A Multi-Task Biomedical Named Entity Recognition Method Based on Data Augmentation

Hui Zhao[1], **Di Zhao**[1,2,3]*, **Jiana Meng**[1], **Shuang Liu**[1], **Hongfei Lin**[2]
[1]School of Computer Science and Engineering, Dalian Minzu University, Liaoning, China
[2]School of Computer Science and Technology, Dalian University of Technology, Liaoning, China
[3]Postdoctoral workstation of Dalian Yongjia Electronic Technology Co., Ltd, Liaoning, China
zhaohui0817@163.com, zhaodi@dlnu.edu.cn, mengjn@dlnu.edu.cn
liushuang@dlnu.edu.cn, hflin@dlut.edu.cn

Abstract

The rapid development of artificial intelligence has led to an explosion of literature in the biomedical field, and Biomedical Named Entity Recognition (BioNER) can quickly and accurately identify key information from unstructured text. This task has become an important topic to promote the rapid development of intelligence in the biomedical field. However, in the Named Entity Recognition (NER) of the biomedical field, there are always some problems of unclear boundary recognition, the underutilization of hierarchical information in sentences and the scarcity of training data resources. Based on this, this paper proposes a multi-task BioNER model based on data augmentation, using four data augmentation methods: Mention Replacement (MR), Label-wise token Replacement (LwTR), Shuffle Within Segments (SiS) and Synonym Replacement (SR) to increase the training data. The syntactic information is extracted by incorporating the input sentence into the Graph Convolutional Network (GCN), and then the tag information encoded by BERT is interacted through a co-attention mechanism to obtain an interaction matrix. Subsequently, NER is performed through boundary detection tasks and span classification tasks. Comparative experiments with other methods are conducted on the BC5CDR and JNLPBA datasets, as well as the CCKS2017 dataset. The experimental results demonstrate the effectiveness of the model proposed in this paper.

1 Introduction

Named Entity Recognition (NER) aims to identify named entities of specific meaning from the text and classify them into pre-defined entity types. In specific areas, such as biomedicine, NER needs to identify entity types, such as chemicals, diseases, genes, and proteins. NER plays a major role in critical information extraction and provides an important basis for other downstream tasks.

In deep learning methods, the forms of biomedical entities is more complex, usually composed of multiple words, and there are many abbreviations, whose forms are changeable and lack of clear rules to follow, which makes it difficult to determine the physical boundaries of entities and increases the difficulty of identification. Moreover, well-annotated datasets in the medical field are relatively scarce, further reducing the identification effect. Biomedical Named Entity Recognition (BioNER) also faces nested overlapping entities, leading to further increased difficulty in identification. Limited use of syntactic information in corporcorpus in BioNER methods, which also leads to the problem of poor results. Based on the above challenges, this paper presents a multi-task BioNER model based on data augmentation, with the following main contributions:

(1)The training data set is expanded by using Mention Replacement (MR), Label-wise token Replacement (LwTR), Shuffle Within Segments (SiS) and Synonym Replacement (SR), so as to solve the problem of scarce training data in the medical field and enhance the robustness of the model.

(2)The BioNER task is transformed from the traditional sequence annotation task to multi-task learning, including boundary detection and span classification joint task, thus solving the problems of unclear boundary recognition and entity nesting.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1075-1086, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

(3)The data set is used for syntactic analysis, and the Graph Convolutional Network (GCN) is used to encode the syntactic analysis graph and integrate it into the multi-task learning, so that the model can fully learn the syntactic components in the text and the relationship between the components, so as to improve the performance of NER.

2 Related Work

NER is one of the important indicators to evaluate the effect of information extraction. This paper introduces a data augmentation-based multi-task BioNER model. Next, data augmentation, syntactic analysis, and multi-task methods will be detailed to get a more comprehensive understanding of the study.

2.1 Data Augmentation Methods

Data augmentation is a method of increasing the amount of data based on different modification strategies, generating more and more diverse data sets by transforming, expanding or changing the original data. When training machine learning models, data augmentation acts as a regularization factor to avoid overfitting and increase the generalization of the model. Data augmentation can generate new training samples to simulate the diversity found in the real world, which helps the model learn more robust feature representations, thereby improving its performance on unseen data. Moreover, among the four data augmentation methods, the Shuffle within Segments method can generate adversarial examples that assist the model in better learning sentence features.

Data augmentation methods have been shown to be very effective in areas such as natural language processing. Wei et al.(Jason Wei and Kai Zou, 2019) enhanced the data by replacing the word with one of the synonyms retrieved from the English lexus. Kobayashi et al.(Sosuke Kobayashi, 2018) proposed a new method of labeled sentence data augmentation, called context augmentation. Wang et al.(Xinyi Wang et al., 2018) brought an extremely simple data augmentation strategy for neural machine translation (NMT): randomly replacing the words in the source and target sentences with other random words in the corresponding vocabulary, named this method as SwitchOut. Gao et al.(Fei Gao et al., 2019) proposed a novel neural machine translation data augmentation method to gently expands randomly selected words in sentences by mixing the context of multiple related words. Min et al.(Junghyun Min et al., 2020) explored several methods by applying syntactic transformations to sentences in the natural language inference (NLI) corpus to enhance the standard training set, increasing accuracy from 0.28 to 0.73. Xia et al.(Mengzhou Xia et al., 2019) used a bilingual dictionary and an unsupervised machine translation model to convert data from high resource languages to low resource languages to extend the machine translation training set of low resource languages.

