Scholar Publishing Group
International Journal of Big Data Intelligent Technology
https://doi.org/10.38007/IJBDIT.2025.060112
ISSN 2790-0932 Vol. 6, Issue 1: 118-125

Research on Medical Named Entity Recognition Technology Based on Prompt BioMRC Model for Deep NLP Algorithm

Xiangtian Hui[1,a*]

[1]*School of Professional Studies, New York University, New York, NY, 10012, U.S.A*

[a]*xiangtian.hui.us@gmail.com*

[*]*corresponding author*

Keywords: Biomedical entity recognition, deep learning methods, character enhanced entity recognition models, part of speech and dependency relationships, natural language processing

Abstract: Biomedical science has seen remarkable technological progress, resulting in vast amounts of research findings and clinical information. However, processing this unstructured data manually presents significant challenges. Natural language processing (NLP), particularly biomedical entity recognition (BioNER), offers promising solutions in this field. BioNER focuses on identifying medical and biological entities within scientific texts but encounters several obstacles. This paper examines common deep learning approaches and introduces two new models: a character-enhanced recognition system (Bert CNN CapsNET) utilizing capsule networks, and a biomedical entity recognition system (BioBERT GAT-NET) that incorporates grammatical structures and word relationships. Testing confirms that both models substantially improve entity identification accuracy. Nevertheless, BioNER technology continues to develop. Future research directions include analyzing combined data types, recognizing entities across different languages, applying knowledge from one domain to another, adapting to specific fields, and developing more transparent and reliable models. These advancements will enhance performance and facilitate broader implementation in both research and clinical environments.

1. Introduction

The biomedical field has witnessed rapid technological advancements, generating vast amounts of research and clinical data, which are extensively published in databases like Medline. As literature volumes surge, manually extracting information from unstructured data (such as scientific research data, electronic health records, and clinical trial data) has become increasingly challenging. Therefore, there is an urgent need to develop automated text retrieval and content extraction

technologies, with Biomedical Named Entity Recognition (BioNER) demonstrating significant potential. BioNER can accurately identify medical entities like genes, proteins, drugs, and diseases from textual data. Despite remarkable progress, BioNER still faces multiple challenges: diverse standards for biomedical entities, varied names or abbreviations for the same entity, complex rules for composite entities, and the involvement of specialized terminology and domain knowledge. Additionally, issues such as nested entities, cross-language processing, and novel entity recognition need to be addressed. Traditional methods relied on rules, dictionaries, or machine learning, but modern research has shifted towards deep learning approaches. By leveraging word embeddings (e.g., Word2Vec, BERT) in conjunction with Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks, deep learning models automatically learn hidden features from large datasets, showcasing powerful learning and processing capabilities. In recent years, deep learning-based BioNER has been widely applied in biomedical text entity recognition, emerging as a research hotspot.

2. Correlation Theory

In recent years, significant achievements have been made in the interdisciplinary field of medicine and natural language processing (NLP). Researchers have fully utilized pre trained Clinical BERT and NLP deep learning models in this field, successfully predicting international disease classification from MIMIIC-III clinical texts and achieving the best level in the field. The research team also delved into the broad applications of NLP and deep learning in medical text analysis, hoping to improve emotional intelligence through artificial intelligence methods. A natural language processing deep learning model specifically designed to distinguish high-grade gliomas from metastases has been constructed, and the important impact of medical report writing methods on model prediction accuracy has been explored in depth. Domestic scholars have also made outstanding contributions in the field of medical named entity recognition. They not only conducted in-depth research on named entity recognition in the medical field based on prompt learning methods, but also proposed a medical named entity recognition model based on knowledge graph enhancement and a recognition method based on ontology attention layer. A research team has also developed a specialized NLP model to achieve accurate recognition of medical named entities, and presented their research results at relevant academic conferences. The researchers also innovatively proposed the MedNER system, which significantly improves the named entity recognition ability in medical corpora by optimizing balance and deep active learning strategies, injecting new vitality into the development of medical NLP field.

