MAJORITY OR MINORITY: DATA IMBALANCE LEARNING METHOD FOR NAMED ENTITY RECOGNITION

Sota Nemoto, Shunsuke Kitada, Hitoshi Iyatomi
Department of Applied Informatics, Graduate School of Science and Engineering, Hosei University
{sota.nemoto.5s@stu., shunsuke.kitada.8y@stu., iyatomi@}hosei.ac.jp

ABSTRACT

Data imbalance presents a significant challenge in various machine learning (ML) tasks, particularly named entity recognition (NER) within natural language processing (NLP). NER exhibits a data imbalance with a long-tail distribution, featuring numerous minority classes (i.e., entity classes) and a single majority class (i.e., $\mathcal{O}$-class). This imbalance leads to misclassifications of the entity classes as the $\mathcal{O}$-class. To tackle this issue, we propose a simple and effective learning method named majority or minority (MoM) learning. MoM learning incorporates the loss computed only for samples whose ground truth is the majority class into the loss of the conventional ML model. Evaluation experiments on four NER datasets (Japanese and English) showed that MoM learning improves prediction performance of the minority classes without sacrificing the performance of the majority class and is more effective than widely known and state-of-the-art methods. We also evaluated MoM learning using frameworks as sequential labeling and machine reading comprehension, which are commonly used in NER. Furthermore, MoM learning has achieved consistent performance improvements regardless of language or framework.

1 INTRODUCTION

Named entity recognition (NER) (Nadeau & Sekine, 2007; Lample et al., 2016) is one of many real-world natural language processing (NLP) tasks with significant data imbalance, especially when applied for business purposes like corporate information-gathering websites (Guo et al., 2009) and extracting drug names and diseases from vast amounts of unstructured medical data (Ramachandran & Arutchelvan, 2021). NER commonly uses a sequential labeling framework, a form of multiclass classification that predicts labels corresponding to the words in a sentence. In sequential labeling, all words are divided into either entity words with information (i.e., proper nouns) or non-entity words without information. Each entity word is labeled as a specific class (PERSON, LOCATION, etc.) to which a few samples belong. In contrast, all non-entity words constitute the majority and are labeled as a single class (i.e., the "others" $\mathcal{O}$-class). This labeling yields a data imbalance with a long-tail distribution. Between the well-known benchmarks CoNLL2003 (Sang & De Meulder, 2003) and OntoNotes5.0 (Pradhan et al., 2013), the number of samples for the $\mathcal{O}$-class significantly exceeds that of the entity class, a condition that often leads to misclassifications of entity classes as the $\mathcal{O}$-class, causing a considerable decline in the prediction performance of the minority classes. Overall, overcoming this data imbalance is a crucial step toward enhancing NER performance.

Conventional machine learning (ML) methods for addressing data imbalances are categorized into sampling-based methods (Pouyanfar et al., 2018; Buda et al., 2018) for inputs and cost-sensitive learning (Adel et al., 2017; Madabushi et al., 2019; Li & Xiao, 2020) for outputs. The sampling-based method, which adjusts the number of sentences in training, has a certain effect on the ML tasks. However, NER uses sequential labeling, which predicts the labels corresponding to each word in a sentence; thus, it does not mitigate the imbalance. By contrast, cost-sensitive learning addresses the imbalance by designing a loss function for the ML model based on the number of samples in each class. While it is effective for binary classification, NER is a multiclass classification requiring extension of this method. This extension will lead to complex weight adjustments for each class and for cases in which it is not fully capable, thus not attaining the desired level of performance.

In this paper, we propose a novel learning method, majority or minority (MoM) learning, to tackle the data imbalance in NER. MoM learning is simple and effective for incorporating the loss computed only for samples whose ground truth is the majority class into the loss of the conventional ML model. Our strategy enables cost-sensitive learning but differs from the concepts of previous studies because it does not depend on the difficulty of the classification or the number of samples in the class. The purpose of MoM learning is to enhance performance by preventing misclassifications of the minority classes (entity classes) as the majority class (the $\mathcal{O}$-class). When incorporating the loss of entity classes instead of the $\mathcal{O}$-class, the model cannot distinguish whether the prediction is misclassified as the $\mathcal{O}$-class or as another entity class. Therefore, MoM learning focuses on the $\mathcal{O}$-class to recognize misclassifications from the $\mathcal{O}$-class to the entity classes.