2.2 Syntactic Analysis

In recent years, syntactic analysis is a task of Natural Language Processing (NLP), and has been successfully applied in many fields. In order to encode the context information of each word in a sentence, Kiperwasser(Kiperwasser E and Goldberg Y, 2016) introduces the Bidirectional Long Short-Term Memory (BiLSTM) into the syntactic analysis system based on transfer model and graph model, respectively. Wei et al.(Wei Wei et al., 2024) proposed a text similarity calculation method based on syntactic analysis, which summarizes the principles of other methods. Ding et al.(Ding Meirong et al., 2024) proposed to build a syntactic dependency tree of text first, introduce richer feature information to the model, and use the feature aggregation module to process the feature aggregation, which improved the accuracy of the model's prediction of word emotion analysis.

2.3 Multi-Task Learning Method

Multi-Task Learning (MTL) is a machine learning paradigm that aims to improve the performance of models by learning multiple related tasks simultaneously. Existing studies have begun to explore the association of NER tasks with other tasks and propose methods based on multi-task learning. Collobert et al.(Collobert R et al., 2011) trained a Window / Sentence method network to jointly perform NER, POS, Chunk, and SRL tasks. Rei(Rei M, 2017) found that the performance of sequence annotation model can be improved by adding unsupervised language modeling objectives to the training process. Zhang(Zhang Xue et al., 2024) et al. proposed a multi-task classification algorithm based on loss optimization of gradient amplitude direction adjustment to transform multiple classification tasks into multiple dichotomy tasks. Song et al.(Song Donghuan et al., 2024) proposed to construct a novel classification model, which demonstrated the effectiveness of the proposed model by designing multi-task experiments and transfer learning experiments. Wang et al.(Wang Jingwei et al., 2024) proposed a multi-task Transformer model for flight state parameter regression and flight state classification, and the experimental results proved the superiority of the proposed model.

3 Method

First, the model uses four methods for data augmentation, with input data and label information processed through BERT encoding, and syntactic information extracted using GCN. New input data and label information, along with syntactic information via an attention interaction matrix, are then applied to the boundary detection and span classification tasks for entity recognition. The BioNER task is transformed from a sequence annotation task into a multi-task framework, specifically addressing the joint tasks of boundary detection and span classification, which solves the problems of unclear boundary recognition and entity nesting. Syntactic analysis is performed on the dataset, and GCN is used to encode the syntactic analysis graph, integrating it into the multi-task framework. This allows the model to fully learn the syntactic components in the text and their interrelationships, thereby improving the performance of NER. The overall architecture of the model is shown in Figure 1.

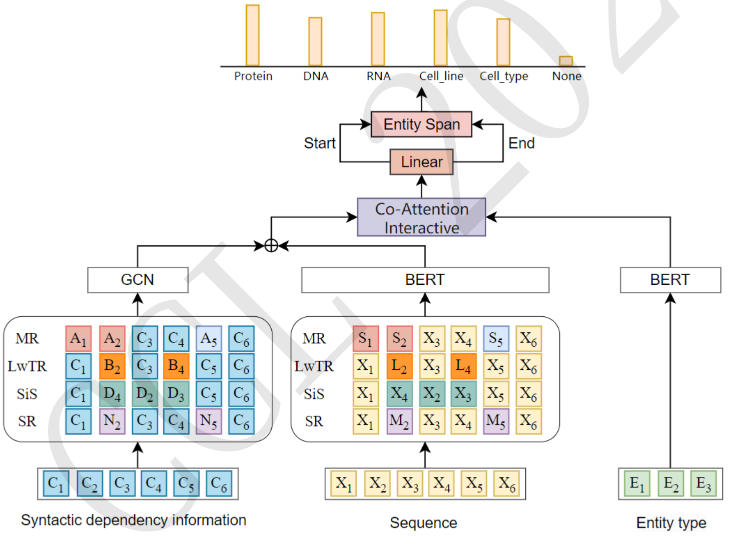

Figure 1: The overall architecture of the model

3.1 Data Augmentation Methods

In this paper, four methods of MR, LwTR, SiS and SR are used to change the input data, and these four methods are used for data augmentatio.

