Low-Resource Named Entity Recognition: Can One-vs-All AUC Maximization Help?

Ngoc Dang Nguyen
Department of Data Science and AI
Monash University
Melbourne, Australia
dan.nguyen2@monash.edu

Wei Tan
Department of Data Science and AI
Monash University
Melbourne, Australia
wei.tan2@monash.edu

Lan Du*
Department of Data Science and AI
Monash University
Melbourne, Australia
lan.du@monash.edu

Wray Buntine
College of Engineering and Computer Science
VinUniversity
Hanoi, Vietnam
wray.b@vinuni.edu.vn

Richard Beare
Faculty of Medicine
Peninsula Clinical School
Melbourne, Australia
richard.beare@monash.edu

Changyou Chen
Department of CS and Engineering
University at Buffalo
New York, USA
changyou@buffalo.edu

2023 IEEE International Conference on Data Mining (ICDM) | 979-8-3503-0788-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICDM58522.2023.00155

***Abstract*—Named entity recognition (NER), a task that identifies and categorizes named entities such as persons or organizations from text, is traditionally framed as a multi-class classification problem. However, this approach often overlooks the issues of imbalanced label distributions, particularly in low-resource settings, which is common in certain NER contexts, like biomedical NER (bioNER). To address these issues, we propose an innovative reformulation of the multi-class problem as a one-vs-all (OVA) learning problem and introduce a loss function based on the area under the receiver operating characteristic curve (AUC). To enhance the efficiency of our OVA-based approach, we propose two training strategies: one groups labels with similar linguistic characteristics, and another employs meta-learning. The superiority of our approach is confirmed by its performance, which surpasses traditional NER learning in varying NER settings.**

***Index Terms*—NER, NLP, One-vs-All, AUC, Low-Budget**

I. Introduction

Named Entity Recognition (NER), a fundamental task in Natural Language Processing (NLP), aims to detect the semantic category of named entities (NE), such as location, organization, or person. As an essential prerequisite for many language applications, NER often plays a pivotal role in a variety of NLP tasks, including information extraction, information retrieval, task-oriented dialogues, and knowledge base construction [1], [2]. NER's relevance extends to numerous real-world applications; for instance, it aids in extracting information from unstructured text in the medical domain to improve patient care, it enhances search engine algorithms to better understand user queries, and it supports intelligence services by identifying crucial entities in large text corpora [2].

Recently, NER has seen significant performance improvements due to the advances of state-of-the-art (SOTA) pre-trained language models (PLMs) [1]. However, these PLMs rely heavily on sizable training datasets, and in specialized low-resource settings (*e.g.*, biomedical domains), the lack of data often leads to sub-optimal performance [3]. This poses a significant challenge, as many languages and specialized domains suffer from a lack of annotated data [2], [4]. Improving NER performance in these low-resource settings has the potential to democratize access to advanced language processing capabilities, thus breaking down barriers to information access and enabling new NLP applications across a multitude of disciplines and industries.

NER models are often heavily dependent on human-annotated data, which can be costly and time-consuming to obtain. To alleviate this dependence on labeled data, existing low-resource machine learning approaches such as domain adaptation [5], and data augmentation [6] have been

* Corresponding Author

(a) AUC (b) BCE

Fig. 1: Illustration of the OVA predicted probability performance of classifiers trained with AUC and BCE losses for each token's BIO-tag on the CoNLL 2003 test data. The binary classifiers trained with the BCE loss exhibit a lack of confidence in token classification, as indicated by the clustering of all predictions rather than their stratification towards their respective label corners. Conversely, when trained with the AUC loss, these binary classifiers effectively classify tokens, showing the potential of AUC for enhanced NER performance.

979-8-3503-0788-7/23/$31.00 ©2023 IEEE
DOI 10.1109/ICDM58522.2023.00155

applied to NER. However, these methods do not account for the imbalanced label distributions commonly found in NER datasets, where the majority of labels in the NER corpora are of the non-entity type "O" (see Table I). These non-entity labels provide limited learning signals for NER models, which presents challenges to the NER classifier in correctly identifying NEs in sentences. This imbalance is particularly prominent in specialized biomedical corpora, such as NCBI [7], and s800 [8], where over 90% of labels are tagged as "O" (see Table I).