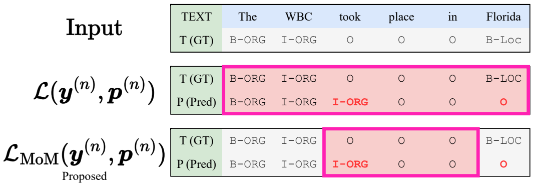
We evaluated MoM learning using four NER datasets and with a ML model, including BERT (Devlin et al., 2019), which have proven successful in various NLP tasks. The evaluation results demonstrated that MoM learning contributes to consistent improvements in performance across languages. We also confirmed that MoM learning is more effective than those introduced in previous state-of-the-art studies, such as focal loss (FL) (Lin et al., 2017) and dice loss (DL) (Li et al., 2020b). Furthermore, beyond common sequential labeling, we demonstrated the effectiveness of MoM learning using the machine reading comprehension (MRC) framework, which is becoming mainstream (Li et al., 2020a; Zhang & Zhang, 2023).

Our contributions are summarized as follows. (1) We propose a novel learning method, majority or minority (MoM) learning, designed to address the data imbalance with a long-tail distribution, which is a significant challenge. (2) We evaluated four common NER datasets (English and Japanese) and demonstrated that MoM learning is more effective than conventional methods of both the sequential labeling framework and MRC. (3) MoM learning improved the performance of the entity classes without compromising the performance of the $\mathcal{O}$-class in a language-agnostic context.

2 Proposed Method

2.1 Notation

First, as a common approach to NER, we introduce the notation for sequential labeling. We consider a dataset comprising a set of input sentences $\boldsymbol{X} = [\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(n)}, \cdots, \boldsymbol{x}^{(N)}]$ and the corresponding training labels $\boldsymbol{Y} = [\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(n)}, \cdots, \boldsymbol{y}^{(N)}]$, where N is the number of sentences in the dataset. The n-th sentence split tokens and corresponding training labels are represented as $\boldsymbol{x}^{(n)} = [\boldsymbol{w}_1^{(n)}, \cdots, \boldsymbol{w}_i^{(n)}, \cdots, \boldsymbol{w}_M^{(n)}]$ and $\boldsymbol{y}^{(n)} = [\boldsymbol{y}_1^{(n)}, \cdots, \boldsymbol{y}_i^{(n)}, \cdots, \boldsymbol{y}_M^{(n)}]$, respectively. M is the number of tokens in the longest sentence in the dataset, and shorter sentences are padded up to M.


Input
$\mathcal{L}(\boldsymbol{y}^{(n)}, \boldsymbol{p}^{(n)})$
$\mathcal{L}_{\mathrm{MoM}}(\boldsymbol{y}^{(n)}, \boldsymbol{p}^{(n)})$
Proposed

TEXT	The	WBC	took	place	in	Florida
T (GT)	B-ORG	I-ORG	O	O	O	B-Loc

T (GT)	B-ORG	I-ORG	O	O	O	B-LOC
P (Pred)	B-ORG	I-ORG	I-ORG	O	O	O

T (GT)	B-ORG	I-ORG	O	O	O	B-LOC
P (Pred)	B-ORG	I-ORG	I-ORG	O	O	O

Figure 1: Concept of MoM learning. The conventional loss function, $\mathcal{L}$ (e.g., cross-entropy loss), calculates the loss for all samples. In MoM learning, $\mathcal{L}_{\mathrm{MoM}}$, the loss associated with the "major" $\mathcal{O}$-class, is added to $\mathcal{L}$.

The training label $\boldsymbol{y}_i^{(n)}$ is annotated using the BIO format (Ramshaw & Marcus, 1999) in sequential labeling. This format consists of entity classes (e.g., PER, LOC, and ORG) and a non-entity class (i.e., the $\mathcal{O}$-class); where the former are represented by prefixing the entity category with B for the first token and I for the rest, as follows: B-PER, I-PER, B-LOC, etc. The sequence of predicted labels is denoted as $\boldsymbol{p}^{(n)} = [\boldsymbol{p}_1^{(n)}, \cdots, \boldsymbol{p}_i^{(n)}, \cdots, \boldsymbol{p}_M^{(n)}]$, where the ML model estimates the predicted probabilities $\boldsymbol{p}^{(n)}$ for each token of the sentence $\boldsymbol{x}^{(n)}$.

