Quantitative Biology 2022, 10(3): 253–263
https://doi.org/10.15302/J-QB-022-0302

RESEARCH ARTICLE

# Deep learning-based large-scale named entity recognition for anatomical region of mammalian brain

**Xiaokang Chai[1], Yachao Di[1], Zhao Feng[1,4], Yue Guan[1,*], Guoqing Zhang[3], Anan Li[1,4,5], Qingming Luo[2]**

[1] Britton Chance Center for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics, MoE Key Laboratory for Biomedical Photonics, Huazhong University of Science and Technology, Wuhan 430074, China
[2] Key Laboratory of Biomedical Engineering of Hainan Province, School of Biomedical Engineering, Hainan University, Haikou 570228, China
[3] CAS Key Laboratory of Computational Biology, Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Shanghai 200031, China
[4] Research Unit of Multimodal Cross Scale Neural Signal Detection and Imaging, Chinese Academy of Medical Sciences, HUST-Suzhou Institute for Brainsmatics, JITRI, Suzhou 215123, China
[5] CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031, China
* Correspondence: yguan@hust.edu.cn

Received June 4, 2021; Revised August 26, 2021; Accepted September 22, 2021

***Background*: Images of anatomical regions and neuron type distribution, as well as their related literature are valuable assets for neuroscience research. They are vital evidence and vehicles in discovering new phenomena and knowledge refinement through image and text big data. The knowledge acquired from image data generally echoes with the literature accumulated over the years. The knowledge within the literature can provide a comprehensive context for a deeper understanding of the image data. However, it is quite a challenge to manually identify the related literature and summarize the neuroscience knowledge in the large-scale corpus. Thus, neuroscientists are in dire need of an automated method to extract neuroscience knowledge from large-scale literature.**

***Methods*: A proposed deep learning model named BioBERT-CRF extracts brain region entities from the WhiteText dataset. This model takes advantage of BioBERT and CRF to predict entity labels while training.**

***Results*: The proposed deep learning model demonstrated comparable performance against or even outperforms the previous models on the WhiteText dataset. The BioBERT-CRF model has achieved the best average precision, recall, and F1 score of 81.3%, 84.0%, and 82.6%, respectively. We used the BioBERT-CRF model to predict brain region entities in a large-scale PubMed abstract dataset and used a rule-based method to normalize all brain region entities to three neuroscience dictionaries.**

***Conclusions*: Our work shows that the BioBERT-CRF model can be well-suited for brain region entity extraction. The rankings of different brain region entities by their appearance in the large-scale corpus indicate the anatomical regions that researchers are most concerned about.**

**Keywords:** brain region; entity extraction; literature mining; WhiteText; deep learning

**Author summary:** In this study, the BioBERT-CRF model was used to extract brain region entities from a large-scale PubMed abstract dataset and a normalization pipeline was created for normalizing all the labeled brain region entities extracted to three neuroscience dictionaries. Prior to entity prediction, the performance of the BioBERT-CRF model was evaluated using the WhiteText dataset. Compared to other deep learning models, the BioBERT-CRF model achieved the best average precision, recall, and F1 score of 81.3%, 84.0%, and 82.6%, respectively. Our work demonstrates how the BioBERT-CRF model can be well-suited for neuroscience brain region entity extraction. The rankings of different brain region entities by their appearance in the large-scale corpus reflect the anatomical regions that researchers are most concerned with.

## INTRODUCTION

In neuroscience, researchers are fascinated with spatial anatomy, neural distribution, neuroanatomical connectivity [1], and will often use image data to understand and further explore these fields. However, image data usually has a specific focus in each experiment, such as a certain type of neuron or region of the brain, thus is not entirely comprehensive. In addition, it is very difficult to continuously carry out experimental research at the mesoscopic level across the entire brain. Therefore, more comprehensive literature knowledge is required to accurately design a sophisticated experiment, which will not only reduce image data collection and processing, but also alleviate the problem from the source. However, it is quite a challenge to manually search for neuroscience-related knowledge from large-scale literature sources, creating an urgent need for an automated knowledge extraction method. In recent years, named entity recognition technology has been one of the most discussed topics since it can automatically extract name mentions from literature sources [2]. This technology can be well qualified for the automatic recognition and labeling of brain region entities in large-scale literature.

Named entity recognition is an information extraction technique that has been widely used to extract a variety of entities from literature, such as genes, proteins [3–6], diseases [4,5,7–9], chemicals [4,6,8–10], mutations [11,12], species [13], and cell types [3,5,14] in the biomedical field. Leaman *et al*. proposed the semi-Markov model [10], which relied on lexical features such as Token, Part-of-Speech (POS), and N-grams to extract two types of entities, diseases, or chemicals. Wang *et al*. proposed the MTM-CW model [3], which allowed for combining different types of biomedical training datasets, such as genes, diseases, and chemicals. The single model error was reduced by learning relevant knowledge in the field, thereby improving recall. Giorgi *et al*. proposed a transfer learning method based on Bi-LSTM-CRF [15], which utilized many silver standard datasets for training the model. By fine-tuning the parameters on the gold standard datasets, the model expanded its data and produced better results. Lee *et al*. added abstracts in PubMed and full texts in PMC for the BERT model [16] for pre-training [17]. They also fine-tuned the parameters of multiple types of entity datasets and obtained multiple SOTA's BioBERT models in the biomedical field. However, WhiteText is the only available gold standard dataset in the neuroscience field [18] as far as we know, which can annotate brain region entities. This project was first proposed by French *et al*. in 2009. They invited several neuroscientists to manually annotate 1,377 literature pieces in *Journal of Comparative Neurology* (JCN) and 18,242 brain entities were labeled. Therefore, the WhiteText dataset was used to extract brain region entities in the neuroscience literature. French *et al*. used dictionary-based and CRF-based methods [19] to evaluate the WhiteText dataset and achieved an F1 score of 44% and 79%. Richardet *et al*. [20] added species information and other features to the CRF-based method to reduce false-positive error and achieved 84.6% precision, 78.8% recall, and 81.6% F1 score. Shardlow *et al*. [21] used the Bi-LSTM-CRF deep learning model to obtain the best evaluation result of 81.8% F1 score on the Whitetext dataset.