Given the inherent imbalanced label distributions in NER corpora, we argue that even though standard learning objectives (*i.e.*, cross entropy loss or conditional random fields), given adequate annotated training data, are capable of producing well-performing token classifiers or sequence labellers, their performance can substantially degrade in low-resource settings. Consequently, we propose to directly address the imbalance issue by training NER models with a surrogate loss that maximizes the area under the ROC curve (AUC). AUC maximization has been demonstrated to greatly improve model prediction for tasks with imbalanced label distribution [9]–[11]. In the NER task, the recent work of [4] has attempted to utilize AUC maximization by reformulating the NER task as a two-task learning problem. While this approach shows significant improvements over traditional methods, it fails to directly address the problem of predicting entity types, such as location, organization, or person, limiting its usage in real-world applications where identifying the entity type is crucial.

Thus, multi-class AUC objectives, which cannot be developed using the aforementioned two-task approach [4], are essential for NER. They can be developed with two strategies. The first is the one-vs-other (OVO) reformulation [12], which is not computationally feasible due to its quadratic relationship with the number of classes. Consequently, we reformulate the standard sequence tagging problem as a one-vs-all (OVA) learning problem. In this context, each unique label (*e.g.*, B-PER, I-ORG, O, *etc.*) will have its own binary classifier, which can be individually learned with an AUC objective function.

However, OVA has two notable weaknesses: **(i)** compared to the traditional multi-class setup, OVA has a longer training time since each binary classifier must be independently trained, and **(ii)** OVA is not data efficient for learning from imbalanced data which results in less accurate classifiers compared to its multi-class counterpart [13]. As shown in Figure 1, the predictive confidence of OVA, learned with the binary cross entropy loss (BCE), is not well-stratified compared to the predictive confidence learned with AUC. To alleviate the first weakness, we propose two new training strategies: one that groups labels sharing similar linguistic characteristics and trains their classifiers together, and another that adapts the idea of meta-learning [14] by selecting a random batch of binary classifiers for the model to learn. For the second weakness, the binary classifier is tuned with an AUC surrogate loss function, which has shown its resilience to imbalanced data [11], [15].

To demonstrate the effectiveness of our proposed method, we conduct thorough empirical investigations to evaluate the efficacy of our OVA-AUC-NER techniques under various scenarios. The results of the experiments, which are provided in section V, show that our techniques exhibit significant performance improvement compared to traditional baselines in terms of entity-F1 score across corpora (*e.g.*, OntoNotes5, CoNLL 2003, NCBI, and s800), domains (*e.g.*, biomedical,general), and label distributions (*e.g.*, NEs percentages of $1, 2,$ or 5%) in low-resource conditions (*e.g.*, training pool size of $\{20, 50, 100, 200, 300, 400, 500\}$). Our contributions include:

- **Reformulation of NER as an OVA task**: We transform NER from a standard multi-class learning problem to an OVA learning problem, with each unique label having its own classifier. This simple OVA reformulation makes AUC maximization feasible for predicting the entity type in NER.
- **Efficient training strategies**: We propose two efficient training strategies to overcome the computational cost of training in the OVA setup; one groups labels sharing similar linguistic characteristics and trains their classifiers together while the other uses a meta-learning approach [14] to randomly select a batch of classifiers for the model to learn.

II. Related Work

The OVA approach is mostly used to expand binary models for multi-class classification, such as logistic regression, support vector machines, etc [18], [19], with the OVO approach seeing little use due to its higher training time. OVA's objective is to divide a K-class problem into K binary problems. For instance, K binary classifiers must be built, where K is the number of classes, and the i-th classifier is trained with positive data from class i and negative samples from the other $K-1$ classes. When the classifier evaluates an unclassified sample, the highest confidence value of the sample is considered to have labelled corresponding to the specified class [18]. Recently, researchers have discovered how OVA methods can accomplish various tasks with deep neural networks. They found that OVA can identify more relevant hidden representations for unidentified instances than the popular Softmax function [20]. Additionally, OVA improves calibration on image classification, outlier detection, and dataset shift tasks, reaching Softmax's predictive performance without increasing the training or the test time complexity [21], [22]. Although the algorithm is simple and straightforward, it shows impressive results, demonstrating that its performance is usually at least as accurate as other multi-class algorithms when the classifiers are properly tuned [23].