2.2 Majority or Minority (MoM) Learning

MoM learning is a simple and effective method that incorporates the loss for samples whose ground truth is a single majority class into the loss of an arbitrary conventional ML model. Fig. 1 illustrates the concept of MoM learning, where conventional loss $\mathcal{L}$ represents an arbitrary loss function of the model, such as cross-entropy, which computes the loss for all samples boxed in red. The $\mathcal{L}_{\mathrm{MoM}}$ term only computes the loss whose ground truth is the $\mathcal{O}$-class (i.e., $\boldsymbol{y}_i^{(n)} =$ "$\mathcal{O}$") framed in red, and incorporates them into the conventional loss. The equation for the $\mathcal{L}_{\mathrm{MoM}}$ is

$$\mathcal{L}_{\mathrm{MoM}}(\boldsymbol{y}^{(n)}, \boldsymbol{p}^{(n)}) = -\frac{1}{M} \sum_{\boldsymbol{y}_i^{(n)}=\text{``}\mathcal{O}\text{''}}^{M} \ell(\boldsymbol{y}_i^{(n)}, \boldsymbol{p}_i^{(n)}), \tag{1}$$

where ℓ is an arbitrary loss function, including cross-entropy, weighted cross-entropy, FL (Lin et al., 2017), DL (Li et al., 2020b), etc. Because $\mathcal{L}_{\mathrm{MoM}}$ focuses only on the $\mathcal{O}$-class, certain entity classes misclassified by the model become inconsequential. Hence, $\mathcal{L}_{\mathrm{MoM}}$ functions as a pseudo-binary classification, distinguishing between the $\mathcal{O}$-class and the entity classes to detect misclassifications of $\mathcal{O}$-class as entity classes. MoM learning enables independence from such factors as the number of class samples, task features, and the model employed, making it adaptable to similarly imbalanced tasks.

For the n-th sentence $\boldsymbol{x}^{(n)}$, the loss function $\mathcal{L}_{\mathcal{S}}$, when applying MoM learning, is written as

$$\mathcal{L}_{\mathcal{S}}(\boldsymbol{y}^{(n)}, \boldsymbol{p}^{(n)}) = \lambda \cdot \mathcal{L}(\boldsymbol{y}^{(n)}, \boldsymbol{p}^{(n)}) + (1-\lambda) \cdot \mathcal{L}_{\mathrm{MoM}}(\boldsymbol{y}^{(n)}, \boldsymbol{p}^{(n)}), \tag{2}$$

where λ is a hyperparameter balancing $\mathcal{L}$ and $\mathcal{L}_{\mathrm{MoM}}$. MoM learning simplifies weights adjustments compared to WCE, with a single hyperparameter λ. Finally, the model loss $\mathcal{L}_{\mathcal{M}}$ is minimized with the training labels for the entire dataset $\boldsymbol{Y}$ and the prediction probabilities $\boldsymbol{P}$:

$$\mathcal{L}_{\mathcal{M}}(\boldsymbol{Y}, \boldsymbol{P}) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{\mathcal{S}}(\boldsymbol{y}^{(n)}, \boldsymbol{p}^{(n)}). \tag{3}$$

3 EXPERIMENTS

This section describes the dataset, followed by the evaluation experiments used, including sequential labeling and MRC, loss functions, and implementation details. Considering data variability, the evaluation was based on the average of the results of 10 random seeds in each condition. In all evaluations, we performed paired t-tests ($\alpha = 0.05$) to identify differences between our method and other leading methods where α is the significance level.

3.1 DATASETS

We used the following four datasets: English CoNLL2003 (Sang & De Meulder, 2003), English OntoNotes5.0 (Pradhan et al., 2013), Kyoto University web document read corpus (KWDLC; in Japanese) (Hangyo et al., 2012), and Stockmark NER wiki (NER wiki; in Japanese) (Omi, 2021). In addition, we evaluated four datasets using sequential labeling. For MRC, we used CoNLL2003, which has been adopted in previous studies, by converting the data from sequential labeling annotations. For the English datasets, we employed the standard training, validation, and test data provided, while for the Japanese datasets without those standard aspects, we randomly split the data on an 8:1:1 basis.

