# BiLSTM-BCRF Model for Low-Resource Named Entity Recognition

**Authors:** Zhong Maosheng, Wu Jiahua, Wu Jiahua

**Date:** 2022-01-02T15:45:41+00:00

## Abstract

[Objective] When annotated data is scarce, existing models are constrained by the limited amount of training data, and their parameters are not adequately fitted, resulting in poor model recognition performance in low-resource named entity recognition tasks. [Method] This paper proposes a novel loss function incorporating Bernoulli distribution (Bernoulli distribution) to enable the model to better fit the data. Furthermore, based on the BiLSTM-CRF model, this work integrates multi-layer character feature information and combines it with the novel loss function based on Bernoulli distribution to construct the BiLSTM-BCRF model. [Results] The BiLSTM-BCRF model proposed in this paper achieves F1 score improvements of 6.16% and 3.35% over the BiLSTM-CRF model on 20% of the CoNLL2003 dataset and 20% of the BC5CDR dataset, respectively. [Conclusion] This model demonstrates good adaptability to low-resource named entity recognition tasks. [Limitation] The model' s performance in recognizing proper nouns requires further improvement.

## Full Text

## BiLSTM-BCRF Model for Low-Resource Named Entity Recognition

**Zhong Maosheng, Wu Jiahua**
School of Computer Information Engineering, Jiangxi Normal University

### Abstract

**[Objective]** When annotated data is scarce, existing models are constrained by limited training data, causing parameters to remain underfitted and resulting in poor recognition performance on low-resource named entity recognition (NER) tasks. **[Methods]** This paper proposes a novel loss function integrated with Bernoulli distribution to enable better model fitting. Furthermore, we

construct the BiLSTM-BCRF model by fusing multi-layer character feature information into the BiLSTM-CRF framework and incorporating the Bernoulli-based loss function. [**Results**] On 20% of the CoNLL2003 dataset and 20% of the BC5CDR dataset, the proposed BiLSTM-BCRF model achieves F1-score improvements of 6.16% and 3.35% respectively over the baseline BiLSTM-CRF model. [**Conclusion**] The model demonstrates strong adaptability to low-resource NER tasks. [**Limitations**] The model' s performance on proper noun recognition requires further enhancement.

**Keywords:** low-resource named entity recognition; neural networks; Bernoulli distribution

Named entity recognition is a fundamental task in natural language processing that aims to automatically identify entities from unstructured text and classify them into predefined categories such as person names, locations, and organizations. For example, the sentence "张无忌，金庸武侠小说《倚天屠龙记》人物角色，中土明教第三十四代教主" contains the person entity "张无忌，金庸", the book title entity "倚天屠龙记", and the organization entity "明教". This illustrates that entity recognition forms the basis of semantic text understanding. NER technology also finds extensive applications in knowledge graph construction, machine translation, and knowledge base development.

In recent years, deep learning methods have been widely adopted for NER. Hammerton [**?**] applied Long Short-Term Memory (LSTM) networks to entity recognition research, establishing the LSTM-CRF architecture as a foundational structure. Lample et al. [**?**] proposed a model combining Bidirectional LSTM (Bi-LSTM) with Conditional Random Fields (CRF) [**?**]. While these methods achieve excellent performance on text entity recognition tasks, they require large-scale annotated data with manual labeling of each word in the training corpus. Under conditions of insufficient annotated data, existing model parameters cannot be adequately fitted, causing the predicted maximum-probability tags to diverge from true labels and degrading model performance. This limitation hinders application in domains such as biology and medicine where annotated corpora are scarce. To address this issue, we propose a novel loss function incorporating Bernoulli distribution that enables better parameter fitting in low-resource scenarios. Additionally, to increase vocabulary coverage and improve recognition of rare words, we fuse multi-layer character feature information into the BiLSTM-CRF model, further enhancing precision and recall.

The remainder of this paper is organized as follows: Section 2 reviews related work in low-resource NER. Section 3 introduces our proposed model, including the input layer, BiLSTM layer, and BCRF layer. Section 4 presents experimental data, methodology, results, and analysis. Finally, we conclude with a summary of our work.

## 2 Related Work

Research methods for named entity recognition primarily include rule and dictionary-based approaches, machine learning methods, and deep learning methods. Rule and dictionary-based methods rely heavily on handcrafted templates from linguists, making them error-prone and poorly portable. Traditional machine learning methods mainly include Hidden Markov Models (HMM), Maximum Entropy (ME) [**?**], Maximum Entropy Markov Models (MEMM) [**?**], and Conditional Random Fields (CRF) [**?**]. These approaches require manual feature engineering and large-scale annotated corpora for training, with performance heavily dependent on the discriminative power of the selected features.