AUC (Area Under the ROC Curve) has been traditionally treated as an important criterion for measuring model's classification performance [24]. Due to its non-convex, and discontinuous nature, most works consider direct optimization of AUC score an NP-hard problem [11]. [24] tried to alleviate this computation difficulty through a pairwise surrogate loss, while [25] implemented a hinge loss. However, both of these surrogate losses lack scalability to large datasets and models. This led to the development of the least-square surrogate loss [9]. Recent research on AUC maximization further optimizes the least-square surrogate loss via deep margin surrogate loss [11] and compositional training [15]. Overall, AUC maximization works well when there exists an imbalanced

TABLE I: Summary of dataset. Please note that we use BIO here instead of the entity type label for ease of reference.

Dataset	Domain	% Training Gold Samples	# Sentences Train/Dev/Test	# Tokens Train/Dev/Test	# Labels	% label (B/I/O) Train	Dev	Test
OntoNotes5 [16]	General	45.59	20,000/3,000/3,000	364,344/54,372/55,754	37	6.2/4.8/89.0	6.2/4.8/89.0	6.1/4.9/89.0
CoNLL 2003 [17]	General	79.25	14,040/3,249/3,452	203,589/51,319/46,376	9	11.5/5.2/83.3	11.6/5.1/83.3	12.2/5.3/82.5
NCBI [7]	Disease	53.89	5,424/923/940	135,597/23,969/24,481	3	3.8/4.5/91.7	3.3/4.5/92.2	3.9/4.5/91.6
s800 [8]	Species	29.51	5,733/830/1,630	147,205/22,166/42,287	3	1.7/2.3/96.0	1.7/2.2/96.1	1.8/2.5/95.7

label distribution, or the AUC score is the default metric for evaluating and comparing different methods [11]. Prior work for AUC maximization in NER includes [4] which fails to predict entity types, such as ORG, LOC, or PER; thus, our work is the first to explore AUC maximization for the NER task while appropriately identifying the types of the entities.

III. AUC Maximization for NER

Notation. Given a set of training data $\mathcal{S} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$, where $\mathbf{x}_i = x_i^1, \ldots, x_i^l$ represents the i-th training example (*i.e.*, a sentence of length l), and $\mathbf{y}_i \in \{\mathrm{B}, \mathrm{I}, \mathrm{O}\}^l$ denotes its corresponding sequence of labels, we would like to learn the objective mapping function $h : \mathcal{X} \rightarrow \mathcal{Y}$. This objective mapping function, traditionally learned with either CRFs or the CE loss, is parameterized with $\mathbf{w} \in \mathbb{R}^d$, *i.e.*, $h_{\mathbf{w}}(\mathbf{x}) = h(\mathbf{w}, \mathbf{x})$. Lastly, different corpora can have different label set, *e.g.*, $y_i \in \{\mathrm{O}, \mathrm{B\text{-}ORG}, \mathrm{B\text{-}PER}, \mathrm{B\text{-}MISC}, \mathrm{B\text{-}LOC}, \mathrm{I\text{-}ORG}, \mathrm{I\text{-}PER}, \mathrm{I\text{-}LOC}, \mathrm{I\text{-}MISC}\}$ for CoNLL 2003 [17]; thus, the use of $\mathbf{y}_i \in \{\mathrm{B}, \mathrm{I}, \mathrm{O}\}^l$ presented in many parts of this study is for ease of reference.

To make direct AUC maximization applicable to NER, we first reformulate the standard NER multi-class setup as a one-vs-all learning problem. Consequently, each unique label is given its own binary classifier that can be learned using the AUC objective function. For instance, given the "O"-tag, the label of this tag is $\mathbf{y}_{\mathbf{O}_i} \in \{-1, 1\}^l$, while the parameter of this task is $\mathbf{w}_{\mathbf{O}} = \{\theta, \omega_{\mathbf{O}}\}$. θ denotes the shared embedding and pretrained language model parameters, while $\omega_{\mathbf{O}}$ denotes the parameters for the binary classifier for the "O"-tag. After the reformulation, we can maximize the AUC score of each binary classifier via the robust deep AUC margin loss (DAM) [11].