Mention Replacement (MR):For each entity in the input sentence, a binomial distribution is used to randomly decide whether it should be replaced. The binomial distribution represents the resulting probability distribution of a series of independent repeated binary experiments. For example, in Figure 2, the chemical type entity ”22-oxacalcitriol [B-chem I-chem I-chem]” was replaced by ”Puromycin aminonucleoside [B-chem I-chem]”, an entity of the disease type ”secondary hyperparathyroidism [B-dise I-disorder]” was replaced by ”Cauda equina syndrome [B-dise I-dise I-disorder]”, ”low bone turnover [B-dise I-dise I-disease]” was replaced by ”neurotoxicity [B-disease]”, and ”renal failure [B-dise I-disorder]” was replaced by ”neurological deficits [B-dise I-disorder]”. The new training data, after entity replacement, along with the original training data, are encoded by the pre-trained model BERT, and then decoded by the CRF to obtain the final label.

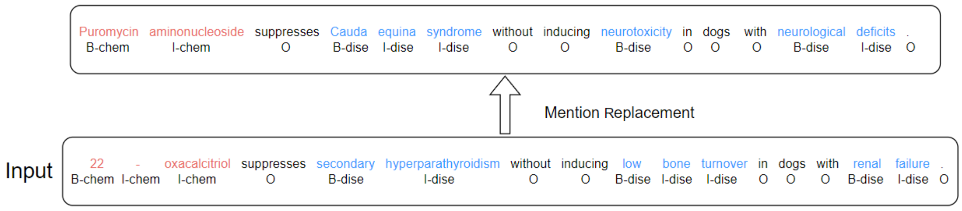
Puromycin aminonucleoside suppresses Cauda equina syndrome without inducing neurotoxicity in dogs with neurological deficits .
B-chem I-chem O B-dise I-dise I-dise O O B-dise O O O B-dise I-dise O
Mention Replacement
Input
22 - oxacalcitriol suppresses secondary hyperparathyroidism without inducing low bone turnover in dogs with renal failure .
B-chem I-chem I-chem O B-dise I-dise O O B-dise I-dise I-dise O O O B-dise I-dise O

Figure 2: The Mention Replacement method

Label-wise Token Replacement (LwTR): For each word, a binomial distribution is also used to randomly decide whether it should be replaced. If the probability result is replacement, use the tag-word distribution (the label-token) distribution built from the original training set to randomly select another word with the same tag and ultimately keep the original tag sequence unchanged. As shown in Figure 3, six words (suppresses, without, inducing, bone, dogs, and renal) were replaced by other words (aminonucleoside, levels, of, none, haven, hypercalcemia) with the same label as the original word.

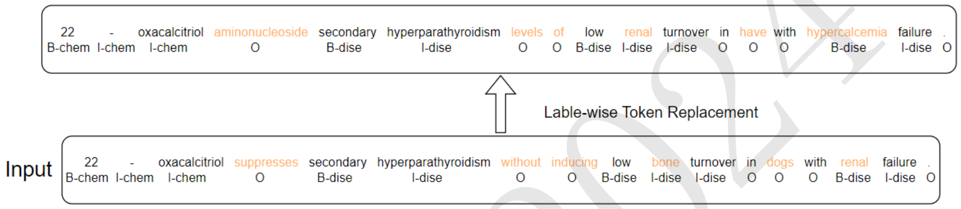
22 - oxacalcitriol aminonucleoside secondary hyperparathyroidism levels of low renal turnover in have with hypercalcemia failure .
B-chem I-chem I-chem O B-dise I-dise O O B-dise I-dise I-dise O O O B-dise I-dise O
Lable-wise Token Replacement
Input
22 - oxacalcitriol suppresses secondary hyperparathyroidism without inducing low bone turnover in dogs with renal failure .
B-chem I-chem I-chem O B-dise I-dise O O B-dise I-dise I-dise O O O B-dise I-dise O

Figure 3: The Label-wise Token Replacement method

Shuffle Within Segments (SiS): First, the word sequence is split into small parts of the same label, where each part corresponds to a word or a series of words of the same type. As shown in Figure 4,secondary hyperparathyroidism, without causing and low bone transformation all three phrases were disrupted, (secondary hyperparathyroidism) were disrupted as (hyperparathyroidism secondary), (without inducing) as (inducing without), and (low bone turnover) as (turnover low bone).

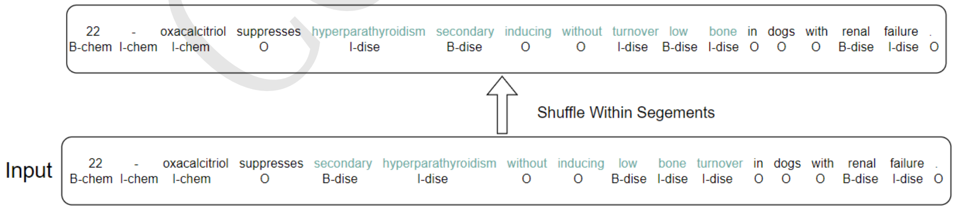
22 - oxacalcitriol suppresses hyperparathyroidism secondary inducing without turnover low bone in dogs with renal failure .
B-chem I-chem I-chem O I-dise B-dise O O I-dise B-dise I-dise O O O B-dise I-dise O
Shuffle Within Segements
Input
22 - oxacalcitriol suppresses secondary hyperparathyroidism without inducing low bone turnover in dogs with renal failure .
B-chem I-chem I-chem O B-dise I-dise O O B-dise I-dise I-dise O O O B-dise I-dise O