3. Method

3.1 Entity recognition method based on rules and dictionaries

Deep learning, as an advanced machine learning method, mimics the neural network structure of the human brain, utilizing multi-layer neural networks for feature extraction and data modeling to achieve efficient representation and analysis of complex data. In entity recognition tasks, deep learning frameworks typically involve three main steps: input distribution representation, feature extraction, and label decoding. They can automatically learn features from large amounts of data without the need for manual feature engineering, significantly reducing the need for manual intervention. Compared with traditional machine learning methods, deep learning exhibits better performance in processing entities in text, capable of handling more complex patterns and relationships. The pre-training and fine-tuning strategy has garnered widespread attention in the field of natural language processing. This strategy adopts the concept of transfer learning, first conducting unsupervised learning on a large-scale unlabeled corpus, and then fine-tuning the

parameters of the pre-trained language model for specific downstream tasks. This approach effectively reduces the training cost for downstream tasks and decreases the reliance on annotated corpora. Currently, mainstream pre-trained language models, such as ELMo, BERT, GPT, and BioBERT, can learn domain-specific semantic information and contextual relationships in entity recognition tasks, thereby enhancing the accuracy and generalization capability of entity recognition. They can capture complex structures and relevant information in text, effectively addressing contextual dependency and ambiguity in entity recognition tasks. Furthermore, pre-trained models can transfer knowledge and representations learned from general domains to specialized fields like biomedicine through transfer learning, accelerating the training and optimization process of models in these fields and improving their performance and efficiency. For example, some studies have incorporated the BERT pre-training method into the BiLSTM-CRF model for geographical entity recognition, while others have combined the BER (a typo, likely meaning BERT) pre-trained language model with BiLSTM-CRP (possibly a variation or typo, typically BiLSTM-CRF) for entity recognition on biomedical datasets, achieving remarkable results. These studies demonstrate the advantages and potential of deep learning methods and pre-trained models in entity recognition tasks.

3.2Rule-based and dictionary-based entity recognition methods

Rule-based and dictionary-based entity recognition methods primarily rely on predefined rules and manually curated dictionaries to identify entities. Specifically, rule-based approaches require pre-defining a series of patterns matching entity characteristics, which are then applied to the target text. Dictionary-based methods depend on exhaustive lists of predefined entities, matching text segments against these lexicons. This approach excels in scenarios with fixed entity types, limited quantities, and standardized expressions, making it widely adopted in early entity extraction tasks. Pioneering systems relied heavily on handcrafted rules and domain-specific dictionaries (e.g., gazetteers), combined with syntax-lexical patterns, to achieve efficient recognition. Such systems demonstrated high performance in specialized domains like Arabic or Urdu, where domain-specific rules and dictionaries could achieve both speed and accuracy when entities were well-defined and consistently structured. For instance, Essayed et al.'s Arabic NER system integrated local grammar, named entity dictionaries, and filtering techniques to attain high F1 scores. Similarly, Riaz's Urdu NER system employed multiple manually designed rules, surpassing prior methods across datasets. Despite their strengths, these methods face inherent limitations. Manual rule/dictionary creation is time-consuming and costly, particularly for large-scale data requiring extensive coverage. Domain specificity further restricts adaptability, necessitating continuous expert maintenance as data evolves. Cross-domain application proves challenging due to the need for complex domain-specific rule integration. While effective in controlled scenarios, these limitations constrain broader applicability in dynamic, multi-domain environments.

3.3Multi layer graph attention network

Graph Neural Networks (GNNs) extend the idea of passing messages over graph-structured data by propagating embedded information through multiple neural network layers among nodes. Each node aggregates messages from its neighbors to update its state, thereby obtaining a richer representation of information about the graph. In Graph Attention Networks (GATs), the introduction of an attention mechanism enables the network to learn the importance of different neighbors and dynamically adjust the influence of neighboring nodes on the target node based on this, effectively capturing complex relationships between nodes. For a single-layer GAT, each node trains its corresponding attention over its neighboring nodes, calculates attention values using learnable weight matrices and vectors, and normalizes the attention weights using softmax. The

normalized attention weights are then used to weight and sum the features of neighboring nodes, followed by a non-linear transformation through an activation function to obtain the output value of the node. A single-layer GAT may not fully extract node features and capture complex relationships. Therefore, multi-layer GATs have emerged. Multi-layer GATs extract node features layer by layer by stacking multiple graph attention layers. The attention weights for each layer are dynamically calculated based on the node features of the current layer, allowing the network to learn complex relationships between different features. For an L-layer GAT, the output of each layer is a weighted sum of the features of the nodes and their neighboring nodes from the previous layer, followed by a non-linear transformation through an activation function. The final layer performs a linear transformation and activation function processing on the output of each layer to obtain the final node representation.