In this study, we first compared the results of a modified deep learning-based named entity recognition model with other deep learning models on the WhiteText dataset. The modified BioBERT-CRF model achieved an F1 score of 82.6%, the best result among these models. Next, we used the BioBERT-CRF model to extract brain region entities from a large-scale PubMed abstract dataset and normalized all extracted brain region entities to three neuroscience dictionaries to solve the problems of synonyms and different naming systems. Finally, we counted the number of different brain region entities that appeared in the large-scale abstracts and selected the top 30 brain region entities. Our study found a suitable deep learning-based method to extract brain region entities from abstracts, reflecting which anatomical regions researchers are most concerned about.

## RESULTS

We used a modified named entity recognition model, the BioBERT-CRF model, as well as four benchmark named entity recognition models, the Bi-LSTM-CRF, MTM-CW, BERT, and BioBERT, to extract brain region entities from the WhiteText dataset. Then, we used the BioBERT-CRF model to extract brain region entities from a large-scale neuroscience literature dataset and normalized all extracted brain region entities to three neuroscience dictionaries.

The WhiteText dataset is a brain region entity gold standard dataset constructed by French *et al*. in 2009 from 1,377 randomly selected abstracts from JCN journals. Several neuroscientists were invited to manually label the brain region entities on the dataset based on the literature content, neuroscience knowledge, reference brain atlases, and dictionaries. Finally, a total of 18,242 brain region entities were labeled. For the labeled brain region entities, the agreement rate of the annotators was 90.7%. We divided part of the WhiteText dataset into a train set, a validation set, and a test set according to 1,000, 150, and 227 abstracts. We

randomly divided them into train set, validation set, and test set five times.

### Comparison of the accuracy of different deep learning-based models for predicting brain region entities

We compared the results of five different named entity recognition models on the WhiteText dataset. Out of the 1,377 abstracts, we randomly selected 1,000 as the train set, 150 as the validation set, and the remaining 227 as the test set. We then performed five cross-validations using the precision, recall, and F1 score as evaluation indicators to show the model prediction results. Evaluation data can be found in the appendix. The box diagrams of the five models are shown in Fig.1.

In addition, we examined specific sentences with predicted labels to compare the recognition results of different deep learning models, as shown in Table 1. The true labels and predicted labels of each model are underlined. In this case, the Bi-LSMT-CRF and MTM-CW model failed to recognize all brain region entities accurately. These two models all recognized adjective words in front. On the contrary, the BERT, BioBERT, and BioBERT-CRF model recognized all entities accurately.

The first model is Bi-LSTM-CRF. With the exception of the third experiment, the precision, recall, and F1 score of Bi-LSTM-CRF on the validation set were higher than 81% and generally higher than the test set results, as shown in Table 2. Its results in the first and fifth experiments were much higher than the test set. This suggests that the model may have a certain degree of overfitting due to its limited dataset size. The overall F1 score of the model on the test set fluctuated around 81%, indicating that the model is robust.

The following model is MTM-CW. The results of this model in Table 3 were consistent with that of the Bi-LSTM-CRF model in Table 2. On the validation and test sets, the model's precision decreased by about 0.65%, recall increased by about 0.6%, and the F1 score remained unchanged. Since the MTM-CW model shared training parameters, it was logical that the precision would decrease, recall would increase, and overall F1 score would remain unaffected compared to the benchmark Bi-LSTM-CRF model. The improvement of the generalization ability can make the model more competent for the extraction of multiple entity datasets. However, the overall training time of the model was much longer than that of the single entity dataset model.

The third model is BERT, pre-trained on Wikipedia, BooksCorpus, and fine-tuned on the WhiteText dataset. The performance of this model was comparable to the previous models on the validation set and test set. Compared to the results of the MTM-CW model in Table 3, the BERT model's recall increased by about