Figure 4: The Shuffle within Segments method

Synonym Replacement (SR): The SR method is similar to LwTR, replacing words with synonyms retrieved using WordNet or Chinese WordNet. But the retrieved synonyms may include more than one word, so for BIO labels, a simple rule can be derived: If the replaced word is the first word in the entity (i.e., the corresponding label is ”B-chem”), we assigned the same label to the first marker of the retrieved multi-word synonyms and assigned ”I-chem” to the other words. If the replaced word is in an entity (i.e., the corresponding label is ”I-chem”), we assign its label to all words of multi-word synonyms. As shown in Figure 5, (suppresses) replaced with (restrain), (low bone turnover) replaced with (suppression of bone turnover), and (renal failure) replaced with (impaired renal failure).

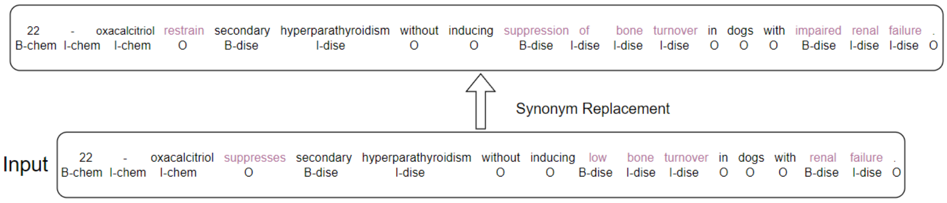
22 - oxacalcitriol restrain secondary hyperparathyroidism without inducing suppression of bone turnover in dogs with impaired renal failure .
B-chem I-chem I-chem O B-dise I-dise O O B-dise I-dise I-dise I-dise O O O B-dise I-dise I-dise O
Synonym Replacement
Input
22 - oxacalcitriol suppresses secondary hyperparathyroidism without inducing low bone turnover in dogs with renal failure .
B-chem I-chem I-chem O B-dise I-dise O O B-dise I-dise I-dise O O O B-dise I-dise O

Figure 5: The Synonym Replacement method

The four methods shown above are used to enhance the input training data. These four methods use the binomial distribution to randomly replace and change. For the binomial distribution, the probability of success in each experiment is p, and the probability of failure is $q = 1 - p$, the probability quality function of the binomial distribution of n experiments is shown as eq 1:

$$P(x = k) = C(n, k) \times p^k \times (1 - p)^{n-k} \tag{1}$$

$P(x = k)$ is the probability of a successful event k, $C(n, k)$ is the number of combinations, indicating the combination of k times in n experiments, p is the probability of success in each experiment, $q(q = 1 - p)$ is the probability of failure in each experiment, n is the total number of independent experiments, k is the number of successful events. A binomial distribution was used each time to calculate whether substitutions were needed.

3.2 Encoding Layer

Syntactic Dependency Tree: Syntactic Dependency Tree is a type of syntactic structure tree that illustrates the grammatical relationships between each word in a sentence and how they are connected within the sentence. There is typically a root node that represents the core of the entire sentence, with other words connected to the root node or to each other through various types of dependency relationships.

GCN: GCN is a deep learning model used to process graph data. In NLP, text data is modeled into a graph where each word is represented by a node in the graph, and the relationships between words, such as dependencies or co-occurrences, are represented as edges in the graph. The context of a named entity typically contains key information, and the GCN can model each node by considering the neighboring nodes and can capture the dependencies between words. In this paper, the GCN is used to extract syntactic information from the input sentences.

The GCN makes the input information and its dependent syntactic tree interact, so that the model fully learns the existing dependent relationship in the sentence. The $h1$ is the input information encoded by BERT, A is the adjacency matrix, D is the degree matrix, and w and b are the learnable parameters. The GCN extracts the syntactic features in the sentence as shown in formula eq 2:

$$G_x = RELU(DAh_1w + b) \tag{2}$$

For the training data augmented with the four data augmentation methods, BERT is used for encoding. In addition to the syntactic information in the sentences, entity-specific information is also encoded within the entity types. To better represent the entity type label information, average pooling is conducted on the BERT-encoded entity type labels, resulting in rich representations of these labels.

3.3 Fusion Layer

The co-attention mechanism is an attention method that processes sequence data and establishes inter-sequence associations. It calculates the weighted attention for each element, making the information transmission of long sequences more efficient. This mechanism enables the model to consider all elements in the input sequence for both encoding and decoding, thereby enabling the model to better understand the long-distance dependencies between sequences. In this paper, we use the co-attention mechanism to learn syntactic information in sentences for improved feature fusion.