4 Results and discussion

4.1 Graph data text topology modeling

This chapter introduces parts of speech and syntactic dependencies in language, aiming to model the topological structure of text using a graph data structure to capture the complex relationships between words. Specifically, let's assume there is an input sequence composed of elements, where each M represents a word or character. To transform this input sequence into a graph structure, we define a label set M = {B, I, O, UNK, PAD}, where B, I, and O represent the beginning, inside, and outside parts of an entity word, respectively, UNK denotes an unknown label, and PAD is used for padding. In this graph structure, each word is treated as a node, and the relationships between words form the edges of the graph. Based on the learned graph structure model, we can predict the label sequence of the input sequence Z, essentially transforming the BioNER (Biomedical Named Entity Recognition) problem into a node classification problem within the graph structure. The model is illustrated in Figure 1.

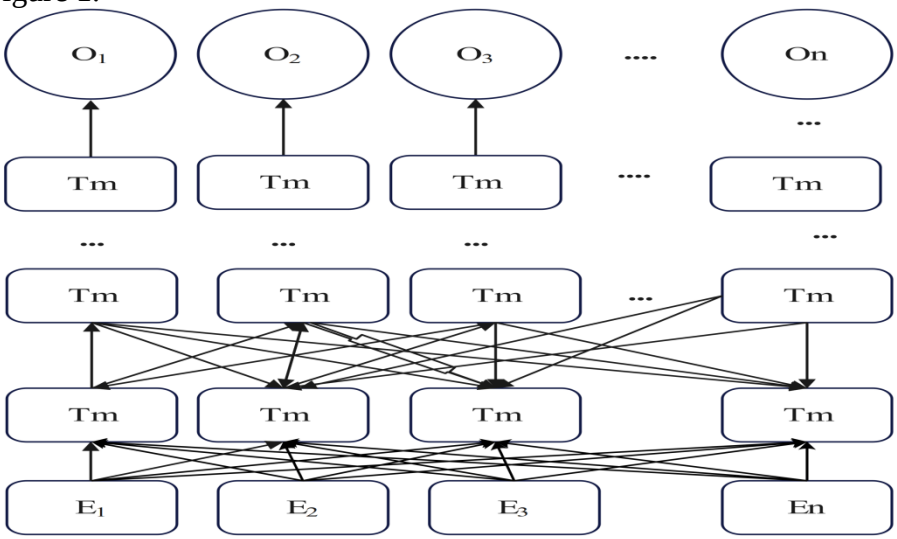

Figure 1 Structure of BioBERT Model

To enhance node classification, this chapter introduces parts of speech (POS) and syntactic dependencies. POS categories (e.g., nouns, verbs, adjectives) help identify specialized biomedical entities like drugs or diseases by leveraging grammatical context, while syntactic dependencies model structural relationships between words (e.g., modifiers, compounds) to clarify sentence roles.We construct POS and dependency tables to encode linguistic features, then model text topology via graph structures. This framework captures complex word interactions, improving BioNER accuracy and offering a transferable approach for NLP tasks requiring deep linguistic analysis.

4.2Experimental setup and data processing process

In the experimental setup, we utilized the en_corew_eb_trf model from the SpaCy library to process biomedical text data. The workflow of this model involves receiving text input and outputting a Doc object, which consists of a sequence of tokens, each accompanied by corresponding annotation information. To optimize the model's performance, we calculated the frequency distribution of parts of speech and syntactic dependencies in the dataset and used a 1% threshold for filtering. Ultimately, we obtained 10 typical parts of speech (such as "NOUN", "PROPN", "PUNCT", etc.) and 18 typical dependencies (such as "compound", "punct", "nmod", etc.).During the data processing phase, we first cleaned the words by removing spaces and illegal characters. Then, we performed sentence segmentation, tokenization, and lemmatization on the text. Sentence segmentation was based on extracting punctuation marks from ASCII encoding values, while English tokenization was separated by spaces, and lemmatization was accomplished using the SpaCy library. We also conducted part-of-speech tagging and syntactic dependency parsing on the sentences.Taking the sentence "Identification of APC2, homology of the adenomatous polyposis coli tumor suppressor" as an example, we demonstrated the annotation results for parts of speech and dependencies. For instance, "Identification" was tagged as a noun and served as the root node of the sentence, while "of" was tagged as a preposition, indicating a prepositional phrase relationship.We constructed training samples, which included token sequences, input_ids (token indices in the BioBERT model vocabulary, with a maximum sequence length of 128 and zero-padding for missing parts), input_mask (a mask where valid positions are 1), segment_ids (used for sentence separation, where all 0s indicate a single sentence), and labelIds (label indices for entity tags). These training samples were fully prepared for the subsequent training process.

