


Ontology Attention Layer for Medical Named Entity Recognition

Yue Zha, Yuanzhi Ke , Xiao Hu and Caiquan Xiong

School of Computer Science, Hubei University of Technology, Wuhan 430068, China; zhayue4@gmail.com (Y.Z.); 1773529577@gmail.com (X.H.); xiongqcq@hbut.edu.cn (C.X.)

* Correspondence: keyuanzhi@hbut.edu.cn

Abstract: Named entity recognition (NER) is particularly challenging for medical texts due to the high domain specificity, abundance of technical terms, and sparsity of data in this field. In this work, we propose a novel attention layer, called the “ontology attention layer”, that enhances the NER performance of a language model for clinical text by utilizing an ontology consisting of conceptual classes related to the target entity set. The proposed layer computes the relevance between each input token and the classes in the ontology and then fuses the encoded token vectors and the class vectors to enhance the token vectors by explicit superior knowledge. In our experiments, we apply the proposed layer to various language models for an NER task based on a Chinese clinical dataset to evaluate the performance of the layer. We also investigate the influence of the granularity of the classes utilized in the ontology attention layer. The experimental results show that the proposed ontology attention layer improved F1 scores by 0.4% to 0.5%. The results suggest that the proposed method is an effective approach to improving the NER performance of existing language models for clinical datasets.

Keywords: named entity recognition; ontology; clinical text mining; attention mechanism



Citation: Zha, Y.; Ke, Y.; Hu, X.; Xiong, C. Ontology Attention Layer for Medical Named Entity Recognition. *Appl. Sci.* **2024**, *14*, 421. <https://doi.org/10.3390/app14010421>

Academic Editors: Jae-Hoon Kim and Kichun Lee

Received: 7 October 2023

Revised: 22 December 2023

Accepted: 26 December 2023

Published: 3 January 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Named entity recognition (NER) is particularly challenging in the medical domain due to its specialized and complex nature. Firstly, there are numerous domain-specific and highly sparse terms in the medical field. Traditional deep learning models generally perform NER based on available datasets and do not precisely adjust for medical domain-specific terms. Secondly, in the medical domain, the relationships between entities are more complex than in other domains, and traditional deep learning models require help to utilize this relational information effectively. Thirdly, medical named entities typically include diseases, drugs, surgeries, symptoms, etc., which can be expressed in various ways within the text, requiring more sophisticated techniques and methods for accurate identification.

In recent years, medical NER has gained increasing attention. For instance, Guillaume et al. [1] applied a BiLSTM-CRF model-based approach for medical named entity recognition, achieving good results. Yang et al. [2] proposed a medical NER method based on knowledge graphs, leveraging entity relationship information to improve entity recognition accuracy.

An ontology is a crucial component often developed prior to tackling an industrial NER task. This preparation proves valuable, particularly within the context of a conventional rule-based NER approach.

Our study highlights that an ontology can also significantly enhance the NER performance of a language model when incorporated through a well-designed attention layer. In this paper, we propose an attention model that incorporates a medical ontology. The model utilizes a medical ontology. It takes text and ontology as inputs and passes them through separate encoders. Using a multi-head attention mechanism, the model calculates and combines attention weights before feeding them into a feed-forward neural network for maximum likelihood estimation. A simple greedy sampling method is applied to generate

tags for NER based on the estimated probabilities. Experimental results show performance improvements across three baseline models.

The main contribution of this paper is the proposal of an approach that better captures specific domain knowledge to recognize medical entities by leveraging knowledge from the medical ontology. We also discuss the effects of different ontology design granularity and abstraction-type properties on NER tasks. Our proposed method is simple, effective, and especially useful in practice.

2. Related Works

2.1. Medical Named Entity Recognition

In the development of medical knowledge graphs, named entity recognition (NER) has received widespread attention as a critical component for constructing medical knowledge graphs. It aims to automatically detect the desired entities from given text.

Conventional general NER frameworks such as a bidirectional-LSTM [3] or a GPT [4] could be used for medical NER tasks. But medical NER has its own challenges, such as the lack of data and named entity normalization (NEN).

For the lack of data, some researches used few-shot or zero-shot learning techniques. Jiang et al. [5] employed a combination of a small amount of robust labeled data and a large amount of weak labeled data to train the model in stages and mitigate the problem of few-shot learning. Li et al. [6] introduced a novel approach that used learnable logical rules for weakly supervised NER. Li et al. [7] introduced a conditional hidden Markov model for multi-source weakly supervised NER. Aly et al. [8] leveraged the naturally occurring facts in the textual descriptions of various entity categories to complete a zero-shot NER task.

Other research has used external additional information. Wu et al. [9] fused the structural information of Chinese characters to improve the performance of Chinese NER. Wang et al. [10] improved the NER task using external context retrieval and cooperative learning. Wei et al. [11] proposed enhancing the knowledge on the basis of the pre-trained model, which improves the effect of NER, NLU, and NLG.

NEN involves standardizing different representations of the same medical entity, such as a disease or drug, and it has a substantial impact on NER performance as well. For NEN, some researchers utilize multi-task learning to address named entity normalization in medical NER [12]. Ji et al. [13] proposed a model based on state transition, which transforms end-to-end disease recognition and normalization tasks into action sequence prediction tasks.

Some research has focused on difficult entity names such as nested names, long names, or discontinuous names. Wang et al. [14] designed an objective function for training neural models to handle nested entity label sequences as suboptimal paths for nested NER tasks. Li et al. [15] developed a network for long names utilizing both segment-level information and word-level dependencies. Wang et al. [16] addressed the issue of discontinuous text in NER tasks by adopting a fragment graph approach. Li et al. [17] presented a span-based model capable of identifying overlapping and discontinuous entities, as well as determining whether entity relationships overlap.