3.2 NER FRAMEWORKS AND LOSS FUNCTIONS

We compare two frameworks. The **sequential labeling framework** classifies at the token level and yields a data imbalance with a long-tail distribution. This framework directly addresses NER as a multiclass classification. Thus, we used the macro F1 score as an evaluation criterion. Compared to sequential labeling, the **MRC framework**, another practical option for NER, has been widely used in recent years (Li et al., 2020a) in binary classification tasks. This framework determines whether each word belongs to a particular class and finds its range. Specifically, for a token $\boldsymbol{w}_i^{(n)}$ in a sentence $\boldsymbol{x}^{(n)}$, the ground truth can be written as $\boldsymbol{y}_i^{(n)} \in \{0,1\}^{\mathcal{Y}}$, where $\mathcal{Y}$ is the set of entity and non-entity classes. Unlike the macro F1 score for sequential labeling, we used the macro F1 score, which matches the index of the predicted start and end points.

Table 1: Performance with BERT in sequential labeling (in macro F1). In all items, MoM had the best score, with a significant difference from FL, which was the next best ($\alpha = 0.05$).

	CoNLL2003			**OntoNotes5.0**			**KWDLC**			**NER Wiki**		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
BERT	90.16	91.86	91.00	87.41	89.07	88.23	70.92	73.96	72.41	77.32	81.04	79.13
w/ WCE-1	89.73	92.15	90.93 (-0.07)	85.66	90.28	87.91 (-0.32)	62.79	78.32	69.70 (-2.71)	73.65	80.28	76.82 (-2.31)
w/ WCE-2	89.94	92.22	91.07 (+0.07)	86.81	89.67	88.22 (-0.01)	68.86	77.23	72.80 (+0.39)	75.72	81.19	78.36 (-0.77)
w/ FL	90.33	92.03	91.17 (+0.17)	87.62	89.15	88.39 (+0.16)	71.88	74.27	73.05 (+0.64)	77.79	81.53	79.61 (+0.48)
w/ MoM (proposed)	90.41	92.27	**91.33** (+0.33)	87.39	89.84	**88.60** (+0.37)	72.54	74.13	**73.32** (+0.91)	78.13	81.61	**79.83** (+0.70)

Table 2: Comparison of performance in each entity in sequential labeling of CoNLL2003.

	w/ MoM			**BERT**		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
MISC	79.18	84.58	**81.78**	79.21	84.04	81.54
LOC	92.91	93.50	**93.20**	92.60	93.49	93.04
ORG	89.87	93.17	**91.54**	89.90	92.70	91.27
PER	97.66	97.88	**97.77**	97.76	97.55	97.65
$\mathcal{O}$	99.72	99.28	**99.50**	99.69	99.31	**99.50**

Table 3: Summary of the performance of MRC on CoNLL2003.

	Prec.	**Rec.**	**F1**
BERT-MRC	92.43	92.22	92.32
w/ FL	92.95	92.10	92.52 (+0.20)
w/ DL	92.69	92.43	92.56 (+0.24)
w/ MoM (proposed)	92.99	92.51	**92.75** (+0.43)

We compared the prediction performance of MoM learning with that of conventional learning methods (i.e., loss functions) that have long been considered state-of-the-art and used widely for data imbalance issues. The **WCE** is one of the most commonly used weighted loss functions, and we consider two variants: the inverse class frequency (WCE-1) and a hyperparameter related to the number of samples (WCE-2). In our experiments, WCE-1 is set as the inverse class frequency and WCE-2 is set as $\log_{10}(\frac{s-s_k}{s_k} + \beta)$, used in a DL paper (Li et al., 2020b), where s_k is the number of samples for class k, s is the total number of train samples, and β is a hyperparameter. The **FL** (Lin et al., 2017) is a more robust and versatile loss (Iikura et al., 2021; Liu et al., 2021) that was proposed after WCE. Because FL was designed for binary classification, we extended the FL with a one-versus-the-rest method in sequential labeling. The **DL** (Li et al., 2020b) has long been considered a state-of-the-art method focused on MRC and was designed to reduce both false positives and false negatives.