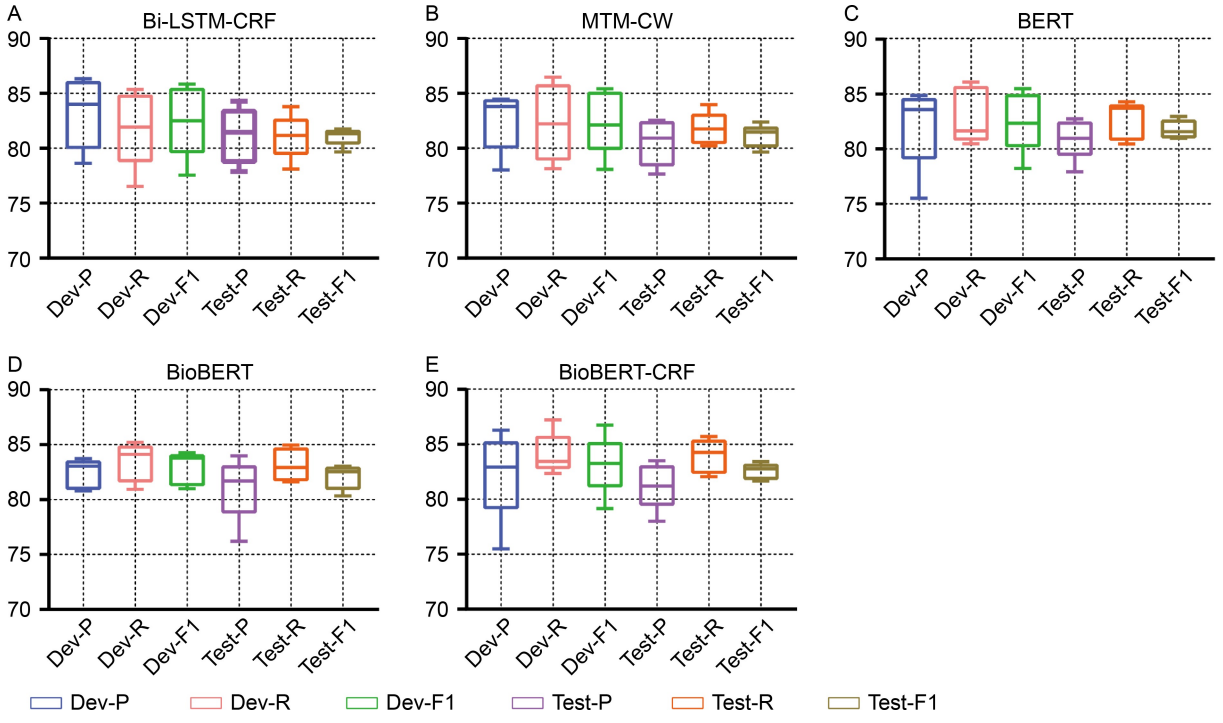


**Figure 1. Box diagrams of the five deep learning-based models.** The box diagrams show the centering, spread, and distribution of the validation set and test set. The mean of (E) is higher than (A), (B), (C), and (D)

**Table 1 Cases of the prediction entities from different models**

| | Models | Brain regions |
|---|---|---|
| Correct prediction | BERT | Connections…in primary somatosensory cortex (area 3b) and adjoining cortex were revealed in owl… |
| | BioBERT | Connections…in primary somatosensory cortex (area 3b) and adjoining cortex were revealed in owl… |
| | BioBERT-CRF | Connections…in primary somatosensory cortex (area 3b) and adjoining cortex were revealed in owl… |
| Wrong prediction | Bi-LSTM-CRF | Connections…in primary somatosensory cortex (area 3b) and adjoining cortex were revealed in owl… |
| | MTM-CW | Connections…in primary somatosensory cortex (area 3b) and adjoining cortex were revealed in owl… |
| True label | | Connections…in primary somatosensory cortex (area 3b) and adjoining cortex were revealed in owl… |

The true labels and the predicted labels of each model are underlined in the sentence.

**Table 2 Entity extraction results based on Bi-LSTM-CRF model**

| Number | Dev-P(%) | Dev-R(%) | Dev-F1(%) | Test-P(%) | Test-R(%) | Test-F1(%) |
|---|---|---|---|---|---|---|
| 1 | 86.32 | 85.36 | 85.84 | 77.89 | 81.54 | 79.67 |
| 2 | 84.01 | 81.06 | 82.51 | 81.46 | 81.17 | 81.32 |
| 3 | 78.63 | 76.52 | 77.56 | 84.29 | 78.11 | 81.08 |
| 4 | 81.33 | 81.92 | 81.62 | 79.39 | 83.79 | 81.53 |
| 5 | 85.84 | 84.31 | 85.07 | 82.76 | 80.76 | 81.75 |
| Average | 83.24 | 81.83 | 82.52 | 81.16 | 81.07 | 81.07 |

**Table 3 Entity extraction results based on MTM-CW model**

| Number | Dev-P(%) | Dev-R(%) | Dev-F1(%) | Test-P(%) | Test-R(%) | Test-F1(%) |
|---|---|---|---|---|---|---|
| 1 | 84.40 | 86.48 | 85.43 | 77.65 | 81.75 | 79.65 |
| 2 | 83.81 | 79.70 | 81.70 | 80.93 | 80.22 | 80.57 |
| 3 | 78.02 | 78.14 | 78.08 | 82.31 | 80.64 | 81.47 |
| 4 | 82.04 | 82.22 | 82.13 | 79.16 | 83.98 | 81.50 |
| 5 | 84.46 | 85.11 | 84.78 | 82.55 | 82.23 | 82.39 |
| Average | 82.55 | 82.33 | 82.42 | 80.52 | 81.76 | 81.12 |

0.6% and 0.91%, while the F1 score increased by about 0.1% and 0.65% on the validation set and test set (Table 4), respectively. The higher recall suggests that the BERT model can effectively reduce false-negative examples.

The fourth model is BioBERT. The average precision, average recall, and average F1 score of BioBERT on the validation set were 1.31%, 0.27%, and 0.84% higher than the test set, as shown in Table 5. The same scores of the Bi-LSTM-CRF benchmark model and MTM-CW model were about 2%, 1%, and 1.5% higher. Through comparison, it can be seen that BioBERT had a smaller overfitting degree. In addition, the various indicators of BioBERT showed less variance, with results between 80%−85%. Compared to BERT, the average precision, average recall, and average F1 score of BioBERT on the test set improved by 0.13%, 0.48%, and 0.3%, respectively, indicating better overall performance.

The final model is BioBERT-CRF, a modified BioBERT model. Compared to the previous four models, the BioBERT-CRF model achieved the best F1 score on both the validation and test set, likely due to the changes in its model structure, such as the change in self-attention heads and the addition of the CRF layer. The average precision of BioBERT-CRF is also identical to that of BioBERT. In contrast, its average recall and F1 score increased by 0.81% and 0.49% on the test set, respectively, as shown in Table 6.