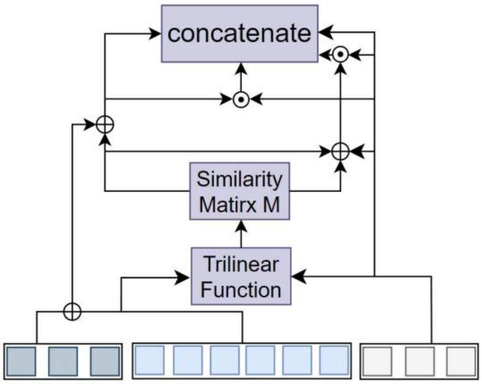

Figure 6: Co-Attention Mechanism Model

In order to better perform feature fusion, the common attention interaction mechanism is used to fuse features. As shown in Figure 6, in the output of the GCN and label after mean pooling output, using the common attention interaction network, make the syntactic information contained in the input sentences and sentences and the entity label contains specific information fully interaction, makes the model can better learn the potential interaction information and specific information, which can better represent the entity span boundary, boundary information as shown in eq 3:

$$M = w^T[h_x, G_x, h_e] \tag{3}$$

Where w is the trainable weight, M is the final attention weight matrix. By performing row-wise and column-wise operations, we respectively obtain the interaction matrices Mr for entity boundaries and Mc for entities. By using eqs 4 and 5, we obtain the interaction between labels and tokens, and tokens and labels.

$$H_l^t = M_r \bullet h_e \tag{4}$$

$$H_t^l = M_r \bullet M_c^T \bullet h_x \tag{5}$$

3.4 Classification Layer

The multi-task NER model allows for sharing and utilizing data across multiple tasks. The method adopted in this paper, which employs boundary detection and span classification, is capable of identifying nested entities. The boundary detection component ascertains the specific boundaries of nested entities. Multi-task learning reduces the complexity of maintaining and debugging individual models, leading to decreased computational resource requirements and improved system efficiency. The specific method is shown in Figure 7.

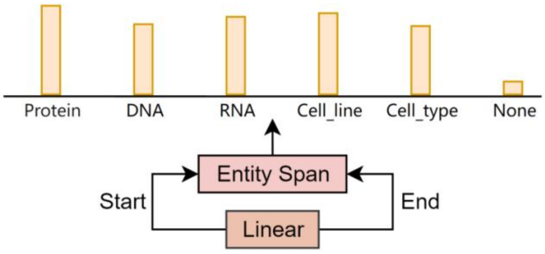

Figure 7: Multi-Task Learning Model

The head and tail positions of entities are identified first, and then the label classification task of entity type is performed. Head and tail entity recognition tasks require the model to perform more fine-grained identification near the boundaries of entities, thus providing more contextual information and being more helpful to accurately determine entity boundaries. The entity classification task identifies entity types, which provides additional information about the entity content. This multitasking method allows the model to understand entities from different perspectives and improves the accuracy of entity identification; the model can better determine the location of the head and tail entities, and the information of the head and tail entities can help each other to improve performance and overall performance.

Specifically, calculating the start or end position of the entity span, as the front task of multi-task mode, calculating the correct entity boundaries, laying the foundation for the following entity type classification task. During training, the cross-entropy loss was used to optimize the boundary of the entity span. The formula is as eqs(6;7;8):

$$B = [h_x; H_l^t h_x \bullet H_l^t; h_x \bullet H_t^l] \tag{6}$$

$$p(\theta)_i = \frac{exp(w_1^T B_i)}{\sum_j exp(w_1^T B_j)} \tag{7}$$

$$L_{ht}(\theta) = -\sum_{i=1}^{N} \hat{y}_i log P(\theta)_i \tag{8}$$

The span classification task involves matching the head and tail labels of each boundary tag to classify entity spans into corresponding semantic labels. If a candidate entity span is not an entity, it is labeled as the ”None” category. An entity span V_{span} is defined as eq 9:

$$V_{span} = [B_i + B_j; B_i - B_j] \tag{9}$$

Finally, the entity span representation is sent to a softmax layer to predict the probability P_{span} of entity labels. The entity label prediction loss L_{span} is minimized through eqs 10 and 11.

$$P_{span}(\theta) = softmax(MLP(w^T V_{span} + b)) \tag{10}$$

$$L_{span}(\theta) = -\sum_{i=1}^{c}(y_{span}^{\Delta i}) log(P_{span}^i(\theta)) \tag{11}$$

In the multi-task approach, during the training phase, the loss of entity type labels and boundaries is minimized to the maximum extent possible using eq 12.

$$L(\theta) = L_{ht}(\theta) + L_{span}(\theta) \tag{12}$$

4 Experiments and Analysis

4.1 Experimental Data

The statistics of the datasets are shown in Table 1:

	JNLPBA	BC5CDR	CCKS2017
Train	14,690	15,935	1,561
Dev	3,856	4,012	334
Test	3,856	4,012	334

Table 1: Dataset Statistics

4.2 Experimental Results and Analysis

Comparative Experiments of Data Augmentation: In order to verify the effectiveness of data augmentation methods, this study conducted experiments on the BC5CDR dataset, the JNLPBA dataset and the CCKS2017 dataset using single data augmentation methods as well as the combination of all four methods under the sequence labeling method. Comparative experiments were also conducted with other methods, and the experimental results are presented in Table 2 :

	BC5CDR			JNLPBA			CCKS2017		
Model	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
BERT+CRF	87.0	81.5	84.2	71.8	77.3	74.5	88.7	93.0	90.9
BERT+BiLSTM+CRF	87.7	82.0	84.8	71.0	79.6	75.0	90.9	91.6	91.2
BERT+CNN+CRF	88.7	82.6	85.5	71.2	80.2	75.4	91.3	92.2	91.7
BERT+CRF+MR+GCN	89.2	87.8	88.5	74.6	79.8	77.2	92.2	92.8	92.5
BERT+CRF+LwTR+GCN	88.6	**89.2**	88.9	74.7	80.3	77.5	92.5	92.9	92.7
BERT+CRF+SiS+GCN	89.5	88.9	89.1	77.8	78.0	77.9	92.7	93.7	93.2
BERT+CRF+SR+GCN	89.1	88.8	89.0	74.2	**81.0**	77.6	92.4	93.4	92.9
BERT+CRF+ALL+GCN	**90.2**	88.7	**89.4**	**78.5**	77.9	**78.2**	**93.2**	**93.8**	**93.5**

Table 2: Baseline Comparison Experiments Results

According to the experimental results shown in Table 2, the model with the added syntactic information exhibits a significant improvement in F1 score, whether using single data augmentation method or all four data augmentation methods combined. Among the four methods of MR, LwTR, SiS and SR, it can be observed that in the experimental results of the three datasets, SiS method performs the best, followed by SR, then LwTR, and finally MR. The reason why the SiS method performs the best is because it increases diversity. Shuffling disrupts the original order of the text, enhancing the diversity of the dataset, which helps the model to learn the associations between words and phrases in different contexts more comprehensively, thus improving its generalization ability and robustness. The advantage of the SR method over other methods lies in its ability to alleviate the problem of data sparsity. Some words may appear less frequently in the dataset, and synonym replacement can increase the number of samples for these words. The relatively poor performance of the LwTR method and the MR method is due to the fact that the replaced words introduce new meanings for the model to learn. However, for the model, the relationships between each word in the context and the sentence are relatively more important. After incorporating syntactic information and data augmentation, the F1 scores of the BC5CDR dataset and the JNLPBA dataset improved by 3.9% and 2.8% respectively compared to the baseline experimental method. On the CCKS2017 dataset, the F1 score improved by 1.8%. After incorporating data augmentation and syntactic information, the model achieved its best performance. The experimental results demonstrate that combining various data augmentation techniques and syntactic information can significantly improve the performance of NER models.

Comparative Experiments of Multi-Task Learning: Experiments were conducted on the BC5CDR dataset, the JNLPBA dataset and the CCKS2017 dataset using the multi-task method, and comparative experiments were conducted with other methods. The experimental results are shown in Tables 3 and 4:

	BC5CDR				JNLPBA		
Model	P(%)	R(%)	F1(%)	Model	P(%)	R(%)	F1(%)
CASN(2021)	89.5	90.9	90.2	CASN(2021)	75.8	80.0	78.0
BERTForTC (2022)	-	-	90.8	SciFive(2021)	-	-	77.5
UniNER (2023)	-	-	89.3	MINER (2022)	-	-	77.0
GoLLIE (2023)	-	-	88.4	BioDistilBERT(2023)	73.5	**85.5**	79.1
Ours	**92.7**	**91.5**	**92.1**	Ours	**78.9**	82.1	**80.5**

Table 3: Comparative Experimental Results on BC5CDR and JNLPBA

Model	CCKS2017		
	P(%)	R(%)	F1(%)
Yu et al. (2022)	-	-	93.8
Lin et al. (2023)	92.6	91.9	92.2
Li et al. (2023)	87.4	92.1	89.7
Shen et al. (2023)	93.3	92.5	92.6
Ours	**94.1**	**96.3**	**95.2**

Table 4: Comparative Experimental Results on CCKS2017

As shown in Tables 3 and 4, after comparative experiments with other methods, the proposed multi-task learning model for BioNER based on data augmentation outperforms other methods. It achieved the best F1 scores on both the BC5CDR dataset and the JNLPBA dataset, with improvements of 1.3% and 1.4% respectively. Similarly, it also achieved the best F1 score on the CCKS2017 dataset, with an improvement of 1.4%. This demonstrates the effectiveness of data augmentation, incorporation of syntactic information, and feature extraction using GCN. The model has successfully learned more dependencies existing in sentences, and its performance in boundary detection and span classification multi-task learning is excellent.