3.3 Implementation Details

Models. We utilized pre-trained models; the input length of these models was determined by the maximum number of tokens in a sentence ($M = 128$), with padding tokens (`[PAD]`) used for filling the remaining space to maintain a consistent length. We fine-tuned in 10 epochs using the Adam optimizer (Kingma & Ba, 2014) for each task. **Sequential labeling**: We used pre-trained BERT (Devlin et al., 2019) as the baseline model. We set D to 768, a learning rate of 2×10^{-5}, and a batch size of 64. **MRC**: We used BERT-MRC (Li et al., 2020a) as the baseline model and set a learning rate of 3×10^{-5} and a batch size of 32.

Hyperparameters. The tree-structured Parzen estimator (TPE) (Bergstra et al., 2011), implemented in the Bayesian optimization library Optuna (Akiba et al., 2019), was used to maximize the F1 score of the validation data. For the sequential labeling experiments, the hyperparameters of WCE-2 (β) and FL (γ) were explored in the predetermined range of 1.0–10.0 and 0.0–10.0, respectively, considering their papers (Lin et al., 2017; Li et al., 2020b). For the MRC experiments, we set the hyperparameters of FL $\gamma = 3.0$ and those of DL $\epsilon = 1.0$ and $\delta = 0.01$, which is based on the DL hyperparameters carefully tuned in Li et al. (2020b). The MoM hyperparameter (λ) was explored in the predetermined range of 0.0–1.0 in both frameworks. Thus, in the sequential labeling, we set the MoM hyperparameter λ for the datasets CoNLL2003, OntoNotes5.0, KWDLC, and NER wiki to 0.175, 0.125, 0.357, and 0.212, respectively, while for MRC CoNLL2003 was set λ to 0.446.

4 Results

Table 1 presents a comparison of the performances of each method using BERT in sequential labeling. We confirmed that MoM learning consistently outperforms other methods across all four datasets. In all datasets, the performance using MoM learning was significant at $\alpha = 0.05$ against the next best method (FL). The results using WCE-1 and WCE-2 demonstrated poor performance compared to the baseline; the reason is discussed in Sec. 5.

Table 2 presents a comparison between the baseline with and without MoM learning for each entity in CoNLL2003. The prefixes of the entity classes `B` and `I` are merged to show the average performance of the respective classes, resulting in nine classes (e.g., `B-PER`, `I-PER`, `B-LOC` and $\mathcal{O}$) becoming five classes (e.g., `PER`, `LOC`, and $\mathcal{O}$). We confirm that MoM learning improves the performance of entity classes without compromising the performance of the $\mathcal{O}$-class.

Table 3 presents the performance with the CoNLL2003 dataset using the MRC. The results confirm MoM learning also demonstrated the best performance and was significant at $\alpha = 0.05$, against the next best method (DL).

5 Discussion

The most important factor in NER is the score of the entity classes, rather than the overall score, including the $\mathcal{O}$-class, as the prediction performance of NER generally concerns the score including the $\mathcal{O}$-class. In practical situations in which entities are extracted and utilized, the performances of the entity classes hold greater significance. Although MoM learning appears to be a marginal improvement, we confirm that MoM learning improves the performance of minor entity classes without sacrificing the performance of the major $\mathcal{O}$-class, regardless of language.

For WCE, we attempted two methods (i.e., WCE-1 and 2); however, a poorer performance than the baseline CE was observed, highlighting the challenges posed by multiclass NER, with its inherent long-tail distribution. As evidenced by various ML tasks, conventional weighting methods struggle with the delicate design of loss functions dependent on specific datasets and tasks (Valverde et al., 2017; Jadon, 2020). Thus, the experiments in Table 1 highlight the difficulty of applying WCE weighting to the sequential labeling of NER.