$$
\begin{aligned}
&\mathrm{AUC_M}(\mathbf{w_O}) \\
&= \mathbb{E}\left[\left(m_{\mathbf{O}} - h_{\mathbf{w_O}}(x) + h_{\mathbf{w_O}}(x')\right)^2 \mid y_{\mathbf{O}} = 1, y'_{\mathbf{O}} = -1\right] && \text{(1a)} \\
&= \min_{a_{\mathbf{O}}, b_{\mathbf{O}}} A_1(\mathbf{w_O}) + A_2(\mathbf{w_O}) + (m_{\mathbf{O}} - a_{\mathbf{O}} + b_{\mathbf{O}})^2 && \text{(1b)} \\
&= \min_{a_{\mathbf{O}}, b_{\mathbf{O}}} A_1(\mathbf{w_O}) + A_2(\mathbf{w_O}) \\
&+ \max_{\alpha_{\mathbf{O}} \geq 0} \left\{2\alpha_{\mathbf{O}}(m_{\mathbf{O}} - a_{\mathbf{O}} + b_{\mathbf{O}}) - {\alpha_{\mathbf{O}}}^2\right\}, && \text{(1c)}
\end{aligned}
$$

where $A_1(\mathbf{w_O}) = \mathbb{E}[h^2_{\mathbf{w_O}}(x) \mid y_{\mathbf{O}} = 1] - a^2_{\mathbf{O}}$, $A_2(\mathbf{w_O}) = \mathbb{E}[h^2_{\mathbf{w_O}}(x) \mid y_{\mathbf{O}} = -1] - b^2_{\mathbf{O}}$, and $m_{\mathbf{O}}$ is the margin that aims to push the expected prediction scores of negative and positive class far from each other [11]. Examining (1), the minimization problem of $a_{\mathbf{O}}$ and $b_{\mathbf{O}}$ is achieved when $a_{\mathbf{O}} = a(\mathbf{w_O}) = \mathbb{E}[h_{\mathbf{w_O}}(x) \mid y_{\mathbf{O}} = 1]$, and $b_{\mathbf{O}} = b(\mathbf{w_O}) = \mathbb{E}[h_{\mathbf{w_O}}(x') \mid y_{\mathbf{O}} = -1]$ respectively [10], [11]. Thus, we expect that minimizing (1) with respect to $\mathbf{w_O}$ can produce a well-tuned binary classifier for the imbalanced "O"-tag prediction. Given K labels, K losses can be defined under OVA and independently minimized to produce an optimal binary classifier for each unique label.

Algorithm 1 OVA-AUC

Input: Training set $\mathcal{S}$, $\theta, \{\omega_1, \ldots, \omega_K\}$, $\{a, b, \alpha\}^K$
Output: $\theta^*, \{\omega_1^*, \ldots, \omega_K^*\}$

```
for epoch in range(num epochs) do
  for prefix in {B,I, O} do
    L := 0
    for i in range(K) do
      if i.startswith(prefix) then
        L+ = Equation (1)
      end if
    end for
    L.optimize()
  end for
end for
```

At prediction time on the test data, we evaluate the individual classifiers by following the maximum confidence strategy [18] to generate the entity tags expected by the corpus label set.

$$\hat{y_i} = \underset{1 \ldots K}{\arg\max} \left[h_{\mathbf{w}_1^*}(x_i), \ldots, h_{\mathbf{w}_K^*}(x_i)\right], \tag{2}$$

where $\mathbf{w}_k^*$ represents the optimal parameters learnt by minimizing (1) for the k-label classifier and $h_{\mathbf{w}_k^*}(x_i)$ represents the probability that the x_i token belongs to class k. We acknowledge that the maximum confidence strategy, although producing the appropriate label predictions, ignores the inherent weakness of OVA, *i.e.*, OVA can lead to confusion areas where **(i)** two or more binary classifiers can be confident that the sample belongs to their classes, or **(ii)** no classifiers are confident enough to claim the sample, especially under the imbalanced settings [13], [23]. However, since AUC maximization can learn a well-tuned binary classifier under imbalanced settings, we argue that this weakness is less concerning, as shown in Figure 1. The results in the figure show that the NER model, trained with the AUC surrogate loss under the OVA setup, leads to smaller confusion areas (few samples in the lower left region) compared to the binary cross entropy loss, indicating that the binary classifiers, learnt with the AUC surrogate loss function, are more well-tuned than those learnt with the binary cross entropy loss. More statistical results and detailed discussions can be found in section V.