Despite the effectiveness of the BioBERT-CRF model, there are still some error cases in the prediction entities, as shown in Table 7. In case 1, the true brain region entity is the “ventral lateral geniculate nucleus”. The BioBERT-CRF model failed to recognize the correct left boundary, probably because of the directional phrases in front. In case 2, the word “rectum” is not a true brain region entity, but the BioBERT-CRF model recognized it as an entity. The reason may be that the model has learned that entity words often follow species words. In case 3, “ependyma” and “choroid plexus” are all true brain region entities, but the

**Table 4 Entity extraction results based on BERT model**

| Number | Dev-P(%) | Dev-R(%) | Dev-F1(%) | Test-P(%) | Test-R(%) | Test-F1(%) |
|---|---|---|---|---|---|---|
| 1 | 84.88 | 86.10 | 85.49 | 77.93 | 84.29 | 80.98 |
| 2 | 84.31 | 80.48 | 82.35 | 80.91 | 81.13 | 81.02 |
| 3 | 75.53 | 81.14 | 78.23 | 82.74 | 80.46 | 81.58 |
| 4 | 82.70 | 81.66 | 82.18 | 80.97 | 83.71 | 82.32 |
| 5 | 83.60 | 85.29 | 84.44 | 82.18 | 83.75 | 82.96 |
| Average | 82.20 | 82.93 | 82.54 | 80.95 | 82.67 | 81.77 |

**Table 5 Entity extraction results based on BioBERT model**

| Number | Dev-P(%) | Dev-R(%) | Dev-F1(%) | Test-P(%) | Test-R(%) | Test-F1(%) |
|---|---|---|---|---|---|---|
| 1 | 83.32 | 85.21 | 84.25 | 76.20 | 84.97 | 80.35 |
| 2 | 80.79 | 82.28 | 81.53 | 81.38 | 81.61 | 81.49 |
| 3 | 81.05 | 80.94 | 80.99 | 83.97 | 81.80 | 82.87 |
| 4 | 83.74 | 84.12 | 83.93 | 81.69 | 84.46 | 83.05 |
| 5 | 83.05 | 84.55 | 83.79 | 82.18 | 82.92 | 82.55 |
| Average | 82.39 | 83.42 | 82.90 | 81.08 | 83.15 | 82.06 |

**Table 6 Entity extraction results based on BioBERT-CRF model**

| Number | Dev-P(%) | Dev-R(%) | Dev-F1(%) | Test-P(%) | Test-R(%) | Test-F1(%) |
|---|---|---|---|---|---|---|
| 1 | 86.30 | 87.23 | 86.76 | 78.00 | 85.70 | 81.67 |
| 2 | 84.17 | 82.36 | 83.25 | 81.20 | 82.67 | 81.93 |
| 3 | 75.49 | 83.21 | 79.16 | 83.50 | 82.06 | 82.78 |
| 4 | 82.79 | 83.44 | 83.11 | 80.91 | 85.09 | 82.94 |
| 5 | 82.93 | 84.25 | 83.59 | 82.62 | 84.26 | 83.43 |
| Average | 82.34 | 84.10 | 83.17 | 81.25 | 83.96 | 82.55 |

**Table 7 Cases of the prediction entities from the BioBERT-CRF model**

| | | Brain regions | Error type |
|---|---|---|---|
| Case1 | True label | Notable diencephalic afferents … of the internal division of the ventral lateral geniculate nucleus | Left boundary error: fail to detect the correct left boundary of the true entity due to some direction words in front |
| | BioBERT-CRF | Notable diencephalic afferents … of the internal division of the ventral lateral geniculate nucleus | |
| Case2 | True label | Neurochemical characterization of extrinsic innervation of the guinea pig rectum | Single vocabulary recognition error: recognize entity error due to species word in front |
| | BioBERT-CRF | Neurochemical characterization of extrinsic innervation of the guinea pig rectum | |
| Case3 | True label | In addition, … (Hsp27 IR) were detected in the ependyma and choroid plexus | Multiple vocabulary recognition error: fail to recognize multiple parallel entities |
| | BioBERT-CRF | In addition, … (Hsp27 IR) were detected in the ependyma and choroid plexus | |

The true labels and the predicted labels of the BioBERT-CRF model are underlined in the sentence. A brief summary of the error type is also included at the end of each example.

BioBERT-CRF model did not recognize them. This result may be because the model had not learned these entities in the train set.

In summary, the BioBERT-CRF model demonstrated good performance in extracting brain region entities, but not without its limitations. When recognizing a single brain region entity or multiple parallel brain region entities, it can easily fail to recognize or recognize the wrong entity. In addition, the model cannot detect boundaries correctly when directional words proceed entities.

We compared the results of five different named

entity recognition models with those reported in the WhiteText dataset literature. The comparison results are shown in Table 8.

The best model reported in literature is the Bi-LSTM-CRF model of Shardlow *et al*., which achieved an F1 score of 81.8%. This result is equivalent to that of the benchmark Bi-LSTM-CRF model in this article. The higher precision may be due to better pre-processing and model parameters. In contrast, the MTM-CW model achieved higher recall due to its generalization ability, while the BERT and BioBERT model achieved higher precision and recall. Among the four benchmark models, the BioBERT model demonstrated state-of-the-art performance. The average recall in 5 repeated experiments was as high as 83.2%, which is 4.4% higher than the traditional CRF model by Richardet *et al*. and 1.7% higher than the Bi-LSTM-CRF model by Shardlow *et al*. However, the average precision is lower than Richardet *et al*., likely because Richardet *et al*. combined several neuroanatomical lexica feature and species feature with linear chain CRF, allowing for higher precision and lower recall. The overall F1 score of the BioBERT model had a certain improvement effect. The BioBERT* model training with special skills further improved the precision rate, resulting in an F1 score of 82.6%. Compared to previous models, the BioBERT-CRF model achieved the highest F1 score, which can be contributed to its beneficial changes, such as the change in self-attention heads and the addition of the CRF layer.