CRF is considered the mainstream model for NER, with the advantage of leveraging both internal and contextual features during the labeling process. With the continuous development of deep learning, research focus has shifted toward deep neural networks. Collobert et al. [**?**] first proposed a neural network-based NER method where each word has a fixed-size window, but this failed to address the long-tail problem. To overcome this limitation, Chiu and Nichols [**?**] proposed a bidirectional LSTM-CNN architecture that automatically detects word and character-level features. Hammerton [**?**] leveraged CRF' s ability to capture contextual features and proposed the LSTM-CRF model.

In recent years, numerous deep learning methods have been applied to low-resource NER, making it a prominent research area. Performance improvements in low-resource NER are prerequisites for widespread practical application. Related research can be broadly categorized into cross-lingual transfer methods, data augmentation methods, integration of automatically annotated corpora, and other approaches.

Cross-lingual transfer methods leverage annotated data from resource-rich languages to assist NER in low-resource languages, and can be divided into data transfer and model transfer approaches. Data transfer methods typically use text translation and label projection to convert annotated data from source languages into target language annotations for model training. Ni et al. [**?**] proposed a label mapping method on corpora to create automatically labeled target language data. Mayhew et al. [**?**] utilized bilingual dictionaries to automatically translate source language annotated text using a phrase-based machine translation approach [**?**]. Model transfer methods typically learn language-independent features first, then train NER models on source language annotated corpora for direct application to target languages. Chen et al. [**?**] employed adversarial learning to extract language-independent features and dynamically compute similarity between source and target languages for more effective knowledge transfer from multiple source languages. Keung et al. [**?**] further used adversarial learning [**?**] on multilingual BERT to learn better language-independent features.

Data augmentation aims to improve model robustness by adding reasonable

noise without increasing manual annotation costs, which significantly benefits model performance in low-data scenarios. Dai et al. [**?**] introduced random word replacement operations to increase training corpus diversity. Chen et al. [**?**] introduced local additivity-based data augmentation for semi-supervised NER. While language transfer and data augmentation effectively alleviate annotated data scarcity, languages with rich annotated resources remain very limited. Consequently, researchers have proposed integrating automatically annotated corpora by first automatically labeling large amounts of data, then integrating them to improve low-resource NER model performance. Yang et al. [**?**] first automatically annotated corpora using dictionary matching, then trained entity recognition models on small amounts of manually annotated data and large amounts of automatically annotated data using Partial-CRF [**?**]. Additionally, they trained a selector based on reinforcement learning [**?**] to filter out noisy annotations.

Beyond these three categories, other methods exist in low-resource NER. Zhang proposed a progressive knowledge distillation method called PDALN [**?**] that effectively adapts high-resource domains to low-resource target domains. Chen proposed a fine-tuning method for low-resource language models [**?**] that uses attention-based fine-tuning strategies to select relevant semantic and syntactic information from pre-trained language models for NER tasks. Our work primarily explores deep learning-based NER methods under low-resource conditions.

## 3 Model

### 3.1 Basic Architecture

The NER task is formulated as a sequence labeling problem. An input sentence is represented as $\langle MATH_0 \rangle$, where $\langle MATH_1 \rangle$ denotes the $i$-th character (including numbers, words, letters, punctuation, etc.). The output label sequence is $\langle MATH_2 \rangle$, where $\langle MATH_3 \rangle \in \{B, M, E, S, O\}$ is the label for $\langle MATH_4 \rangle$, with $B$, $M$, $E$, $S$, and $O$ representing the beginning, middle, end, single-character entity, and non-entity respectively. NER essentially classifies each character into one of these five categories.

We integrate Bernoulli distribution into the loss function of the BiLSTM-CRF model and fuse multi-layer character information to construct the BiLSTM-BCRF model. The basic structure is shown in Figure 1

. The model consists of three main components: an input layer, a Bi-LSTM layer, and a BCRF layer.

### 3.2 Input Layer

As illustrated in Figure 2 [FIGURE:2], the input layer structure uses word embeddings $\mathbf{x}$ generated from GloVe English word vectors [**?**], and character embeddings $\mathbf{c}$ trained by BiLSTM. The word and character vectors are concatenated and fed into the BiLSTM layer.