2.2. Attention Mechanism

Early research on attention mechanism in neural networks used it with conventional neural networks such as recurrent neural networks or convolutional neural networks. Luong et al. [18] subsequently introduced global and local attention mechanisms. Yang et al. [19] integrated hierarchical thinking and introduced a hierarchical attention mechanism. Zhang et al. [20] proposed a top-down attention mechanism, making the neural network more targeted when learning attention. Gehring et al. [21] then proposed a multistep attention mechanism, using an attention mechanism at each decoder layer.

Subsequently, Vaswani et al. [22] proposed the transformer model, which relies entirely on the attention mechanism to represent the global dependencies between inputs and outputs, avoiding convolution and recurrence and promoting attention mechanisms.

Following that work, Eldete et al. [23] proposed a model called AttnSleep, which uses multi-head attention to capture temporal dependencies among extracted features. Xiong et al. [24] introduced a new regularization transfer learning framework called DELTA, which uses a feature map with attention for deep learning transfer and preserves the outer layer output of the target network.

Despite building upon these models, medical NER remains challenging due to the specialized nature, numerous rare words, and high sparsity of named entities in medical-related texts compared to general texts. Therefore, in this study, we propose incorporating the relevance between the text to be recognized and the concepts in the knowledge graph ontology as features in medical NER to improve its efficiency and accuracy.

3. Proposed Methods

3.1. Overall Architecture

The overall architecture of our proposed method for medical NER is shown in Figure 1. The input text and its ontology are encoded into vectors. Then, their output is input into a multi-head attention layer to compute cross-attention. Afterwards, the input text vector and the output of the attention layer are added, and the result is input into a feed-forward network with a softmax activation function to obtain the likelihood of NER tags.

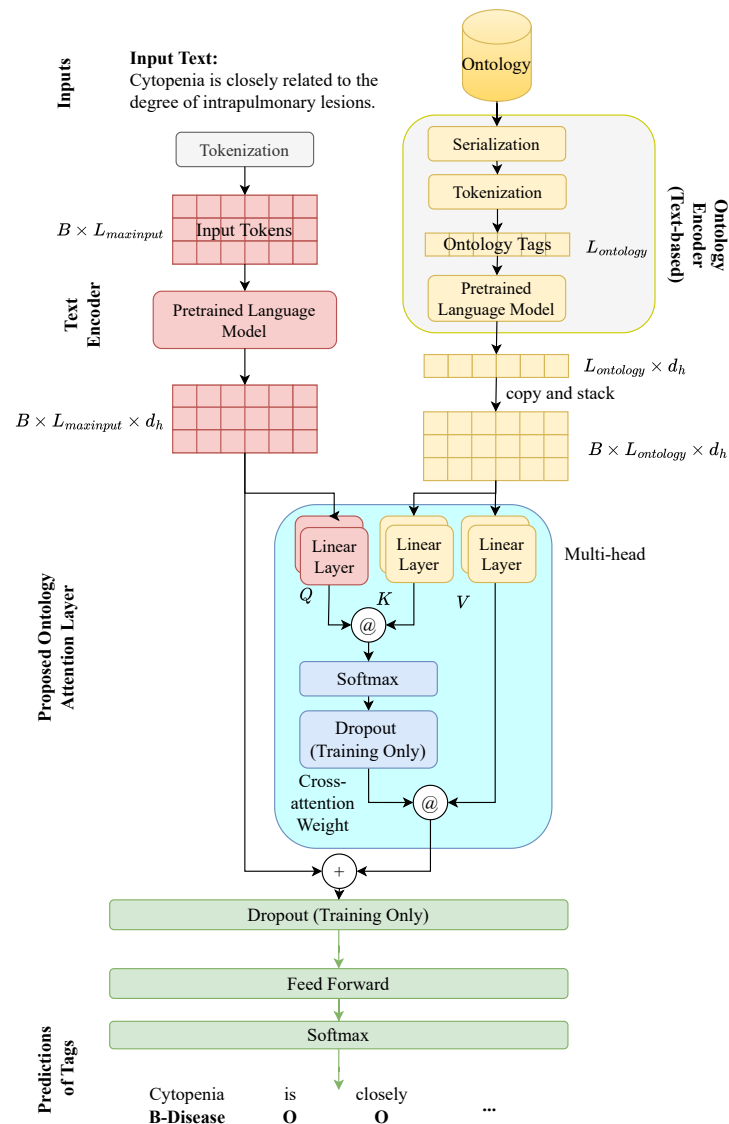


Figure 1. The architecture of our proposed model for medical NER. B refers to the batch size. $L_{maxinput}$ refers to the max length of an input text. We pad every input text into the same length and use a

mask to preserve the correctness of attention results. $L_{ontology}$ refers to the length of the serialized and tokenized ontology tags, which was fixed because the whole unchanged ontology was used in the experiment. d_h is the dimension of the hidden vectors. "@" refers to matrix multiplication operator. "+" refers to the element-wise add operator. The multi-head attention is described in Equations (1) and (2).

3.2. Text Encoder

The proposed method utilizes a BERT-based language model as the text encoder, such as BERT [25], BERT-WWM [26,27], and RoBERTa [28]. A typical BERT model employs a transformer to extract features, effectively leveraging the contextual information of the text, which is beneficial for entity extraction and classification in downstream tasks.