## Brain region entity extraction in a large-scale PubMed abstract dataset

After comparing the accuracy of different deep learning-based models to predict brain region entities, our results showed that the BioBERT-CRF model demonstrated the best performance in brain region entity extraction. Therefore, the trained BioBERT-CRF model was chosen to extract brain region entities from a large-scale PubMed abstract dataset. A total of 27,436 abstracts were downloaded from PubMed using 399 brain regions from the Allen Brain Atlas (ABA) ontology [22]. To ensure the brain regions can be found in neuroscience-related literature, certain highly specific brain regions were removed from the ABA ontology, leaving 399 brain regions. Then, a maximum of 200 abstracts were randomly selected for each brain region and the trained BioBERT-CRF model was used to predict brain region entities in a database containing 27,436 PubMed abstracts. The abstracts were pre-processed into the CoNLL-U format [2] and inputted into the BioBERT-CRF model. The results contained a total of 153,069 predicted brain region entities from the abstracts, which were all normalized according to three neuroscience dictionaries, specifically ABA ontology, BAMS ontology [23], and NeuroNames lexicon [24]. This step can solve the problems posed by synonym use and different naming systems. For example, the term "midbrain" has various abbreviations. In ABA and BAMS, "midbrain" is abbreviated as "MB", but in NeuroNames it is abbreviated as "MBr". Since the different abbreviations correspond to the same anatomical region, a rule-based method was used to match the brain region entities to their corresponding terms in the three neuroscience dictionaries. A total of 85.2% of the brain region entities were matched. For the remaining unmatched entities, all directional terms were removed and rematched. Finally, the number of different brain region entities appeared in all 27,436 abstracts was counted. Different colors were used to represent the different brain regions, gradients were used to represent

**Table 8 Comparison of five different entity extraction models and literature report models**

| Literature | Model | Test-P(%) | Test-R(%) | Test-F1(%) |
|---|---|---|---|---|
| French *et al*. [18] | CRF | 0.813 | 0.761 | 0.786 |
| Richardet *et al*. [20] | CRF | **0.846** | 0.788 | 0.816 |
| Shardlow *et al*. [21] | Bi-LSTM-CRF | 0.821 | 0.815 | 0.818 |
| Wang *et al*. [3] | Bi-LSTM-CRF | 0.812 | 0.811 | 0.811 |
| | MTM-CW | 0.805 | 0.818 | 0.811 |
| Devlin *et al*. [16] | BERT | 0.810 | 0.827 | 0.818 |
| Lee *et al*. [17] | BioBERT | 0.811 | 0.832 | 0.821 |
| | BioBERT* | 0.823 | 0.830 | **0.826** |
| Our models | BioBERT-CRF | 0.813 | **0.840** | **0.826** |

Precision (P), Recall (R), and F1 scores (F1) on each model are reported. Best scores are in bold texts. The BioBERT* model was trained with special skills.

the different levels of each brain region in the mammalian sagittal plane, and different sizes were used to represent the total number of brain region entities that appeared in the literature, as shown in Fig. 2. The results suggest that researchers are more interested in the cerebral cortex and brainstem, which highly reflect the reality of neuroscience research.

The top 30 brain region entities from the 27,436 abstracts are shown in Table 9. Results show the cerebrum as the anatomical region researchers are most

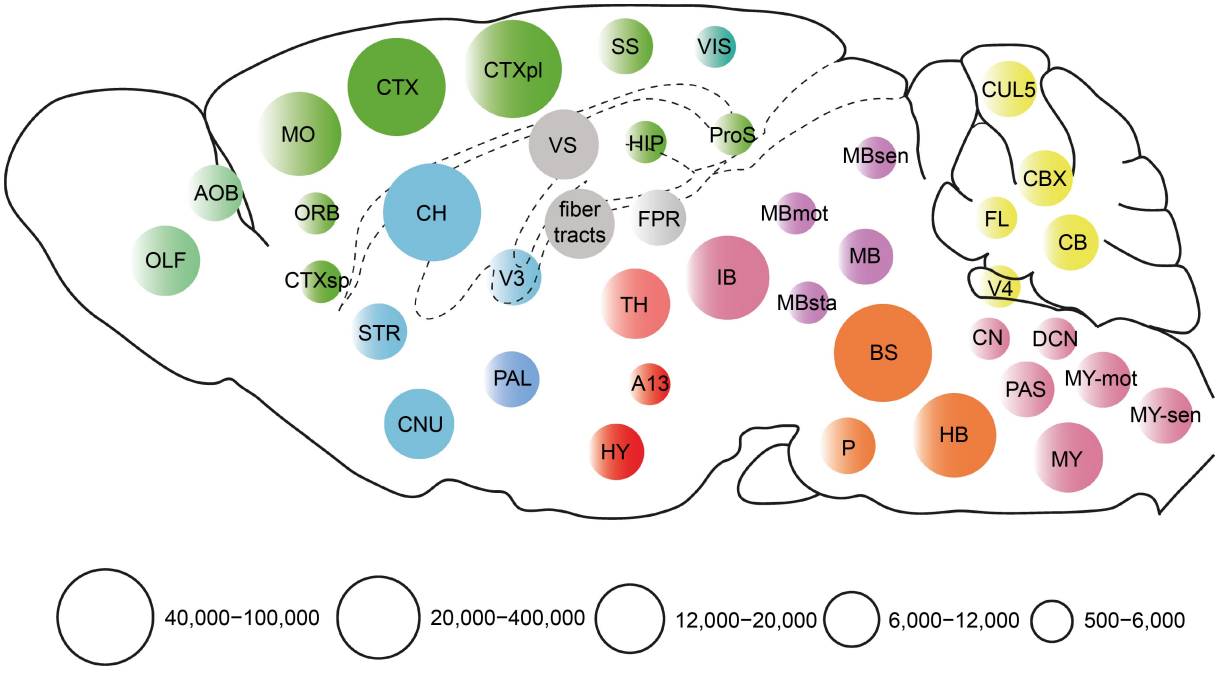


**Figure 2. The number of different brain region entities in the mammalian sagittal plane.** The different colors represent different brain regions in the mammalian sagittal plane. The different sizes represent the number of brain region entities that appeared in the literature.