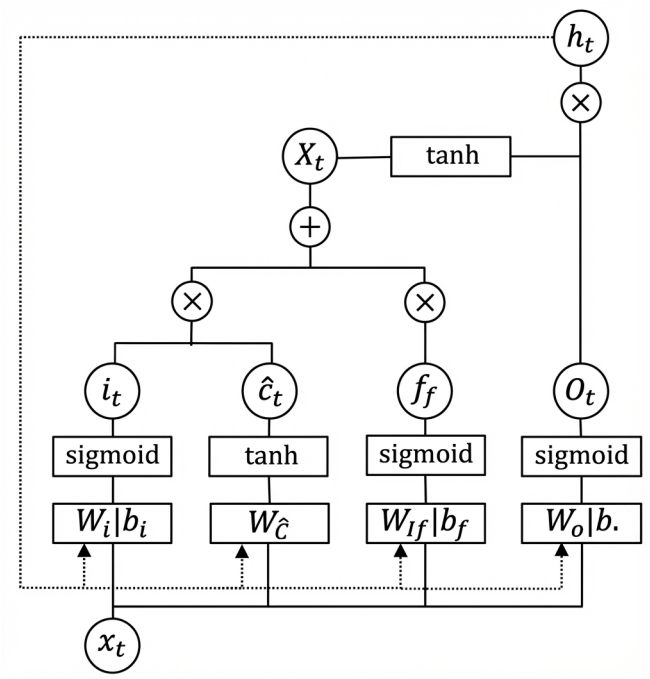


Figure 1: Figure 1

### 3.3 BiLSTM Layer

LSTM neural networks demonstrate strong modeling capabilities for NER tasks, effectively learning feature information of words and characters in text. The BiLSTM layer comprises two LSTMs. The LSTM network structure consists of three stages: forget, selective memory, and output. The LSTM unit structure is shown in Figure 3 [FIGURE:3], where $\langle MATH_5 \rangle$, $\langle MATH_6 \rangle$, and $\langle MATH_7 \rangle$ represent the input, forget, and output gates respectively in the LSTM unit at time $t$. $\langle MATH_8 \rangle$ denotes the hidden state at time $t-1$, and $\langle MATH_9 \rangle$ represents the cell memory state at time $t$. $\sigma$ denotes the sigmoid activation function, and tanh denotes the hyperbolic tangent activation function, as shown in equations (1)-(6):

The hidden state outputs in the BiLSTM neural network are $\langle MATH_{10} \rangle$, representing the forward and backward outputs respectively.

### 3.4 BCRF Layer

The decoding layer in conventional NER models is primarily the Conditional Random Field (CRF). CRF consists of state feature functions (also called emission probabilities) and state transition feature functions, which can be represented by a transition matrix in the model. The final conditional probability is given by equation (7):

$$\langle MATH_{11} \rangle$$

where $\langle MATH_{12} \rangle$ is a mapping of feature vectors for $\mathbf{x}$ and $\mathbf{y}$. $\langle MATH_{13} \rangle$ represents the probability of obtaining label sequence $\mathbf{y}$ given text sequence $\mathbf{x}$. The loss function is calculated as shown in equation (8):

$$\langle MATH_{14} \rangle$$

The advantage of CRF is its ability to consider dependencies among sequence labels. During training, the Viterbi algorithm is used for maximum likelihood estimation to predict the label sequence with maximum probability for input text, as shown in equation (9):

$$\langle MATH_{15} \rangle$$

where $\langle MATH_{16} \rangle$ represents the maximum probability of the model' s predicted labels. However, in low-resource scenarios, models constrained by limited annotated data suffer from underfitted parameters, causing the predicted maximum-probability label sequence to not necessarily be the true label sequence, which degrades final recognition performance. Drawing inspiration from Jie et al. [**?**], who employed cross-validation for training on incompletely annotated NER

tasks, we integrate Bernoulli distribution into the CRF loss function to construct a novel loss function, with the corresponding decoding model termed BCRF.

The Bernoulli distribution, also known as the two-point or 0-1 distribution, describes a discrete random variable $X$ with parameter $P$ $(0 < P < 1)$ that takes value 1 with probability $P$ and value 0 with probability $1 - P$. It is a special case of the binomial distribution when $N = 1$. Integrating the Bernoulli distribution function into CRF yields the new loss function shown in equation (10):

$$\langle MATH_{17} \rangle$$

A distribution function $q$ is added to the original loss function calculation, as shown in equation (11). The distribution function $q$ takes values of 0 or 1, following a Bernoulli distribution.

$$\langle MATH_{18} \rangle$$

where $\langle MATH_{19} \rangle$ represents the model' s maximum-probability predicted label and $\langle MATH_{20} \rangle$ represents the true label. In equation (10), when the predicted labels match the true labels for words in a sentence, the resulting loss value is smaller. Conversely, more incorrect predictions in a sentence yield larger loss values. Using this new loss function, model parameters can be better fitted even with limited training samples, improving precision and recall and ultimately enhancing recognition effectiveness.