3.3. Ontology Encoder

For the ease of computing the co-relevance between the concepts in the ontology and the tokens in the input, we serialize the ontology into texts and then encode it using a pre-trained language model, rather than employing popular graph encoders such as TransE [29], ConvE [30], GCN [31], CompGCN [32], etc. This encoding approach facilitates the learning of the mapping between the text and ontology concepts, as both the input text and the ontology can be encoded by the same model, ensuring natural alignment. Furthermore, because the ontology in our experiments contained many properties but few entity-to-entity edges, the methods based on triplets or link prediction perform less powerfully than in other tasks.

3.3.1. Serialization and Tokenization

The primary challenge is how to serialize the ontology to preserve relationship information between the concepts in the resulting sequence. We hypothesize that since a BERT-based language model takes the position of each input token into account by exploiting the position vectors [25], if the position of each class tag or property in the ontology matches the conditions as follows, the language model is able to learn the corresponding relationship information by fine-tuning:

1. The relative position of two tokens is able to show the relationship between the corresponding class tags or properties.
2. The positions are fixed in every training epoch.

Treating the ontology as a graph, the relationship of the classes and their properties can be represented by their topological location. Thus, we use alphabetical pre-order depth-first traversal (DFS) to traverse the ontology and output and concatenate the tags of each visited node. In our preliminary study, we found that using pre-order traversal is slightly better than post-order traversal or the combination of both of them. The comparison of pre-order traversal, post-order traversal and the combination of them is provided in Appendix A.

For the small example in Figure 2, which shows a part of the ontology in our experiment (to be described in Section 4.3), the subtrees are sorted alphabetically from left to right. In the example, "i" is before "r", so the subtree of "disease" is to the left of that of "drug". We start from the root, and then the "disease" subtree. For the root, we output nothing. For the "disease" subtree, at first, we obtain the tag of the node "disease". Let us assume that all the tags are equal to the English node name here for simplicity; thus, the sequence is now "disease". Then we go to the left-most subtree of the "disease" subtree, which has only one node, "cause". We obtain its tag and update the sequence to "disease cause". Similarly, we obtain "disease cause clinical manifestations diagnostic criteria disease name" for the "disease subtree". Then we process the "drug" subtree and update the sequence to "disease cause clinical manifestations diagnostic criteria disease name drug contraindication dosage drug component drug name indications".

In this way, each node will have a meaningful position that implicitly corresponds to its topological position in the ontology, which is also fixed in every training epoch. We can see that “clinical manifestations”, which are properties of a “disease”, are closer to “disease” than to “drug”. Furthermore, the class “disease” is closer to its own “disease name” than the “drug name” property of the class “drug”.

Formally, the ontology is considered as a directed tree, where the classes and their properties are treated as nodes, and the relationships, e.g., “is_property”, are treated as unweighted edges. Denote the tree as $G = (V, E)$, where V and E refer to the nodes and edges, respectively. Depth-first traversal is employed, utilizing alphabetically ordered depth-first search to traverse the ontology and generate a sequence of node tags $S = (s_1, s_2, \dots, s_i, \dots, s_n)$, where i is the position and n is the number of nodes in the ontology. Note that the relationship tags are dropped here because of the limitation of the proposed method.

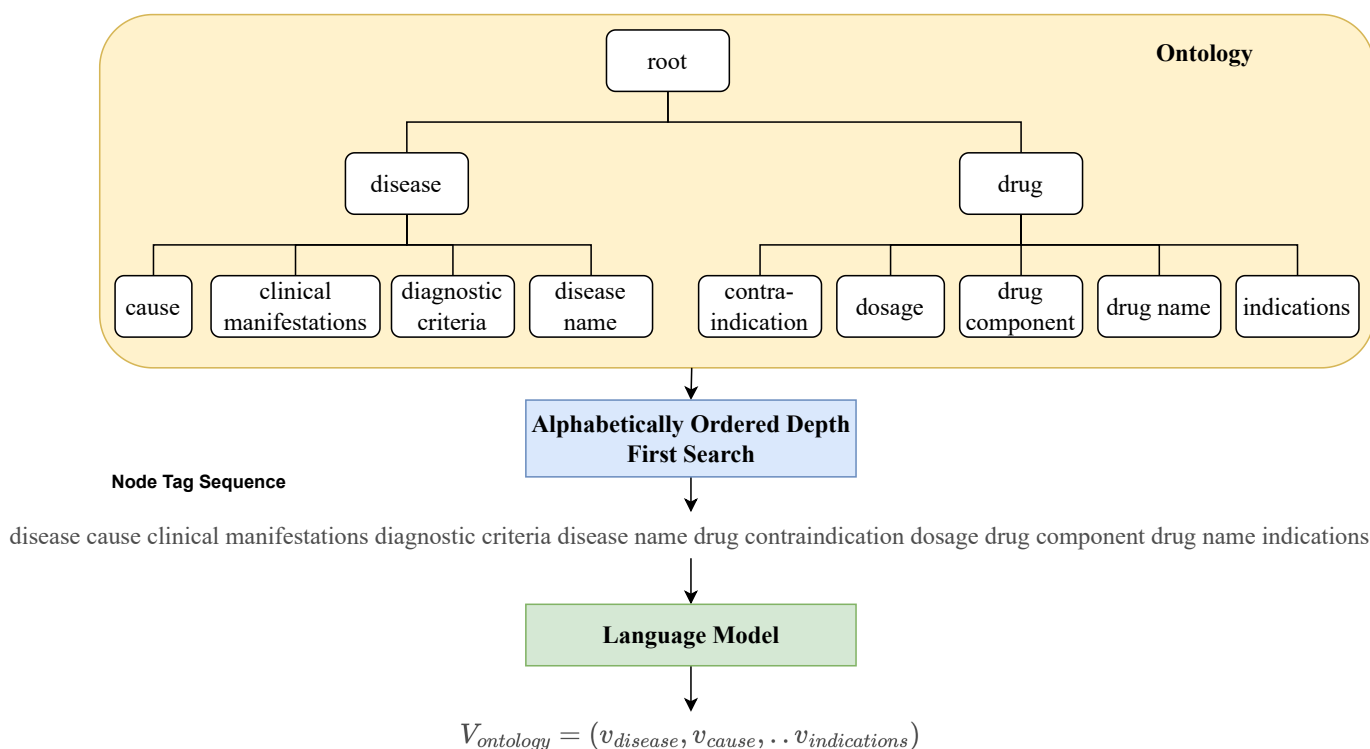


Figure 2. The proposed method to encode the ontology with a small example.