4.3Analysis of Experimental Results and System Implementation

Table 1 Comparison of NER models on different datasets

Method	Corpus	Precision (%)	Recall (%)	F1 (%)
DTranNER	BC2GM	84.05	84.12	84.08
L-BioBERT	BC2GM	83.59	83.39	83.74
Proposed	BC2GM	83.94	84.44	84.19
DTranNER	BC4Chemd	91.94	92.04	91.99
Proposed	BC4Chemd	92.12	92.31	92.21
DTranNER	BC5CDR-Chemical	92.28	92.04	92.16
L-BioBERT	BC5CDR-Chemical	92.74	92.79	92.77
Proposed	BC5CDR-Chemical	93.09	93.13	93.11
DTranNER	BC5CDR-Disease	84.10	86.62	85.34
Proposed	BC5CDR-Disease	84.15	86.66	85.39
DTranNER	NCBI-Disease	88.21	89.04	88.62
L-BioBERT	NCBI-Disease	86.92	89.27	88.08
Proposed	NCBI-Disease	86.46	91.62	88.97

This chapter offers a comprehensive analysis of the BioBERT-GAT-NET model's performance in biomedical entity recognition tasks and further substantiates the significance of part-of-speech and syntactic dependency features through feature ablation experiments. A biomedical entity recognition system has been developed utilizing the proposed model, capable of identifying

biomedical entities like genes, compounds/drugs, and diseases within sentences, and generating a biomedical knowledge graph based on syntactic relationships. In the assessment of model effectiveness, the BioBERT-GAT-NET model exhibited outstanding performance across multiple biomedical corpora, including BC2GM, BC4Chemd, BC5CDR, and NCBI Disease. Compared to similar models such as DTranNER, Kocaman, and L-BioBERT, the model presented in this chapter demonstrates significant improvements in accuracy, recall, and F1 score. Notably, on the BC4Chemd and BC5CDR Chemical datasets, the model's F1 scores surpassed 90%, highlighting its robust capability in recognizing compound entities. The results are presented in Table 1.

The proposed model has shown satisfactory results in identifying disease and gene entities in biomedical texts. Feature ablation experiments highlight the importance of part-of-speech and syntactic dependency features, with performance dropping significantly upon their removal, especially dependency relationships. Even without these features, the model, relying solely on BioBERT's token embeddings, outperforms the previous model, attesting to BioBERT's strength in biomedical text processing. A web-based biomedical entity recognition system has been implemented using the Flask framework. Users can input text, and the system processes it to output annotated entities. The system also recognizes entities in real-time, extracts relationships, and constructs knowledge graphs, supporting biomedical knowledge mining and clinical services. Future research aims to refine and enhance the system's practicality and accuracy.

5. Conclusion

With the rapid advancement of biomedical research, an enormous amount of unstructured literature and data emerges daily, making manual processing of this knowledge-rich content infeasible. As a result, Biomedical Named Entity Recognition (BioNER) technology has emerged, effectively assisting researchers in quickly extracting key information from these vast texts, accurately identifying diseases and related genetic variations, thereby aiding medical researchers in gaining a deeper understanding of disease mechanisms, providing personalized treatment plans for patients, and accelerating the process of drug discovery and development. This paper focuses on the task of biomedical named entity recognition, conducting an in-depth study of widely used deep learning methods and proposing innovative solutions to address existing issues and challenges. Specifically, we first design a character-enhanced entity recognition model, Bert CNN CapsNET, based on capsule networks. This model significantly enhances entity recognition performance by capturing the internal character structure features of words. Experimental results demonstrate that the model excels in character feature extraction, confirming the importance of convolutional neural networks and capsule networks in capturing entity character features. Furthermore, this paper investigates the impact of part-of-speech and syntactic features on entity recognition performance and proposes a biomedical entity recognition model, BioBERT-GAT-NET, that integrates part-of-speech and dependency relationships. This model utilizes graph attention networks to process dependency relationships within syntactic features, further enhancing the model's semantic understanding capabilities and significantly improving the accuracy of entity recognition. However, biomedical named entity recognition technology is still evolving. Future research can explore areas such as multimodal data processing, cross-language entity recognition, transfer learning, domain adaptation, and improving model interpretability and trustworthiness. These efforts aim to continuously enhance technical performance, adapt to various needs, and promote the widespread application of BioNER in medical research and clinical practice.

References:

[1] Shi C. Research on Deep Learning Algorithms for Predicting DNA-Binding Proteins Based on Sequence Information[C]//2024 IEEE 2nd International Conference on Electrical, Automation

and Computer Engineering (ICEACE). IEEE, 2024: 1566-1570.

[2] Yang J. Research on the Strategy of MedKGGPT Model in Improving the Interpretability and Security of Large Language Models in the Medical Field[J]. Academic Journal of Medicine & Health Sciences, 5(9): 40-45.

