding all entities above this threshold. This decision was made to strike a balance between model complexity and computational efficiency. Due to this imposed constraint, our model cannot predict spans that surpass the 10-token threshold resulting in a reduced recall score. Particularly, with 140 and 19 spans surpassing the threshold in the training and validation set respectively, the maximum attainable recall score is 99.72 and 99.73 for the training and validation set respectively.

Acknowledgements

This work is partially supported by a public grantoverseen by the French National Research Agency (ANR) as part of the program Investissements d'Avenir (ANR-10-LABX-0083). This work was granted access to the HPC/AI resources of [CINES/IDRIS/TGCC] under the allocation 2023AD011013096R1 made by GENCI.

References

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.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. 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.

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.

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

Sławomir Dadas and Jarosław Protasiewicz. 2020. A bidirectional iterative algorithm for nested named entity recognition. *IEEE Access*, 8:135091–135102.

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.

Niama El Khbir, Nadi Tomeh, and Thierry Charnois. 2022. Arabie: Joint entity, relation and event extraction for arabic. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 331–345.

Jinlan Fu, Xuanjing Huang, and Pengfei Liu. 2021. SpanNER: Named entity re-/recognition as span prediction. 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 7183–7195, Online. Association for Computational Linguistics.

Chadi Helwe, Ghassan Dib, Mohsen Shamas, and Shady Elbassuoni. 2020. A semi-supervised BERT approach for Arabic named entity recognition. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 49–57, Barcelona, Spain (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, Marseille, France. European Language Resources Association.

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.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified mrc framework for named entity recognition.

Liyuan Liu, Jingbo Shang, and Jiawei Han. 2019. Arabic named entity recognition: What works and what's next. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 60–67, Florence, Italy. Association for Computational Linguistics.

Yijia Liu, Wanxiang Che, Jiang Guo, Bing Qin, and Ting Liu. 2016. Exploring segment representations for neural segmentation models. In *IJCAI*.

Cedric Lothritz, Kevin Allix, Lisa Veiber, Tegawendé F. Bissyandé, and Jacques Klein. 2020. Evaluating pretrained transformer-based models on the task of fine-grained named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3750–3760, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Dan Shen, Jie Zhang, Guodong Zhou, Jian Su, and Chew-Lim Tan. 2003. Effective adaptation of hidden Markov model-based named entity recognizer for biomedical domain. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 49–56, Sapporo, Japan. Association for Computational Linguistics.

Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2849, Brussels, Belgium. Association for Computational Linguistics.

Koichi Takeuchi and Nigel Collier. 2002. Use of support vector machines in extended named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Jue Wang, Lidan Shou, Ke Chen, and Gang Chen. 2020. Pyramid: A layered model for nested named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5918–5928, Online. Association for Computational Linguistics.

Congying Xia, Chenwei Zhang, Tao Yang, Yaliang Li, Nan Du, Xian Wu, Wei Fan, Fenglong Ma, and Philip Yu. 2019. Multi-grained named entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1430–1440, 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 *ACL*.

Songlin Yang and Kewei Tu. 2022. Bottom-up constituency parsing and nested named entity recognition with pointer networks. In *ACL*.

Zhiwei Yang, Jing Ma, Hechang Chen, Yunke Zhang, and Yi Chang. 2021. HiTRANS: A hierarchical transformer network for nested named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 124–132, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352, Online. Association for Computational Linguistics.

Urchade Zaratiana, Niama Elkhbir, Pierre Holat, Nadi Tomeh, and Thierry Charnois. 2022a. Global span selection for named entity recognition. In *Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)*, pages 11–17, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2022b. GNNer: Reducing overlapping in span-based NER using graph neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 97–103, Dublin, Ireland. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2022c. Named entity recognition as structured span prediction. In *Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)*, pages 1–10, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jie Zhang, Dan Shen, Guodong Zhou, Jian Su, and Chew-Lim Tan. 2004. Enhancing hmm-based biomedical named entity recognition by studying special phenomena. *Journal of Biomedical Informatics*, 37(6):411–422. Named Entity Recognition in Biomedicine.

Changmeng Zheng, Yi Cai, Jingyun Xu, Ho-fung Leung, and Guandong Xu. 2019. A boundary-aware neural model for nested named entity recognition. 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 357–366, Hong Kong, China. Association for Computational Linguistics.

GuoDong Zhou and Jian Su. 2002. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 473–480, USA. Association for Computational Linguistics.

A Example of Remaining Errors

Table 5 and 6 present examples of errors related to span boundaries and examples where the model predicts spans that are not part of the gold standard annotations, respectively.

Sentence	Gold Span	Predicted Span	Label
ضوابط حماية صغار المستثمرين	المستثمرين	صغار المستثمرين	NORP
لقد خلقت هده الأزمة انشقاقا في منطقة الشرق الأوسط وشمال افريقيا …	الشرق الأوسط	منطقة الشرق الأوسط	LOC
شملت هذه القضايا سكانا سابقين في الغوطة، وحلب، وريف دمشق …	دمشق	وريف دمشق	GPE
عندما التقينا في مدينة عدن الساحلية …	مدينة عدن الساحلية	مدينة عدن	GPE
كوفئ اليمنيون، بمن فيهم المتظاهرون الذين شجعوا التغيير، باستفتاء في فبراير …	المتظاهرون الذين شجعوا التغيير	المتظاهرون	NORP
بعد أسابيع قليلة في نيويورك، …	بعد أسابيع قليلة	بعد أسابيع	DATE
إفراجات وغرامات مالية ضد المعتقلين الفلسطينيين (١٩٨٦)	الفلسطينيين	المعتقلين الفلسطينيين	NORP
برنامج خاص حول قمة واشنطن الرباعية تم إذاعته	قمة واشنطن	قمة واشنطن الرباعية	EVENT

Table 5: Example of Span Boundary Errors from the validation set for flat NER. The model predicts the correct label but fails to capture the gold span offsets.

Sentence	Predicted Span	Predicted Label
معرض المال والأعمال الأول	الأول	ORDINAL
ساعاتا بتزكّر نوزيا تبع سارتر	سارتر	PERS
اهتم بالاثنين فكلاهم مهم	بالاثنين	NORP
اهل قرية بيت نبالا اليوم	اليوم	DATE
في تخصصك: لنقل أنك مصمم	مصمم	OCC
المهم هسا الدكتور يرقص بقناة غنوه	الدكتور	OCC
welcome to geeksforgeeks	geeksforgeeks	WEBSITE
… والقرضاوي عم بيحرم الصلاة بالمسجد الاقصى	والقرضاوي	PERS
انا مسلم (رائع جدا)	مسلم	NORP
protect yourself from hackers	hackers	NORP
كيف أتابع ناروتو ؟	ناروتو	PERS
كان يا ما كان في قديم الزمان، ملك عنده ثلاث بنات	ثلاث	CARDINAL
الجميع يضغطون على الجرس ما عدا الخليجي	الخليجي	NORP
تنتهي الاغنيه يرن هاتف المدير	المدير	OCC
بحسب نيرما جيلاسيتش، ممثلة اللجنة، تجمع اللجنة أدلة على الجرائم	ممثلة اللجنة	OCC
… فإن المجتمع الإيزيدي رفض الادعاءات	المجتمع الإيزيدي	NORP

Table 6: Example of False Negatives with Novel Entities from the validation set for flat NER. These are entities predicted by the model but not annotated in the dataset. All reported entities do not manifest any overlap with gold ones.