Algorithm 2 OVA-AUC-MAML First Order Approximation

Input: Training set $\mathcal{S}$, $\theta, \{\omega_1, \ldots, \omega_K\}$, $\{a, b, \alpha\}^K$
Output: $\theta^*, \{\omega_1^*, \ldots, \omega_K^*\}$

```
for epoch in range(num epochs) do
    set ℒ := 0 and sample m classes
    for i in range(K) do
        if i in m then
            ℒ+ = Equation (1)
        end if
    end for
    ℒ.optimize()
end for
```

IV. Experimental Settings

Domains and Corpora: We used corpora from both general and biomedical domains to benchmark the NER performance. Table I summarizes the label distribution statistics for these corpora. Both CoNLL 2003 [17] and OntoNotes5 [16] are benchmark corpora from the general domains. Whereas NCBI [7], and s800 [8] have been widely used in biomedical named entity recognition (bioNER) task [5], [26]. Since these corpora vary in terms of label distribution and linguistic characteristics, they serve to substantiate the versatility of our NER methods regardless of the underlying corpora or domains.

Model Architecture and Embedding: For embeddings and model architectures, we focused on state-of-the-art NER and bioNER architectures and embeddings. CoNLL 2003 and OntoNotes5 were trained with "bert-base-cased" [1], while NCBI and s800 were trained with "biobert-base-cased-v1.1" [26] to avoid out-of-vocabulary (OOV) issues in bioNER [5].

Baselines & OVA AUC NER Methods:

- **CE**: The standard multi-class cross entropy loss, most commonly used in almost all existing NER works, was used as one of the major baselines to verify and establish the significance and impact of our OVA AUC NER methods.
- **CRFs**: As our second baseline, CRFs were traditionally used in many NER works, such as those of [5], [27]. As CRFs produce a sequence labeller instead of a token classifier, we also used BiLSTM to push the performance of this baseline.
- **OVA-BCE**: Instead of using the AUC surrogate loss, this baseline simply uses the binary cross entropy loss (BCE) for each label. This baseline is to study the performance difference between BCE and the AUC surrogate loss under low-resource and imbalanced settings for ablation purposes.
- **OVA-AUC**: As OVA has higher training time compared to multi-class CE since each binary classifier should be independently trained, we thus consider grouping the binary classifiers for labels that share similar linguistic characteristics (*e.g.*, B-PER, B-ORG, B-MISC and B-LOC for CoNLL 2003) and train their classifiers together (see Algorithm 1).
- **OVA-AUC-MAML**: Additionally, we also reduce the training time by applying first-order meta-learning (MAML) [14], [28] and sample a random batch of m binary classifiers in each iteration for the model to optimize (see Algorithm 2).

It is noteworthy that the goal of our experiments is to compare the baseline objective functions and our OVA AUC objective functions; thus, to make the comparison fair, we used the same embeddings and language model for all the objective functions *e.g.*, Bert-CE, Bert-CRF, and Bert-BiLSTM-CRF v.s. Bert-OVA-AUC and Bert-OVA-AUC-MAML.

Evaluation Setups:

- The size of $\mathcal{S}$: In order to simulate the low-resource scenarios, we used training set $\mathcal{S}$ with size of $\{20, 50, 100, 200, 300, 400, 500\}$. We then trained our methods and all the baselines on 10 random training partitions of the same size and reported/analysed their average entity F1 performance.
- Imbalance entity generator: This was derived from [4] to sample $\mathcal{S}$ containing $\{1, 2, 5, 10\}$ percentage of entity tokens, *i.e.*, tokens with labels that are not "O". This is to investigate the robustness of our methods under different distributions.

V. Experimental Results & Discussions

A. Low-resource Studies

Using Figure 2, we observe the followings:

- **CE vs. OVA-AUC**: Under extreme low-resource scenarios (*i.e.*, size $\{20, 50\}$), OVA-AUC outperforms CE by a significant margin, with the average performance difference reaching 30%. When the training pool increases, OVA-AUC still exhibits substantial gains compared to CE at the 95% level of confidence for the vast majority of the time, indicating that OVA-AUC is a superior alternative to the standard multi-class CE loss function for low-resource NER.
- **CRF and BiLSTM-CRF vs. OVA-AUC**: It is apparent that OVA-AUC significantly outperforms CRF in all scenarios. As CRF is a sequence labeler, we further adopt BiLSTM to improve this baseline's performance. Nevertheless, Figure 2 suggests that OVA-AUC should remain a superior solution to both CRF and BiLSTM-CRF under the low-resource NER.
- **OVA-BCE vs. OVA-AUC**: The performance of OVA-BCE is close to that of CE for CoNLL 2003, s800, and NCBI datasets. However, its performance is sub-optimal for the Ontonotes 5 dataset, resulting in a noticeable gap compared to our OVA-AUC training strategies. This difference can be attributed to the higher number of classes in the Ontonotes 5 dataset ($K = 37$), which leads to an extremely imbalanced situation. One of the aforementioned weaknesses of the OVA approach is its inefficiency in learning from imbalanced data, often resulting in less accurate classifiers compared to multi-class counterparts. Therefore, these findings underscore the importance of combining both the OVA approach and AUC maximization to achieve optimal results for imbalanced NER.
- **OVA-AUC vs. OVA-AUC-MAML**: Although the performance of OVA-AUC can be higher on average than that of OVA-AUC-MAML under some settings, the difference between the two approaches is not significant, except for OntoNotes5 with a 200 training size. As OVA-AUC groups the labels that share similar linguistic characteristics and trains their classifiers together, the difference can be attributed to longer training time. As OVA-AUC-MAML performs on