[3] Xu, Yue. "Research on Maiustream Web Database Development Techncl ogy." Journal of Computer Science and Artificial Intelligence 2.2 (2025): 29-32.

[4] Guo X. Research on systemic financial risk early warning based on integrated classification algorithm[C]//2024 IEEE 2nd International Conference on Electrical, Automation and Computer Engineering (ICEACE). IEEE, 2024: 1586-1591.

[5] Yang J. Research on the Application of Medical Text Matching Technology Combined with Twin Network and Knowledge Distillation in Online Consultation[J].

[6] Xu Y. Research on UAV Navigation System Based on Behavioral Programming[C]//2024 IEEE 7th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE). IEEE, 2024: 419-425.

[7] Chen, H., Wang, Z., & Han, A. (2024). Guiding Ultrasound Breast Tumor Classification with Human-Specified Regions of Interest: A Differentiable Class Activation Map Approach. In 2024 IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium (UFFC-JS) (pp. 1-4). IEEE.

[8] Liu, Boyang. "Research on Holographic Retrieval and Analysis System for Scientific Research Data Based on SSH Framework and Lucene Engine."

[9] Cao, Y., Cao, P., Chen, H., Kochendorfer, K. M., Trotter, A. B., Galanter, W. L., ... & Iyer, R. K. (2022). Predicting ICU admissions for hospitalized COVID-19 patients with a factor graph-based model. In Multimodal AI in healthcare: A paradigm shift in health intelligence (pp. 245-256). Cham: Springer International Publishing.

[10] Varatharajah, Y., Chen, H., Trotter, A., & Iyer, R. K. (2020). A Dynamic Human-in-the-loop Recommender System for Evidence-based Clinical Staging of COVID-19. In HealthRecSys@ RecSys (pp. 21-22).

[11] Varatharajah, Y., Chen, H., Trotter, A., & Iyer, R. K. (2020). A Dynamic Human-in-the-loop Recommender System for Evidence-based Clinical Staging of COVID-19. In HealthRecSys@ RecSys (pp. 21-22).

[12] Fan, Sunjia, et al. "Defense methods against multi-language and multi-intent LLM attacks." International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI 2024). Vol. 13403. SPIE, 2024.

[13] Liu, Boyang. "Study on the Frequency of Computer Language Use Based on Big Data Analysis." Academic Journal of Computing & Information Science 7.10 (2024).

[14] Nag P K , Bhagat A , Priya R V ,et al.Emotional Intelligence Through Artificial Intelligence : NLP and Deep Learning in the Analysis of Healthcare Texts[J]. 2024.

[15] Abdul K R S , Tilford T , Stoyanov S .Fine-grained food image classification and recipe extraction using a customized deep neural network and NLP[J].Computers in Biology and Medicine, 2024:175.

[16] Sadigov R , Yildirim E , Kocacinar F C C .Deep learning-based user experience evaluation in distance learning[J].Cluster computing, 2024, 27(1):443-455.

[17] Alshathry N I , Alghamdi M , Aldobaian A S ,et al.INTEGRATING OPTIMAL DEEP LEARNING WITH NATURAL LANGUAGE PROCESSING FOR ARABIC SPAM AND HAM TWEETS RECOGNITION[J].Fractals, 2024, 32(9/10).DOI:10.1142/S0218348X25400523.

*[18] Yang, **zhu. "Research on the Strategy of MedKGGPT Model in Improving the Interpretability and Security of Large Language Models in the Medical Field." Academic Journal of Medicine & Health Sciences 5.9: 40-45.*

[19] Liu, Boyang. "Data Analysis and Model Construction for Crew Fatigue Monitoring Based on Machine Learning Algorithms." optimization 8.5: 48-52.

[20] Liu, Yu. "Build an Audit Framework for Data Privacy Protection in Cloud Environment." Procedia Computer Science 247 (2024): 166-175.

[21] Zhao, Fengyi. "Risk Assessment Model and Empirical Study of in Vitro Diagnostic Reagent Project Based on Analytic Hierarchy Process." International Journal of New Developments in Engineering and Society 8.5 (2024), 76-82

[22] Yang J. Application of Multi-model Fusion Deep NLP System in Classification of Brain Tumor Follow-Up Image Reports[C]//The International Conference on Cyber Security Intelligence and Analytics. Cham: Springer Nature Switzerland, 2024: 380-390.

[23] Wang Y. Design and Implementation of a General Data Collection System Architecture Based on Relational Database Technology[C]//The International Conference on Cyber Security Intelligence and Analytics. Cham: Springer Nature Switzerland, 2024: 561-572.