## 4 Experiments

### 4.1 Datasets and Evaluation Metrics

We evaluate our model on the CONLL2003 [**?**] and BC5CDR [**?**] datasets. CONLL2003 contains four entity types across English and German languages, while BC5CDR includes two entity types from 1,500 medical articles. Since NER requires correct identification of both entity boundaries and categories, we use Precision, Recall, and F1-score to measure model performance, as shown in equations (12)-(14):

$$\langle MATH_{21} \rangle$$

where $N$ represents the total number of entities predicted by the model, $M$ represents the number of correctly predicted entities, and $K$ represents the total number of annotated entities in the dataset. Hyperparameter settings are shown in Table 1 .

### 4.2 Experimental Results and Analysis

Experiments are conducted on the CONLL-2003 English dataset and the BC5CDR specialized medical domain dataset. Comparative results between BiLSTM-BCRF and BiLSTM-CRF are shown in Figures 4 [FIGURE:4] and 5 [FIGURE:5]. The results demonstrate that our model outperforms BiLSTM-CRF on small annotated datasets, making it suitable for low-resource NER tasks. Notably, the model achieves an F1-score of 89.32% on 30% of the CONLL2003 dataset, indicating strong recognition performance without requiring large-scale annotated corpora.

Comparisons with the TMN model are presented in Tables 2 and 3 . On 20% of CONLL2003, BiLSTM-BCRF achieves higher F1-score than TMN. On 20% of BC5CDR, BiLSTM-BCRF outperforms BiLSTM-CRF by 3.35% in F1-score, though it is 0.7% lower than TMN, primarily due to slightly inferior performance on proper nouns. However, BiLSTM-BCRF does not require annotated trigger words, reducing manual annotation costs to only 3/4 of TMN' s requirements, while achieving 2.05% higher F1-score than TMN on CONLL2003. Improving proper noun recognition remains a focus of our future work.

### 4.3 Analysis of Word and Multi-layer Character Information Fusion

Section 3.2 proposed the input layer structure shown in Figure 2. To verify which factors in the input layer affect NER performance, we conduct the following experiments:

**Impact of concatenation order:** Table 4 investigates the effect of word vector and character vector concatenation order on model performance using the BiLSTM-BCRF model on 20% of BC5CDR. Here, "word" denotes word embedding information with dimension 100, "char" denotes character vectors with dimension 50, "char*2" represents concatenation of two character vector matrices, and "+" indicates concatenation. Results show that reversing the concatenation order of one word vector and one character matrix slightly decreases F1-score. When concatenating word vectors with two character matrices, placing the word vector between the two character matrices reduces precision, recall, and F1-score. This demonstrates that concatenation order significantly impacts model performance.

**Impact of character matrix quantity:** Table 5 explores how the number of concatenated character vector matrices affects performance. In low-resource scenarios, concatenating character vectors after word vectors improves the model' s ability to handle rare words and enhances recognition performance. Experiments reveal that with word vector dimension set to 100, concatenating two character vector matrices yields optimal results. Further experiments with word vector dimensions of 50, 200, and 300 confirm that setting word vector dimension to 100 with two character vector matrices provides the greatest performance improvement.

### 4.4 Ablation Study

To investigate the impact of Bernoulli distribution and multi-layer character information, we use BiLSTM-CRF as the baseline. First, we integrate Bernoulli distribution into the baseline' s loss function to create BiLSTM-BCRF(1). Then we fuse multi-layer character information into BiLSTM-CRF(1) to construct the full BiLSTM-BCRF model. Results on both datasets are shown in Table 6 .

The ablation results demonstrate that BiLSTM-BCRF(1) outperforms BiLSTM-CRF on both datasets, confirming the effectiveness of the novel loss function. The final BiLSTM-BCRF model achieves superior precision, recall, and F1-score, validating its strong recognition capability.

### 4.5 Qualitative Analysis

To better illustrate differences between BiLSTM-BCRF and BiLSTM-CRF, we select two example sentences from the datasets: "Only France and Britain backed Fischler' s proposal." and "Rare Hendrix song draft sells for almost $17000" . Annotation results from manual labeling, BiLSTM-CRF, and BiLSTM-BCRF are shown in Tables 7 and 8 .

Table 7 shows a sentence with three entities, where the original BiLSTM-CRF model identifies only two, while our BiLSTM-BCRF model correctly identifies all three. Table 8 contains a single person entity, but BiLSTM-CRF incorrectly labels the first two words as a person entity, whereas our model accurately identifies the true person entity.