Another issue is how to separate the node in the resulting sequence. It depends on the target language. In the experiment, the tags of nodes were separated by a space, because most experiments were performed on a Chinese dataset where space was rarely used.

3.3.2. Encoding the Serialized Ontology

After we serialize the ontology into a text token sequence, this sequence is then inputted into a BERT-based language model for encoding. For each node tag s_i , when it is inputted into a BERT language model in the next phase (introduced in Section 3.3.2), a position embedding PE_i is generated and inputted into the language model along with it. The method to generate the position embedding depends on the language model. A typical method is to use sine and cosine functions, proposed by Vaswani et al. [22]. Additionally, the self-attention layers in a BERT model enable each node tag to attend to all positions. Thus, the encoder is capable of extracting the meaning of the positions.

We use the same model used for the input text so that the embedding spaces are aligned. Formally, we take the hidden states of the last layer $V_{ot} = v_1^o, v_2^o, \dots, v_n^o$ as a vector representation of the ontology. Because the outputs of a BERT-based language model

implicitly contain the context information, each vector v_i^o both represents the meaning of the tag of the node and the topological position of the node in the original ontology.

3.4. Ontology Attention

The ontology attention layer utilizes attention mechanisms to capture the relevance between the concepts in the ontology and input tokens.

Denote the encoded input text as $V_t = \{v_1, v_2, \dots, v_i, \dots, v_{L_{maxinput}}\}$, where v_i is the vector corresponding to the i_{th} token. We employ multi-head attention. For $head_i$, we initialize parameters W_i^Q, W_i^K, W_i^V , and obtain the query, key, and value vectors for multi-head attention as follows:

$$Q = V_t W_i^Q, K = V_{ot} W_i^K, V = V_{ot} W_i^V \quad (1)$$

Then, let all the vectors be of dimension d_h . Then, we have

$$\begin{aligned} head_i &= \text{Attention}(Q, K, V) \\ &= \text{Softmax}\left(\frac{QK^T}{\sqrt{d_h}}V\right). \end{aligned} \quad (2)$$

We concatenate the results from all the heads, apply another linear projection, and add it to V_t , as follows:

$$V_f = V_t + \text{Concat}(head_1, \dots, head_i, \dots, head_h)W, \quad (3)$$

where W and V_f are the parameters of the linear projection and the output of this ontology attention layer, respectively.

3.5. Feed-Forward Layer and Output

Finally, the feed-forward layer applies a linear transformation to the output of the ontology attention layer, followed by a softmax activation function, to produce the likelihood distribution over the predefined tags for the NER task, as follows:

$$P = \text{Softmax}(V_f W^P + b^P), \quad (4)$$

where P is the predicted distribution of the NER tags. W^P and b^P are trainable parameters.

4. Experiments

4.1. Dataset

We conducted experiments on two datasets: the CMeEE dataset from the Chinese Biomedical Language Understanding Evaluation (CBLUE) [33] and the NCBI disease dataset [34].

The CMeEE dataset task is a multi-class NER task, requiring the recognition of “疾病” (disease), “临床表现” (symptom), “医疗程序” (process), “医疗设备” (equipment), “药物” (drug), “医学检验项目” (test item), “身体” (anatomical location), “科室” (hospital department), and “微生物类” (microbe) in clinical texts. The texts are collected from clinical trials, electronic health records, medical forums, textbooks, and search engine logs. The ground truth is annotated by three to five domain experts.

This NCBI disease dataset contains the disease name and concept annotations of the NCBI disease corpus. The corpus comprises 793 abstracts in PubMed (<https://pubmed.ncbi.nlm.nih.gov/>, accessed on 6 October 2023).

4.2. Labeling Policy

For both the datasets, we followed their own original labeling policy.

For CMeEE, the BIO format is used for data annotation. Each character was labeled “B-X”, “I-X”, or “O”. Specifically, “B-X” indicates that the character belongs to the X type and is the beginning of a sequence. “I-X” indicates that the character belongs to the X type and is in the middle of a sequence. “O” indicates that the character does not belong to any type. The corpus is annotated according to three categories: disease name, specific symptoms, and causes. For example, for the term “cell reduction”, which is a symptom, the initial character “c” is labeled as “B-Symptom” and the other characters from “e” to “n” are labeled “I-Symptom”.

For the NCBI disease dataset, the start of a mention of diseases is tagged by “1”. The subsequent disease tokens are tagged by “2”. The tag “0” indicates no disease mentioned. Furthermore, as the original tokens are not in the subword format for the language models in the experiment, we reprocessed the data to fit the inputs to the models. For example, before we input the mention of disease “adenomatous polyposis coli tumour”, tagged as “1, 2, 2”, to a BERT model, we reprocessed the term into “ad, ##en, ##om, ##at, ##ous, po, ##ly, ##po, ##sis, co, ##li, t, ##um, ##our”, and the tags into “1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2”, respectively. The first five “1”s are the tags to “adenomatous”, which was divided into subwords “ad, ##en, ##om, ##at, ##ous”. The “2”s are for the other words in this mention of a disease.