Ablation Experiments: In order to verify the effectiveness of the multi-task module, syntactic information, and data augmentation methods on the model's performance, we conducted ablation experiments on the BC5CDR dataset, the JNLPBA dataset and the CCKS2017 dataset. The experimental results are shown in Table 5:

Model	BC5CDR			JNLPBA			CCKS2017		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Ours	**92.7**	91.5	**92.1**	**78.9**	**82.1**	**80.5**	**94.1**	**96.3**	**95.2**
W/O MT	90.2	88.7	89.4	78.5	77.9	78.2	93.2	93.8	93.5
W/O GCN	91.3	90.1	90.7	78.4	79.4	78.9	93.5	94.3	93.9
W/O DA	90.2	**92.4**	91.3	78.0	81.7	79.8	93.9	95.6	94.7

Table 5: Comparative Results of Ablation Experiments

As shown in Table 5, it is evident that removing either the multi-task module, syntactic information, or data augmentation leads to a decrease in the model's performance. After removing the multi-task module, syntactic information, and data augmentation, the F1 scores for the BC5CDR dataset decreased by 2.7%, 1.4%, and 0.8% respectively. For the JNLPBA dataset, the F1 scores decreased by 2.3%, 1.6%, and 0.7% respectively. Similarly, for the CCKS2017 dataset, the F1 scores decreased by 1.7%, 1.3%, and 0.5% respectively. The experimental results demonstrate that the multi-task module has the greatest impact on the model, followed by syntactic information, and finally data augmentation methods. Using the multi-task learning approach with entity boundary detection and span classification tasks enables the recognition of all entities and effectively addresses the out-of-vocabulary problem. This is crucial for NER tasks as it ensures comprehensive entity recognition. The syntactic relationships provided by syntactic information serve as shallow features and have a beneficial effect on NER. Data augmentation methods enhance the robustness of the model, albeit to a lesser extent compared to other factors. The experimental results provide ample evidence of the positive impact and effectiveness of the multi-task learning approach, syntactic information, and data augmentation methods on the model.

4.3 Analysis of Errors

On the BC5CDR dataset, the CASN method is a BioNER model based on a multi-task approach. BioLinkBERT utilizes inter-document connections for self-supervised training. BERTForTC is a BERT-based model specifically designed for text classification tasks. UniNER leverages the large model GPT to enhance the performance of the NER model. GoLLIE utilizes a large model guide for information extraction to improve model effectiveness, showing significant results in NER.

On the JNLPBA dataset, SciFive is a domain-specific T5 model that has been pre-trained on extensive biomedical corpora and has delivered superior performance in the context of NER models. MINER proposes a new learning framework for NER, approaching the challenge from an information-theoretic viewpoint and addressing the issue of out-of-vocabulary (OOV) entities. BINDER improves the efficiency of NER by applying contrastive learning to project candidate text spans and entity types into a shared vector representation space. The BioDistilBERT method, based on the DistilBERT model, has been fine-tuned on biomedical text and has yielded excellent outcomes.

On the CCKS2017 dataset, Yu et al. proposed a multi-task learning approach for entity recognition, which recognizes entities through the interaction of two tasks: entity recognition and participle training. The two tasks are not both NER in this method, while the boundary recognition and entity type classification proposed in this paper are both related to NER, and the experimental results also proved that the proposed method is better. Duan et al. tested the micro-F1 values and the changes of the F1 values of each entity after augmenting the word embedding with Chinese character features, and experimented the effect of different methods on the CCKS2017 dataset. Lin et al. proposed a pre-trained model multi-layer dynamic fusion method. The multi-layer dynamic fusion method can fully utilize the knowledge of each layer of the pre-trained model, so that the results contain rich syntactic, semantic and other feature information, improve the representation ability of the model in the task, and enhance the flexibility of the model. Li et al. proposed a novel entity recognition method combined with CNN. The method introduces an external vocabulary through SoftLexicon and combines character and vocabulary vectors to enhance character feature representation using word frequency normalization of associated characters. The experimental results show that this method outperforms SoftLexicon on multiple datasets, effectively improving the performance of NER. Shen et al. proposed a model that combines whole-word masking and BERT pre-processing, extracting features through BiLSTM and attention mechanisms, and using CRF decoding to enhance the recognition effect of Chinese medical text. The experimental results on different datasets show that the model achieved high F1 scores in the task of BioNER, proving its effectiveness.

These methods have overlooked the shallow-level features inherent in sentences and also did not have an adequate amount of training data for thorough training. In this paper, the model has learned the shallow-level features within sentences and has been enhanced with data augmentation methods to strengthen the training data. The results have also confirmed the effectiveness of this model.