## 5 Conclusion

To address the challenge of limited annotated corpora in low-resource domains, we propose the BiLSTM-BCRF model for low-resource NER. By integrating Bernoulli distribution into the loss function, the model achieves better parameter fitting in low-resource scenarios. Fusing multi-layer character feature information enhances the model' s ability to handle rare words, enabling strong recognition performance even with limited annotated data. However, proper noun recognition remains an area for improvement. Future work will focus on enhancing proper noun recognition capabilities, as well as improving knowledge transfer and cross-domain performance.

We thank the reviewers for their constructive feedback. This work was supported by the National Natural Science Foundation of China (NO.61877031).

### References

[1] Hammerton J. Named entity recognition with long short-term memory[C]//Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003. 2003: 172-175.
[2] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for

named entity recognition[J]. arXiv preprint arXiv:1603.01360, 2016.
[3] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.
[4] Yin Z, Li X, Huang D, Li J. Research on Chinese named entity recognition integrating character and word models[J]. Journal of Chinese Information Processing, 2019, 33(11): 95-100+106.
[5] Lin G, Zhang S, Lin H. Research on named entity recognition based on fine-grained word representations[J]. Journal of Chinese Information Processing, 2018, 32(11): 62-71+78.
[6] Rathaparkhi A. A Maximum Entropy Part of Speech Tagger[J]. Proceedings of EMNLP' 96, 1996.
[7] McCallum A, Freitag D, Pereira F C N. Maximum entropy Markov models for information extraction and segmentation[C]//Icml. 2000, 17(2000): 591-598.
[8] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J].
[9] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12(ARTICLE): 2493−2537.
[10] Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.
[11] Ni J, Dinu G, Florian R. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection[J]. arXiv preprint arXiv:1707.02483, 2017.
[12] Mayhew S, Tsai C T, Roth D. Cheap translation for cross-lingual named entity recognition[C]//Proceedings of the 2017 conference on empirical methods in natural language processing. 2017: 2536-2545.
[13] Toolkit O S, BCK Học, MHN Ngữ, et al. Báo cáo khoa học: "Moses: Open Source Toolkit for Statistical Machine Translation" [J]. tailieu vn.
[14] Chen X, Awadallah A H, Hassan H, et al. Multi-source cross-lingual model transfer: Learning what to share[J]. arXiv preprint arXiv:1810.03552, 2018.
[15] Keung P, Lu Y, Bhardwaj V. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER[J]. arXiv preprint arXiv:1909.00153, 2019.
[16] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.
[17] Dai X, Adel H. An analysis of simple data augmentation for named entity recognition[J]. arXiv preprint arXiv:2010.11683, 2020.
[18] Chen J, Wang Z, Tian R, et al. Local additivity based data augmentation for semi-supervised NER[J]. arXiv preprint arXiv:2010.01677, 2020.
[19] Yang Y, Chen W, Li Z, et al. Distantly supervised NER with partial annotation learning and reinforcement learning[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 2159-2169.
[20] Tsuboi Y, Kashima H, Mori S, et al. Training conditional random fields using incomplete annotations[C]//Proceedings of the 22nd International

Conference on Computational Linguistics (Coling 2008). 2008: 897-904.
[21] Feng J, Huang M, Zhao L, et al. Reinforcement learning for relation classification from noisy data[C]//Proceedings of the aaai conference on artificial intelligence. 2018, 32(1).
[22] Zhang T, Xia C, Philip S Y, et al. PDALN: Progressive Domain Adaptation over a Pre-trained Model for Low-Resource Cross-Domain Named Entity Recognition[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 5441-
[23] Chen S, Pei Y, Ke Z, et al. Low-Resource Named Entity Recognition via the Pre-Training Model[J]. Symmetry, 2021, 13(5): 786.
[24] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.
[25] Jie Z, Xie P, Lu W, et al. Better modeling of incomplete annotations for named entity recognition[C]//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). 2019: 729-734.
[26] Sang E F, De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition[J]. arXiv preprint cs/0306050, 2003.
[27] Li J, Sun Y, Johnson R J, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction[J]. Database, 2016,
[28] Lin B Y, Lee D H, Shen M, et al. TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition[C]//arXiv. arXiv, 2020.

### Author Contributions

Zhong Maosheng: Conceptualized the model, wrote and revised the manuscript.
Wu Jiahua: Conducted experiments and revised the manuscript.
All authors declare no competing interests.

### Data Availability

Supporting data is self-archived by the authors and available at 202041600071@jxnu.edu.cn