4.3. Construction of the Ontology

In this study, a hierarchical structure was adopted to construct a tree-like system as a medical ontology. The detailed design is as shown in Table 1. We designed this ontology for the CBLUE dataset. We extracted seven classes: “疾病” (disease), “临床表现” (symptom), “医疗设备” (equipment), “药物” (drug), “身体” (anatomical location), “医院部门” (hospital department), and “微生物” (microbe) from the dataset. Note that these classes were not the same as the targets of the CMeEE dataset, although there was some overlap. We also used the translated version of this ontology for the NCBI disease dataset. Even though the NCBI disease task only requires recognizing mentions of diseases, we believe that including additional classes could help the models distinguish diseases from potentially confusing words.

The properties were classified into three types based on their characteristics: entity-type properties, attribute-type properties, and relationship-type properties.

The entity-type properties are used to model and describe the nouns of the ontology in more detail. A specific description of the noun ontology can be obtained at the entity-type property. For example, in Figure 1, the entity-type property would describe the unique descriptions of the noun “disease”, such as the name and cause.

The attribute-type properties describe the characteristics of the noun ontology. For example, in Figure 1, the attribute-type property of the disease includes the clinical manifestations and diagnostic criteria of the disease.

The relationship-type properties represent the relationships between the entity and attribute-type properties, describing various behaviors between various entities. For example, a treatment method can treat specific diseases, and certain diseases can cause certain symptoms.

We used the tags of the ontology classes, their entity-type properties, and their attribute-type properties for the NER tasks in the experiments because none of the datasets in our experiments required relationship extraction.

Table 1. Detailed design of the ontology in the experiment for the Chinese Biomedical Language Understanding Evaluation (CBLUE) dataset. The ontology is presented in Chinese, and we offer the corresponding English translation here.

Ontology Class	Entity-Type Property	Attribute-Type Property	Relationship-Type Property
疾病 disease	疾病名称 disease name 疾病病因 cause	临床表现 clinical manifestations 诊断标准 diagnostic criteria	诊断 diagnosis 治疗 treat 预防 prophylaxis
临床表现 symptom	症状 symptom signs	表现的性质 symptom property 表现的时序 symptom timing	表现与疾病的关系 diagnosis and treatment of the relationship between manifestations and disease
医疗设备 medical equipment	设备名称 equipment name 设备功能 equipment function	适合使用的症状 symptoms suitable for use 操作要求 operational requirements	设备使用 device use 设备操作 device operation
身体 anatomical location	部位名称 part name 解剖结构 anatomy	部位性质 anatomical property	部位与疾病的临床表现 location and clinical manifestations of the disease
医院部门 hospital department	部门名称 department name 部门职能 departmental functions	服务对象 service object 作用范围 range of action	部门与疾病治疗的关系 the relationship between the department and the treatment of diseases
微生物 microbe	微生物名称 microbial name 微生物工作职能 microbiology functions	生物学特征 biological characteristics 致病性 pathogenicity	微生物与药物之间的关系 the relationship between microorganisms and drugs
药物 drug	药物名称 drug name 药物成分 drug component	适应症 indications 禁忌症 contraindication 剂量 dosage	药物治疗疗程与疾病的关系 the relationship between the course of drug therapy and the disease

The serialization result of the ontology described above can be found in Table A2 in Appendix A.

4.4. Hyperparameters

The hyperparameters were tuned by grid search on the vanilla BERT model in the CMeEE task. Because the purpose of the experiments was to validate the effectiveness and functional feasibility of the proposed method, we used the same hyperparameters for all the groups to reduce the effects from the hyperparameters. A detailed description is provided in Table 2.

Table 2. Hyperparameter settings. The hyperparameters are tuned by grid search on the BERT model.

Hyperparameter	Value
Training Batch Size	16
Evaluation Batch Size	64
Learning Rate	5×10^{-5}
Epochs	10

5. Experimental Results and Discussion

5.1. Performance over Different Pretrained Language Models as the Encoders

We validated the effectiveness of the proposed method using different pretrained language models as the encoders. We tested our method with BERT, BERT-WWM, and RoBERTa. In detail, for the CMeEE dataset, which is in Chinese, we used “bert-base-chinese” (<https://huggingface.co/bert-base-chinese>, accessed on 6 October 2023), “chinese-bert-wwm-ext” (<https://huggingface.co/hfl/chinese-bert-wwm-ext>, accessed on 6 October 2023) [26,27], and “chinese-roberta-wwm-ext” (<https://huggingface.co/hfl/chinese-roberta-wwm-ext-large>, accessed on 6 October 2023). For NCBI disease, which is in English, we used “bert-base-uncased” (<https://huggingface.co/bert-base-uncased>, accessed on 6 October 2023) [25] and roberta-base (<https://huggingface.co/roberta-base>, accessed on 6 October 2023) [28]. Because BERT-WWM is not for English, we did not use a BERT-WWM model for the NCBI disease dataset.

To assess the impact of the proposed algorithm on model recognition efficiency, all models were trained and tested on the same dataset using identical hyperparameters. The hyperparameters are tuned by grid search on the BERT model. The training and testing data for the medical corpus were also extracted from the same dataset.

We compare the test results achieved by our proposed method, those achieved by the language models with the conventional self-attention layer, and those achieved by the vanilla language models. We used micro F1 score for CMeEE and macro F1 for NCBI, respectively, as the metric. The results in CMeEE and NCBI are shown in Figures 3 and 4, respectively.

For CMeEE, the proposed ontology attention improved efficiency from 0.1% to 0.5% compared to the conventional self-attention layer. Furthermore, compared to the vanilla models, the proposed method improved efficiency from 0.1% to 1.5%.