Fig. 2: The average entity-F1 performance with 95% error bands on the test set taken from 10 random training partitions of each training set $\mathcal{S}$ size for each loss function. The title of the plot indicates the corpus, and the label distribution in BIO format.

Fig. 3: The average entity-F1 performance with error bands on test set taken from 10 random training partitions of each training set size for each loss function. The entity tag size indicates the percentage of entity-tokens to the total tokens in the training set.

par with OVA-AUC, it aptly outperforms all the selected baselines in most low-resource NER across all the corpora.

B. Imbalanced Data Distribution Studies

We deployed an imbalance entity tag generator [4] to demonstrate the robustness of our methods on the diverse training set distributions for the NER task. This generator simulates scenarios in which the training set $\mathcal{S}$ data distribution changes from that of $\mathcal{S}^{\text{test}}$ in order to test the resilience of the baselines and our methods. Figure 3 illustrates the performance differences for those methods according to the size of the entity tag. From these results, we provide the following observations:

- **CE vs. OVA-AUC**: Across all entity tags settings, we observed that OVA-AUC consistently outperforms CE. On the basis of the performance on CoNLL, NCBI, and s800, OVA-AUC is significantly superior to CE in the most extreme imbalanced scenarios (*i.e.*, entity label size of 1 and 2%). As the amount of entity tags rises in OntoNotes5, OVA-AUC performance improves greatly. Conversely, CE performs inadequately when the entity tag and training size are small.
- **CRF and BiLSTM-CRF vs. OVA-AUC**: OVA-AUC outperforms CRF and BiLSTM-CRF by a significant margin in all scenarios. Similar to CE, neither CRF nor BiLSTM-CRF performs when the entity tag size is extremely small as the sequence labelers are not fed with enough learning signals.
- **OVA-BCE vs. OVA-AUC**: The results further reinforce the previous discussion that AUC trained binary classifiers under OVA are important to the low-resource and imbalanced NER.

- **OVA-AUC vs. OVA-AUC-MAML**: We observe no significant difference between OVA-AUC and OVA-AUC-MAML based on the performance in most settings across all datasets.

Overall, while the results from subsection V-A suggests that OVA AUC NER methods are important under the low-resource NER, this subsection brings new evidence to the versatility of our works under different NER imbalanced label distributions.

VI. Conclusion

In this study, we offer efficient approaches to the challenges of low-resource and imbalanced data that are ubiquitous in many NER tasks via restructuring the conventional NER multi-class learning problem into a one-vs-all learning scenario, and utilizing an AUC loss to instruct the binary classifiers. Our experiments on diverse datasets in low-resource and data imbalance situations prove the superiority of our OVA AUC NER strategies over traditional methods like CE and CRF, regardless of the NER models, embeddings, or domains being employed. These findings not only present advancements in handling low-resource and imbalanced NER but also pave the way for future exploration, with potential ramifications in wider NLP areas grappling with data imbalance and resource scarcity.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[2] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, 2020.

[3] U. Yaseen and S. Langer, "Data augmentation for low-resource named entity recognition using backtranslation," *ArXiv*, vol. abs/2108.11703, 2021.

[4] N. D. Nguyen, W. Tan, W. Buntine, R. Beare, C. Chen, and L. Du, "AUC maximization for low-resource named entity recognition," 2022. [Online]. Available: https://arxiv.org/abs/2212.04800

[5] N. D. Nguyen, L. Du, W. Buntine, C. Chen, and R. Beare, "Hardness-guided domain adaptation to recognise biomedical named entities under low-resource scenarios," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 4063–4071. [Online]. Available: https://aclanthology.org/2022.emnlp-main.271

[6] R. Zhou, X. Li, R. He, L. Bing, E. Cambria, L. Si, and C. Miao, "MELM: Data augmentation with masked entity language modeling for low-resource NER," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 2251–2262.