5 Summary and Outlook

This paper proposes a multi-task BioNER model based on data augmentation. It employs four data augmentation methods, MR, LwTR, SiS and SR, to augment the training data. The model utilizes GCN to extract syntactic features from sentences. These multiple features are then interacted through a co-attention mechanism and combined with boundary detection and span classification in a multi-task mode for NER. The experimental results confirm the effectiveness and positive impact of adopting multi-task learning and data augmentation methods.

In the future, we will explore transfer learning and large language models. Large language models, with more parameters and deeper structures, have the capability to capture richer and more complex data patterns and features. This enables them to learn more abstract and advanced representations, leading to better performance across various tasks.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities (0854-53) and the Natural Science Foundation of Liaoning Province, China (2022-BS-104).

References

Chen, P., Peng, et al.: *Co-Attentive Span Network with Multi-Task Learning for Biomedical Named Entity Recognition.* In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2021, pp. 649-652.

COLLOBERT R,WESTON J,BOTTOU L,et al. *Natural language processing(almost) from scratch* [J]. Journal of Machine Learning Research, 2011,12:2493-2537.

Ding Meirong, Lai Jin Qian, Zeng Biqing, etc. *Level-level sentiment analysis based on local global context guidance* [J]. Software Guide, 2024,23 (01): 190-196.

Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. 2019. *Soft contextual data augmentation for neural machine translation.* In ACL, pages 5539–5544, Florence, Italy.

Jason Wei and Kai Zou. 2019. *EDA: Easy data augmentation techniques for boosting performance on text classi- fication tasks.* In EMNLP-IJCNLP, pages 6382–6388, Hong Kong, China.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. *Syntactic data augmentation increases robustness to inference heuristics.* In ACL, pages 2339–2352, Online.

KIPERWASSER E,GOLDBERG Y. Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations [J].TACL,2016,4:313–327.

Kocaman, V., Talby, D.: *Accurate Clinical and Biomedical Named Entity Recognition at Scale.* Software Impacts 13, 100373 (2022).

Li Mingjian, Li Weijun, Wang Hairong. *Identification of fused association information with CNN* [J]. Journal of Zhengzhou University (Science edition), 2023,55 (05): 53-59.

Lin Lingde, Liu Na, Xu Zhenshun, et al. *Chinese medical named entity identification based on multi-layer dynamic fusion* [J]. Computer Engineering and Application, 2023:1-13.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. *Generalized data augmentation for low-resource translation.* In ACL, pages 5786–5796, Florence, Italy.

Phan L N, Anibal J T, Tran H, et al. *SciFive: a text-to-text transformer model for biomedical literature* [J]. 2021.

REI M. *Semi-supervised multitask learning for sequence labeling* [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017:2121-2130.

Rohanian, O., Nouriborji, M., Kouchaki, S., et al.: *On the Effectiveness of CompactBiomedical Transformers.* Bioinformatics 39 (3), btad103 (2023).

Sainz O, García-Ferrero I, Agerri R, et al. *GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction* [C]//The Twelfth International Conference on Learning Representations. 2023.

Shen Tongping, Yu Lei, Shu Jianhua, etc. *Medical text entities based on the multi-head attention mechanism* [J]. Journal of Anqing Normal University (Natural Science Edition), 2023,29 (01): 85-91.

Song Donghuan, Hu Maodi, Ding Jielan, etc. *Research on cross-type text classification techniques based on multi-task learning* [J/OL]. Data analysis and knowledge discovery: 1-19 [2024-03-20].

Sosuke Kobayashi. 2018. *Contextual augmentation: Data augmentation by words with paradigmatic relations.* In NAACL, pages 452–457, New Orleans, Louisiana.

Wang Jingwei, Gao Yan, Song Xing, etc. *Aircraft state prediction based on the Transformer* [J]. Computer Engineering and Design, 2024,45 (02): 477-483.

Wang, X., Dou, S., Xiong, L., Zou, Y., Zhang, Q., Gui, T., Qiao, L., Cheng, Z., Huang, X.: *MINER: Improving Out-of-Vocabulary Named Entity Recognition from an Information Theoretic Perspective.* In:

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022, vol. 1, pp. 5590-5600.

Wei Wei, lilac incense, Guo Mengxing, etc. *Summary of text similarity calculation methods [J/OL].* Computer Engineering: 1-19 [2024-03-20].

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. *SwitchOut: an efficient data augmentation algorithm for neural machine translation.* In EMNLP, pages 856–861, Brussels, Belgium.

Yu, P., Chen, Y., Xu, J., et al.: *Entity Recognition Method for Electronic Medical Records Based on Multi-task Learning.* In: Computer and Modernization, 2022 (09), 40-50.

Zhang Xue, Tian LAN, Zeng Ming, etc. *An ECG signal multitask classification algorithm based on gradient amplitude direction adjustment* [J/OL]. Computer Science: 1-10 [2024-03-20].

Zhou W, Zhang S, Gu Y, et al. *UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition* [C]//The Twelfth International Conference on Learning Representations. 2023.