For NCBI disease, the proposed method consistently outperformed self-attention. However, it failed to outperform the groups that used the plain base model. This outcome may be attributed to the fact that the ontology was designed for the CBLUE dataset in Chinese and was translated into English for the NCBI disease task by non-medical professionals. A more professional translation may yield better results.

Table 3 shows some typical outputs for the NCBI disease dataset. All groups encountered challenges in the recognition of certain difficult named entities, such as “T-PLL” in data #11. In texts without disease mentions, the plain base model and the self-attention group tended to generate more false positives, while the proposed method showed a reduction in false positives, as seen in data #18 and #81. Notably, the proposed method demonstrated superior efficacy in the identification of long complex named entities, exemplified in data #94, #138, and #155.

The results above indicate that the proposed ontology attention mechanism outperforms existing self-attention mechanisms in recognizing challenging medical named entities and reducing false positives. The integration of ontology enhances the model’s comprehension of medical terms. Consequently, the model can better capture context and semantic information related to medical entities, recognizing more complex medical terms and mislabeling fewer non-medical terms. This leads to an improvement in the performance of medical named entity recognition (NER).

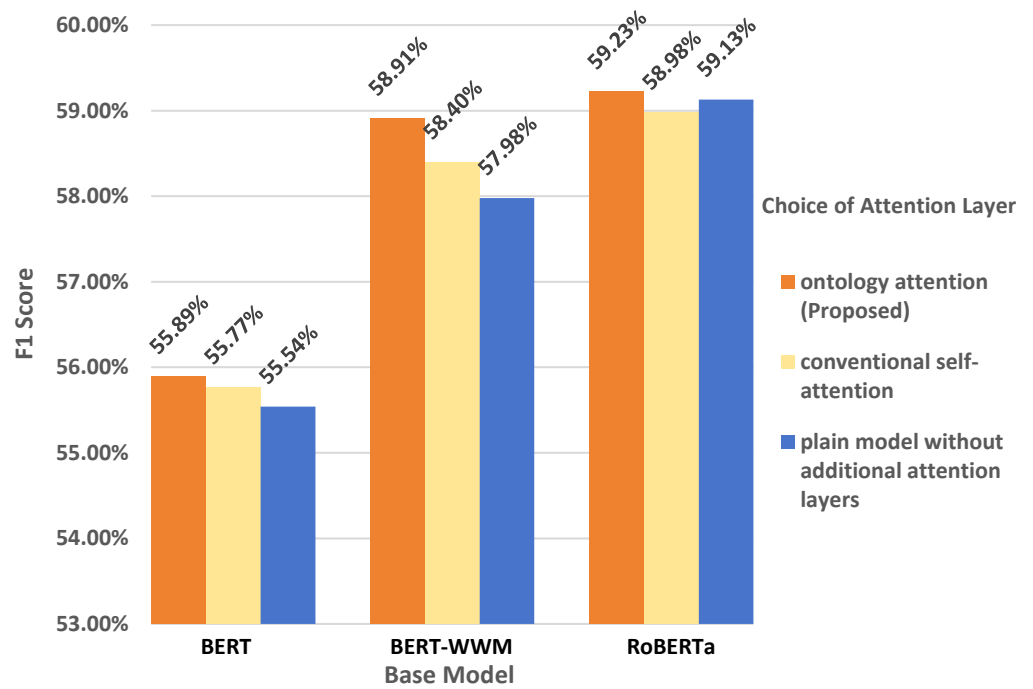


Figure 3. Comparison of the test results in CMeEE (Chinese dataset) task achieved by our proposed method (in orange), those by the language models with the conventional self-attention layer (in yellow), and those by the vanilla language models (in blue). The results with BERT, BERT-WWM, and RoBERTa used as the base language model are provided.

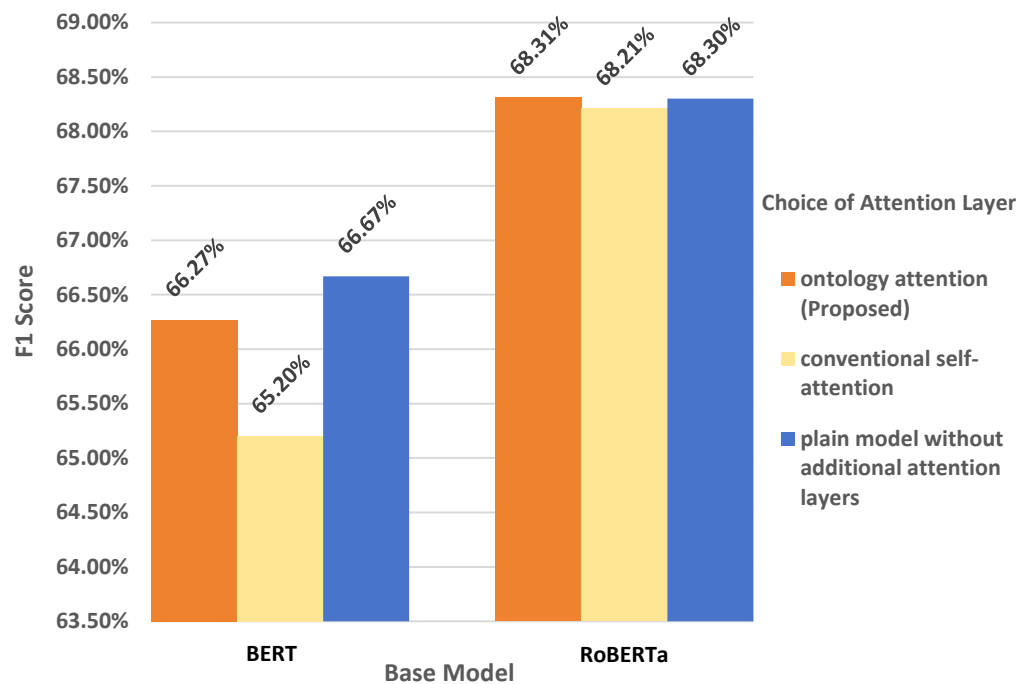


Figure 4. Comparison of the test results in NCBI disease (English dataset) task achieved by our proposed method (in orange), those by the language models with the conventional self-attention layer (in yellow), and those by the vanilla language models (in blue). The results with BERT and RoBERTa used as the base language model are provided. Macro F1 Scores are reported.