**Table 9 The top 30 of the brain region entities in the literature**

| Number | Brain region | Frequency | Number | Brain region | Frequency |
|---|---|---|---|---|---|
| 1 | Cerebrum | 103,149 | 16 | Striatum | 9,340 |
| 2 | Cerebral cortex | 87,722 | 17 | Primary motor area | 9,319 |
| 3 | Cortical plate | 85,701 | 18 | Secondary motor area | 8,588 |
| 4 | Brain stem | 51,640 | 19 | MY-mot | 8,376 |
| 5 | Somatomotor areas | 25,374 | 20 | Somatosensory areas | 7,799 |
| 6 | Hindbrain | 21,996 | 21 | Hypothalamus | 7,710 |
| 7 | Interbrain | 20,109 | 22 | CUL5 | 7,700 |
| 8 | Olfactory areas | 17,318 | 23 | V3 | 7,688 |
| 9 | Medulla | 15,869 | 24 | Accessory olfactory bulb | 7,537 |
| 10 | Cerebral nuclei | 15,427 | 25 | MY-sen | 7,493 |
| 11 | VS | 14,713 | 26 | Midbrain | 7,399 |
| 12 | fiber tracts | 13,694 | 27 | Fasciculus proprius | 7,349 |
| 13 | Thalamus | 12,399 | 28 | Parasolitary nucleus | 7,212 |
| 14 | Cerebellum | 10,736 | 29 | Pons | 6,127 |
| 15 | Cerebellar cortex | 9,570 | 30 | Pallidum | 6,087 |

concerned about, appearing a total of 103,149 times. We can also see that, within the cerebrum, the cerebral cortex is researched more often than the cerebral nuclei. In addition, researchers are also interested in the brainstem, somatomotor areas, and hindbrain, among other anatomical regions.

## CONCLUSIONS

This study used a modified and four benchmark deep learning models, specifically the BioBERT-CRF, Bi-LSTM-CRF, MTM-CW, BERT, and BioBERT model, to extract brain region entities on the WhiteText dataset. The benchmark model, Bi-LSTM-CRF, achieved an 81.1% F1 score using the semantic information of the text to predict the brain region entity. The MTM-CW model demonstrated improvement in neuroscience brain region entity recognition due to its multiple added training datasets in the biomedical field, resulting in an 81% recall of the recognition task. The BERT, BioBERT, and BioBERT-CRF models were trained in two steps: pre-training and fine-tuning. Unlike BERT, the BioBERT model pre-trained with the PubMed abstracts and PMC full texts, which resulted in a better recall of 83.2% and a better F1 score of 82.1%. On this basis, we added a CRF layer and changed the self-attention heads in the BioBERT model. The modified BioBERT-CRF model achieved the best recall and F1 score of 84.0% and 82.6%. The pre-training using biomedical literature data helped the model learn the semantics of the field. The fine-tuning using the WhiteText dataset helped the model learn the semantics of the brain region entity. The transition matrix of the CRF layer helped reduce error examples. This combination achieved the best prediction results. However, there are still limitations with the BioBERT-CRF model. When recognizing a single vocabulary brain entity or multiple parallel vocabulary brain entities, it can easily fail to recognize or recognize incorrectly. There are also cases of inaccurate boundary detection when directional phrases proceed the entity. The error examples are shown in the results section. The entity recognition boundary issues and single entity non-recognition issues should be addressed and improved in future work.

We used the trained BioBERT-CRF model to predict brain region entities in the 27,436 PubMed abstract dataset and obtained a total of 153,069 entities. We then used a rule-based method to normalize all brain region entities according to three neuroscience dictionaries, specifically ABA ontology, BAMS ontology, and NeuroNames lexicon. A total of 85.2% of the brain region entities were matched to the terms found in these dictionaries. Although the normalization pipeline solved the problems of synonyms and different naming systems, the accuracy of standardization required further improvement by introducing additional annotated terms into the dictionary. The brain region entity recognition model and rule-based normalization pipeline can then extract brain regions from a large set of neuroscience literature, such as the PubMed abstract dataset, and construct a brain region-to-abstract map in a context-sensitive approach to serve as the basis for a neuroscience oriented search engine. In addition, the brain region entities were sorted in order of their frequency in the corpus, reflecting the anatomical regions researchers are most concerned about. In the future, we will extract the relationship between different brain region entities recognized by the BioBERT-CRF model described in this work. The relation extraction will allow us to obtain further information from the literature. Searching for two or more brain entities will provide us with a more accurate literature list, while extracting information on the relationship between different brain entities will help us learn about the neuroanatomical connectivity among brain structures from a bird's eye view.

## MATERIALS AND METHODS

### Dataset

We used the White Text dataset to evaluate our proposed framework. The WhiteText dataset was a brain region entity gold standard dataset collected by French *et al*. in 2009. The dataset first selected 1,377 abstracts from a multitude of JCN journals with brain connections through keyword search. Then, it used an abbreviation amplification algorithm to replace the brain abbreviations mentioned in the abstracts with their corresponding full names. Next, they invited several neuroscientists to manually label brain region entities based on neuroscience knowledge, reference brain atlases, and dictionaries [18]. The 1,377 articles contained 18,242 labeled brain entities, and the agreement rate among labelers was 90.7%. We converted the XML format of the WhiteText dataset to the most commonly used CoNLL-U format [2] and acquired 17,585 brain region entities. The segmented text and corresponding tags were represented line by line, with tab intervals, and sentences were separated with blank lines. The beginning of the document was marked with "-DOCSTART- -X- -X- -X- O". The label can select either the BIO or BIOES format, where "B" represents the beginning of the entity, "I" represents the middle of the entity, "E" represents the end of the entity, "O" represents a non-entity, and "S" represents a single entity.