MoM learning was effective at sequential labeling and MRC, especially the latter, where we observe the role of MoM learning in monitoring the number of entities. For example, in the sentence “Estadio Santiago Bernabéu opened in 1974.”, “Estadio”, “Santiago”, and “Bernabéu” are assigned to the classes `B-LOC`, `I-LOC`, `I-LOC`, and the other words are assigned to the $\mathcal{O}$-class. When considering a basic sentence, it is highly likely that the word (“opened”) following the last word (“Bernabéu”) of the entity belongs to be `I-LOC` or $\mathcal{O}$-class because sequences of different entity words are extremely rare, such as `LOCATION` after `PERSON`. Because MoM learning focuses more on $\mathcal{O}$-class words, the model can learn whether the final word belongs to the `I-LOC` or $\mathcal{O}$-class. In other words, the MoM can monitor how many entity words are consecutive, which is a factor in the improved performance of the words at the end of the entity.

6 Conclusion

In this paper, we have proposed a novel learning method, MoM learning, to address NER tasks characterized by a data imbalance with a long-tail distribution consisting of a single class with many samples (the majority class) and multiple classes with a few samples (the minority classes). MoM learning is a simple and effective method that suppresses misclassifications of majority as minority classes by incorporating the loss of samples in which ground truth is the majority class into the loss of conventional ML models. Evaluation experiments using four datasets (two each in English and Japanese) showed that MoM learning outperforms existing and even state-of-the-art methods in addressing data imbalances regardless of language or framework.

References

Heike Adel, Francine Chen, and Yan-Ying Chen. Ranking convolutional recurrent neural networks for purchase stage identification on imbalanced Twitter data. In *Proc. of EACL*, pp. 592–598, 2017.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proc. of KDD*, pp. 2623–2631, 2019.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL-HLT*, pp. 4171–4186, 2019.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. Named entity recognition in query. In *Proc. of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 267–274, 2009.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. Building a diverse document leads corpus annotated with semantic relations. In *Proc. of PACLIC*, pp. 535–544, 2012.

Riku Iikura, Makoto Okada, and Naoki Mori. Improving bert with focal loss for paragraph segmentation of novels. In *Distributed Computing and Artificial Intelligence, 17th International Conference*, pp. 21–30. Springer, 2021.

Shruti Jadon. A survey of loss functions for semantic segmentation. In *Proc. of CIBCB*, pp. 1–7, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

Jinfen Li and Lu Xiao. syrapropa at semeval-2020 task 11: Bert-based models design for propagandistic technique and span detection. In *Proc. of SemEval*, pp. 1808–1816, 2020.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified mrc framework for named entity recognition. In *Proc. of ACL*, pp. 5849–5859, 2020a.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. In *Proc. of ACL*, pp. 465–476, 2020b.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proc. of ICCV*, pp. 2980–2988, 2017.

Jianyi Liu, Xi Duan, Ru Zhang, Youqiang Sun, Lei Guan, and Bingjie Lin. Relation classification via bert with piecewise convolution and focal loss. *Plos one*, 16(9):e0257092, 2021.

Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. Cost-sensitive bert for generalisable sentence classification on imbalanced data. In *Proc. of NLP4IF*, pp. 125–134, 2019.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

Takahiro Omi. stockmarkteam/ner-wikipedia-dataset: Japanese Named Entity Extraction Dataset using Wikipedia, 2021. URL `https://github.com/stockmarkteam/ner-wikipedia-dataset`.

Samira Pouyanfar, Yudong Tao, Anup Mohan, Haiman Tian, Ahmed S Kaseb, Kent Gauen, Ryan Dailey, Sarah Aghajanzadeh, Yung-Hsiang Lu, Shu-Ching Chen, et al. Dynamic sampling in convolutional neural networks for imbalanced data classification. In *Proc. of MIPR*, pp. 112–117, 2018.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards robust linguistic analysis using ontonotes. In *Proc. of CoNLL*, pp. 143–152, 2013.

R Ramachandran and K Arutchelvan. Named entity recognition on bio-medical literature documents using hybrid based approach. *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–10, 2021.

Lance A Ramshaw and Mitchell P Marcus. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pp. 157–176. 1999.

Erik Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proc. of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003.

Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, and Xavier Lladó. Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach. *NeuroImage*, 155:159–168, 2017.

Yuzhe Zhang and Hong Zhang. Finbert–mrc: Financial named entity recognition using bert under the machine reading comprehension paradigm. *Neural Processing Letters*, pp. 1–21, 2023.