Table 3. Output examples for the NCBI disease dataset when BERT-base was used as the base model. “ID” refers to the data ID in the NCBI disease dataset. In the “Text” column, we display the input texts. The bold words in the “Text” column represent the ground truth. If no words are bold, it indicates that there was no disease mention in it. “Plain Base Model” refers to the group that only used BERT-base. “Proposed Method” refers to the group that utilized the proposed method. “Self-attention” refers to the group that incorporated an additional self-attention layer to the base model for NER.

ID	Text	Plain Base Model	w/ Proposed Method	w/ Self-Attention
11	As constitutional DNA was not available, a putative hereditary predisposition to T-PLL will require further investigation.	(None)	T-	T-
18	Neither the content nor the activity of Na ⁺ /K ⁺ ATPase and sarcoplasmic reticulum Ca ²⁺ -ATPase are affected by DMPK absence.	(None)	(None)	DMPK absence
81	Saamis (2%) and Mordvinians (1.8%) had significantly lower frequencies of the Tyr allele.	Saamis	(None)	(None)
94	Numerous cytogenetic and allelotype studies have reported frequent loss of heterozygosity on chromosomal arm 10q in sporadic prostate cancer .	prostate cancer	sporadic prostate cancer	prostate cancer
138	We conclude that paternal transmission of congenital DM is rare and preferentially occurs with onset of DM past 30 years in the father.	DM, DM	congenital DM, DM	DM, DM
155	Mutations in the SMAD4/DPC4 tumor suppressor gene, a key signal transducer in most TGFbeta-related pathways, are involved in 50 % of pancreatic cancers .	pancreatic cancers	tumor, pancreatic cancers	tumor, pancreatic cancers

5.2. Discussion on the Granularity of the Ontology

After the last round of experiments, we found that the proposed attention mechanism worked better than the self-attention mechanism. Considering that our ontological attention mechanism reduces sparsity by subdividing text, the more words it contains, the greater the sparsity. In this round of experiments, the original structure of the ontology attention mechanism was not changed. Instead, the input body was replaced with ontology of different granularity. Because the input body was different, the mask calculated during the calculation was also different. As a result, the calculated ontological similarities are also inconsistent, resulting in different F1 scores. The main points of this study are based on medical ontology. We defined three classes of grain properties: coarse, medium, and fine. Details regarding the design of this experiment are provided in Table 4, and the results of the comparison are shown in Figure 5.

Table 4. Setup of groups for the experiment on the influence of the granularity type properties.

Group	Contained Property Types
Coarse-grained	class name only without properties
Medium-grained	entity-type properties
Fine-grained	entity-type properties, attribute-type properties

According to the experimental results, the NER performance of the proposed model with a medium-grained ontology is approximately 0.2% higher than that with a coarse-grained ontology. When the granularity increases further, there is a decrease in the results. A possible reason of the latter result is that the distance of words in semantic space becomes far when the granularity of the ontology increases too much, resulting in increased sparsity, although appropriately increasing the granularity can enhance the semantic richness, resulting in better representation of domain knowledge.

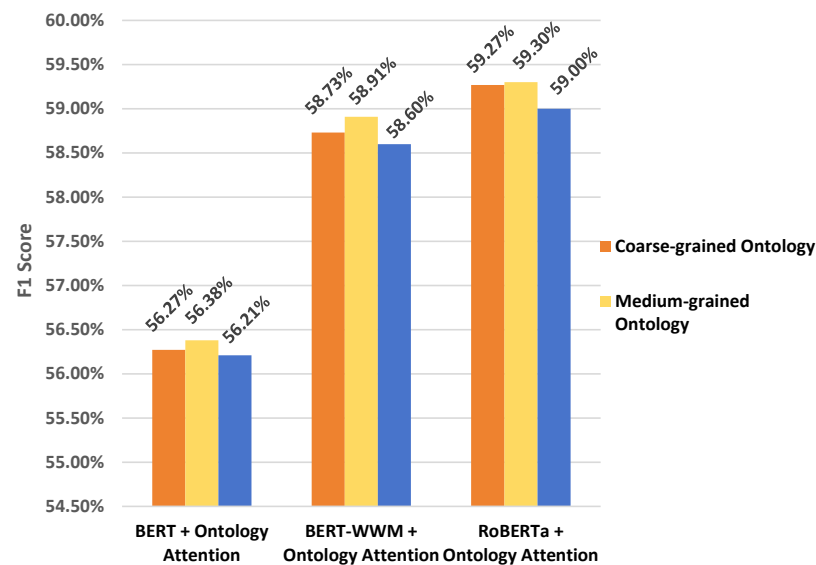


Figure 5. Results of the proposed method with different ontology setups. The values are micro F1 scores for the NER task on the CMEE dataset. Regardless of what language model is used as the encoder, the medium-grained ontology performs best.