### Word embeddings

In the Bi-LSTM-CRF and MTM-CW model, we initialized the word embedding matrix with pre-trained word vectors from word2vec, obtained by Pyysalo *et al*. [25], using Wikipedia corpus, PubMed abstracts, and PMC full-text training. These word embeddings were trained using a skip-gram model, as described by Mikolov *et al*. [26]. In the BERT and BioBERT model, word2vec was not used to initialize the word embedding matrix since they learned the WordPiece embeddings [27] from scratch during pre-training. The WordPiece tokenization divided each word into a limited set of standard sub-word units and learned all the sub-word unit's vectors. The pre-training step was also carried out using the Wikipedia corpus, PubMed abstracts, and PMC full-text. The embedding dictionary saved its vectors for each word.

### Evaluation

In NER, the predicted results can be true positive (TP), false positive (FP), true negative (TN), or false negative (FN), based on the true label. For the prediction results, the most common evaluation indicators are precision, recall, and F1 score. The precision represents the number of entities that the model predicts correctly to the total number of entities whose predictions are positive, as defined:

$$P = \frac{TP}{TP + FP}$$

Recall refers to the ratio of the number of entities that the model correctly predicts to the number of real entities marked on the dataset, as defined:

$$R = \frac{TP}{TP + FN}$$

The F1 score compromises the precision and recall, thus being more representative. It is defined as follows:

$$F1 = \frac{2PR}{P + R}$$

### Model

BioBERT-CRF: in natural language processing, transfer learning has been used to pre-train neural network language models with a large amount of unstructured text data and subsequently fine-tune the target task to solve the missing gold standard in the target domain. BERT is a representative result of the pre-training model [16]. The BERT model can achieve better performance by pre-training on the corpus and fine-tuning on the specific dataset. As a pre-training model in the biomedical field, BioBERT has achieved better results than the Bi-LSTM-CRF model and BERT model in many natural language processing tasks [17]. Therefore, we applied the BioBERT model to entity extraction in the neuroscience field, fine-tuned it, and evaluated the results of the neuroscience dataset extraction. In addition, we added a CRF layer to the BioBERT model's output to calculate label probability.

The BioBERT-CRF model first obtains the feature vector of the input text through three embedding features, specifically WordPiece embedding [27], position embedding, and segmentation embedding. WordPiece tokenization divides the word into a limited set of standard sub-word units. Position embedding encodes the position information of the word into a feature vector. Segmentation embedding distinguishes between two sentences. For sentence pairs, segmentation embedding sets the feature value of the first sentence to 0 and second sentence to 1. Then, the feature vector is inputted into the bidirectional transformers. The transformer's encoder is composed of multi-head attention and a full connection is used to convert the input corpus into a feature vector. The transformer's decoder inputs the output of the encoder and predicts the result. The decoder is composed of masked multi-head attention, multi-head attention, and a full connection, which are used to output the conditional probability of the final result. Then, the model receives the predicted label sequence through the linear classifier. Next, the predicted label sequence is inputted into the CRF layer, which will multiply the output of the BioBERT layer with the parameter matrix to obtain the state transition matrix $A$, where $A_{ij}$ represents the transition probability from the $i^{th}$ label position to the $j^{th}$ label position. The label output at different positions of the sequence is calculated by the sum of the output $p_i$ of the BioBERT layer and the transition matrix $A$ of the CRF layer. The probability value is normalized by the softmax activation function. The maximum likelihood function will calculate the loss and the gradient descent algorithm is used to calculate the model parameters. Finally, the model parameters are updated layer by layer through the backpropagation algorithm and the neural network is fine-tuned. The architecture of the model is shown in Fig. 3.

The parameters of the BioBERT-CRF model are: the initial learning rate is $5 \times 10^{-5}$, the epoch is 10, the batch size is 32, the dropout size of the attention layer and hidden layer is set to 0.1, the optimization algorithm is Adam, the self-attention heads are 24, and the remaining parameters are the default BioBERT model settings.