[7] R. I. Doğan, R. Leaman, and Z. Lu, "Special report: NCBI disease corpus: A resource for disease name recognition and concept normalization," *J. of Biomedical Informatics*, vol. 47, p. 1–10, Feb. 2014.

[8] E. Pafilis, S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, and L. J. Jensen, "The species and organisms resources for fast and accurate identification of taxonomic names in text," *PLOS ONE*, vol. 8, no. 6, pp. 1–6, 06 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0065390

[9] W. Gao, R. Jin, S. Zhu, and Z.-H. Zhou, "One-pass AUC optimization," in *International conference on machine learning*. PMLR, 2013, pp. 906–914.

[10] Y. Ying, L. Wen, and S. Lyu, "Stochastic online AUC maximization," *Advances in neural information processing systems*, vol. 29, 2016.

[11] Z. Yuan, Y. Yan, M. Sonka, and T. Yang, "Large-scale robust deep AUC maximization: A new surrogate loss and empirical studies on medical image classification," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2021, pp. 3020–3029. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00303

[12] Z. Yang, Q. Xu, S. Bao, X. Cao, and Q. Huang, "Learning with multiclass AUC: Theory and algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[13] H. Liu, W. Zheng, G. Sun, Y. Shi, Y. Leng, P. Lin, R. Wang, Y. Yang, J.-f. Gao, H. Wang *et al.*, "Action understanding based on a combination of one-versus-rest and one-versus-one multi-classification methods," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2017, pp. 1–5.

[14] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[15] Z. Yuan, Z. Guo, N. Chawla, and T. Yang, "Compositional training for end-to-end deep AUC maximization," in *International Conference on Learning Representations*, 2021.

[16] R. Weischedel, S. Pradhan, L. Ramshaw, M. Palmer, N. Xue, M. Marcus, A. Taylor, C. Greenberg, E. Hovy, R. Belvin *et al.*, "Ontonotes release 5.0," *LDC2011T03, Philadelphia: Linguistic Data Consortium*, 2014.

[17] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, pp. 142–147. [Online]. Available: https://www.aclweb.org/anthology/W03-0419

[18] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes," *Pattern Recognition*, vol. 44, no. 8, pp. 1761–1776, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320311000458

[19] H. Liu, W. Zheng, G. Sun, Y. Shi, Y. Leng, P. Lin, R. Wang, Y. Yang, J.-f. Gao, H. Wang, K. Iramina, and S. Ge, "Action understanding based on a combination of one-versus-rest and one-versus-one multi-classification methods," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2017, pp. 1–5.

[20] J. Jang and C. O. Kim, "One-vs-rest network-based deep probability model for open set recognition," *CoRR*, vol. abs/2004.08067, 2020. [Online]. Available: https://arxiv.org/abs/2004.08067

[21] S. Padhy, Z. Nado, J. Ren, J. Z. Liu, J. Snoek, and B. Lakshminarayanan, "Revisiting one-vs-all classifiers for predictive uncertainty and out-of-distribution detection in neural networks," *CoRR*, vol. abs/2007.05134, 2020. [Online]. Available: https://arxiv.org/abs/2007.05134

[22] K. Saito and K. Saenko, "Ovanet: One-vs-all network for universal domain adaptation," *CoRR*, vol. abs/2104.03344, 2021. [Online]. Available: https://arxiv.org/abs/2104.03344

[23] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *The Journal of Machine Learning Research*, vol. 5, pp. 101–141, 2004.

[24] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of machine learning research*, vol. 4, no. Nov, pp. 933–969, 2003.

[25] P. Zhao, S. C. H. Hoi, R. Jin, and T. Yang, "Online AUC maximization," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ser. ICML'11. Madison, WI, USA: Omnipress, 2011, p. 233–240.

[26] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 09 2019. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz682

[27] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "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*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 260–270. [Online]. Available: https://www.aclweb.org/anthology/N16-1030

[28] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *CoRR*, vol. abs/1803.02999, 2018. [Online]. Available: http://arxiv.org/abs/1803.02999