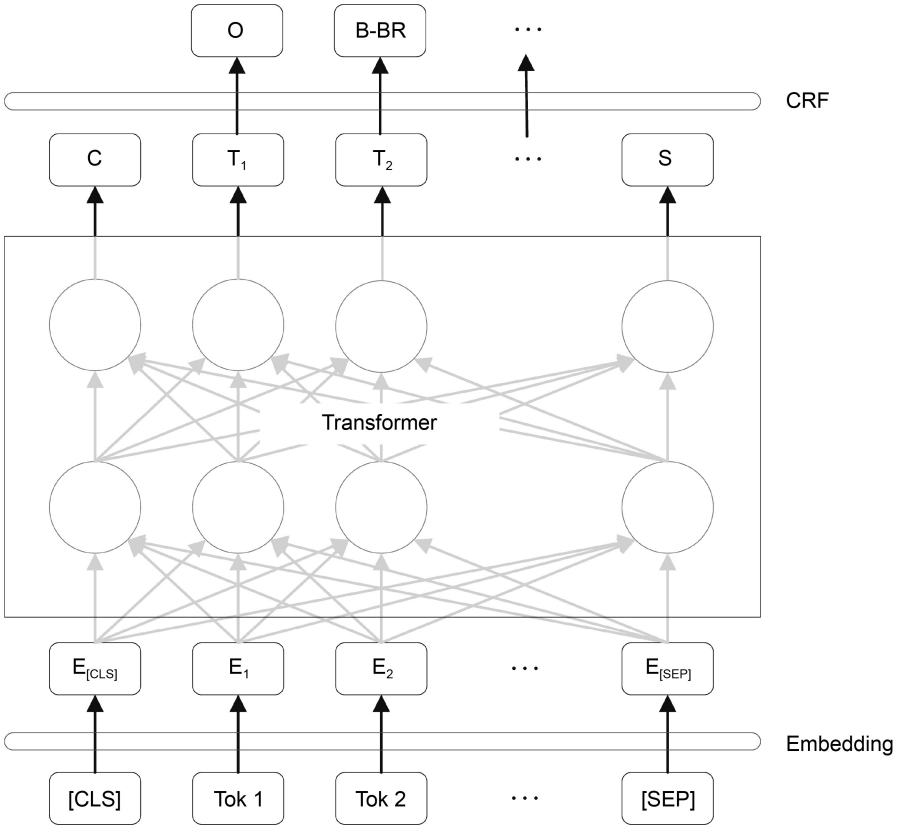


**Figure 3. The architecture of BioBERT-CRF model.**


## ACKNOWLEDGEMENTS

We thank the MOST group members of Britton Chance Center for Biomedical Photonics for assistance with experiments and comments on the manuscript. We also thank Nannan Li for the helpful discussions. This work was supported by the National Science and Technology Innovation 2030 Grant (No. 2021ZD0201002), the National Natural Science Foundation of China (Nos. T2122015 and 61890954), CAMS Innovation Fund for Medical Sciences (No. 2019-I2M-5-014) and Suzhou Prospective Application Research Project (No. SYG201915).


## COMPLIANCE WITH ETHICS GUIDELINES

The authors Xiaokang Chai, Yachao Di, Zhao Feng, Yue Guan, Guoqing Zhang, Anan Li and Qingming Luo declare they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

## OPEN ACCESS



# REFERENCES


1. Sun, L., Patel, R., Liu, J., Chen, K., Wu, T., Li, J. and Ye, J. (2009) Mining brain region connectivity for Alzheimer's disease study via sparse inverse covariance estimation. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1335−1344
2. Sang, E. F., and De Meulder, F. (2003) 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, pp. 142−147
3. Wang, X., Zhang, Y., Ren, X., Zhang, Y., Zitnik, M., Shang, J., Langlotz, C. and Han, J. (2019) Cross-type biomedical named entity recognition with deep multi-task learning. Bioinformatics, 35, 1745−1752
4. Yoon, W., So, C. H., Lee, J. and Kang, J. (2019) CollaboNet: collaboration of deep neural networks for biomedical named entity recognition. BMC Bioinformatics, 20, 249
5. Cho, M., Ha, J., Park, C. and Park, S. (2020) Combinatorial

feature embedding based on CNN and LSTM for biomedical named entity recognition. J. Biomed. Inform., 103, 103381
6. raj Kanakarajan, K., Kundumani, B. and Sankarasubbu, M. (2021) Bioelectra: Pretrained biomedical text encoder using discriminators. In: Proceedings of the 20th Workshop on Biomedical Language Processing, pp. 143−154
7. Leaman, R., Islamaj Doğan, R. and Lu, Z. (2013) DNorm: disease name normalization with pairwise learning to rank. Bioinformatics, 29, 2909−2917
8. Dang, T. H., Le, H. Q., Nguyen, T. M. and Vu, S. T. (2018) D3NER: biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information. Bioinformatics, 34, 3539−3546
9. Peng, Y., Chen, Q. and Lu, Z. (2020) An empirical study of multi-task learning on bert for biomedical text mining. In: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, pp. 205−214
10. Leaman, R. and Lu, Z. (2016) TaggerOne: joint named entity recognition and normalization with semi-Markov models. Bioinformatics, 32, 2839−2846
11. Wei, C. H., Harris, B. R., Kao, H. Y. and Lu, Z. (2013) tmVar: a text mining approach for extracting sequence variants in biomedical literature. Bioinformatics, 29, 1433−1439
12. Wei, C. H., Phan, L., Feltz, J., Maiti, R., Hefferon, T. and Lu, Z. (2018) tmVar 2. 0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. Bioinformatics, 34, 80−87
13. Wei, C. H., Kao, H. Y. and Lu, Z. (2012) SR4GN: a species recognition software tool for gene normalization. PLoS One, 7, e38460
14. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X. and Poon, H. (2020) Domain-specific language model pretraining for biomedical natural language processing. arXiv, 200715779
15. Giorgi, J. M. and Bader, G. D. (2018) Transfer learning for biomedical named entity recognition with neural networks. Bioinformatics, 34, 4087−4094
16. Devlin, J., Chang, M. W., Lee, K. and Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv, 181004805
17. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H. and Kang, J. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36, 1234−1240
18. French, L., Lane, S., Xu, L. and Pavlidis, P. (2009) Automated recognition of brain region mentions in neuroscience literature. Front. Neuroinform., 3, 29
19. Lafferty, J., McCallum, A. and Pereira, F. C. (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282−289
20. Richardet, R., Chappelier, J. C., Telefont, M. and Hill, S. (2015) Large-scale extraction of brain connectivity from the neuroscientific literature. Bioinformatics, 31, 1640−1647
21. Shardlow, M., Ju, M., Li, M., O'Reilly, C., Iavarone, E., McNaught, J. and Ananiadou, S. (2019) A text mining pipeline using active and deep learning aimed at curating information in computational neuroscience. Neuroinformatics, 17, 391−406
22. Oh, S. W., Harris, J. A., Ng, L., Winslow, B., Cain, N., Mihalas, S., Wang, Q., Lau, C., Kuan, L., Henry, A. M., *et al*. (2014) A mesoscale connectome of the mouse brain. Nature, 508, 207−214
23. Bota, M. and Swanson, L. W. (2008) Bams neuroanatomical ontology: Design and implementation. Front. Neuroinform., 2, 2
24. Bowden, D. M. and Dubach, M. F. (2003) NeuroNames 2002. Neuroinformatics, 1, 43−59
25. Pyysalo, S., Ginter, F., Moen, H., Salakoski, T. and Ananiadou, S. (2013) Distributional semantics resources for biomedical text processing. In: Proceedings of LBM, pp. 39−44
26. Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) Efficient estimation of word representations in vector space. arXiv, 13013781
27. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W. and Dean, J. (2016) Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv, 160908144

 