5.3. Ablation Study on Effects of Different Property Types

Besides the granularity, different property types also contribute differently to the performance because they are parallel to ontology words and have different emphases when describing ontology words. In this round of experiments, we studied the effects by the attribute-type properties and entity-type properties. To achieve this, we modified the medium-grained group described in Section 5.2 to contain either only entity-type properties or only attribute-type properties, and then compared the NER performance of the proposed model with the two modified ontologies. The results are shown in Figure 6.

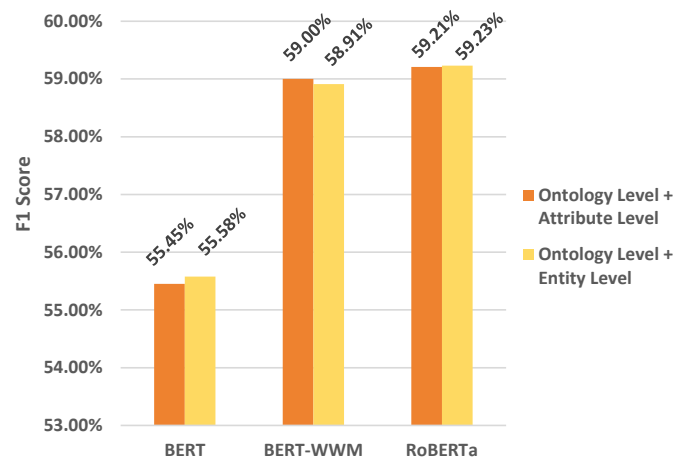


Figure 6. Results of the ablation study on ontology property types. The values are micro F1 scores for the NER task on the CMEE dataset.

This round of experiments confirmed that the two layers have parallel relationships in the ontology, but have different descriptive properties and manifestations in the ontology. Entity-type attributes provide only a specific description of the entity, while attribute-type attributes include ontology properties and capabilities, which describe ontology information for different dimensions. This ultimately leads to inconsistencies in the calculations and results between the two layers. Both are descriptions of the corresponding ontology information, but there are slight differences in the results due to differences in emphasis.

5.4. Discussion on the Number of Ontology Classes

In the experiments introduced in Section 5.2, the refinement of granularity improved the model to a certain extent. Meanwhile, the ontology also had an impact on the model effect. In this experiment, we investigated the effect of adding additional ontology-type attributes to the same type attribute of the ontology. The ontology was expanded from seven classes to nine classes by adding “medical procedures” and “medical testing items” and their corresponding entity-type properties, attribute-type properties, and relationship-type properties, as shown in Table 5. The results are as shown in Table 6.

Table 5. The two new classes added to the experimental ontology to investigate the impact of the number of classes.

Ontology Class	Entity-Type Property	Attribute-Type Property	Relationship-Type Property
医疗程序 medical process	医务人员 personnel 检验样本 sample	医疗程序的执行 execution of medical process 医学检验项目的目的 the purpose of the test	医疗程序的操作步骤 the process of performing medical procedure 医疗程序的选择 choice of medical procedure
医疗检验项目 test items	医疗设备 equipment 检测设备	操作技术 operational technique 医学检验项目的指标 indicator	检验流程 inspection process 质量控制 quality control

Table 6. A comparison of NER performance (micro F1 scores) in the CMEE dataset of the proposed method using ontology of 7 or 9 classes under different base language models and different granularity settings.

Settings	Coarse-Grained	Medium-Grained	Fine-Grained
w/ BERT & original ontology (7 classes)	56.27%	56.41%	56.21%
w/ BERT & extended ontology (9 classes)	56.35%	56.20%	56.21%
w/ BERT-WWM & original ontology (7 classes)	58.73%	58.91%	58.60%
w/ BERT-WWM & extended ontology (9 classes)	58.78%	58.71%	58.75%
w/ RoBERTa & original ontology (7 classes)	59.27%	59.23%	59.00%
w/ RoBERTa & extended ontology (9 classes)	59.18%	59.15%	59.15%

For the fine-grained setting in which both entity-type properties and attribute-type properties were involved, adding two ontology classes consistently improved the NER performance of the proposed model. For the coarse-grained setting, it also improved most variations of the proposed model. The results suggest that adding two additional ontology classes allows for more comprehensive descriptions within the ontology, which can lead to better NER performance in some cases.

For the medium-grained setting in which only entity-type properties were involved, the extending ontology classes consistently harmed the NER performance of the proposed model. The most probable reason was that the entity-type properties of these two new classes were too similar to each other. The results suggest that the classes in the ontology for the proposed model should be as unique as possible, which can be critical for the practical use of the proposed method.

5.5. Discussion on the Usage of the Ontology Class Name

In the experiments introduced above, we used the ontology class names with corresponding properties. As introduced in Section 3.3, the encoding policy for the ontology is based on depth-first search. Thus, the class name is to be read first, followed by the properties. To study that whether the properties themselves are descriptive enough for the proposed model, in this round of experiments, we directly eliminated class name before the ontology was encoded to expose the properties without the class name. This was

performed to investigate the changes in the micro F1 scores of the proposed model before and after removing the class names in the ontology encoding step. The experimental results are shown in Table 7.

Table 7. Effects of the usage of the ontology class name on the proposed model under different base language models and different granularity settings. Micro F1 scores are reported.

Settings	Medium-Grained	Fine-Grained
w/ BERT and class name	56.41%	56.21%
w/ BERT, w/o class name	56.14%	56.57%
w/ BERT-WWM and class name	58.91%	58.60%
w/ BERT-WWM, w/o class name	58.53%	58.92%
w/ RoBERTa and class name	59.23%	59.00%
w/ RoBERTa, w/o class name	59.38%	59.19%

If the medium-grained setting was used, the NER performance of the proposed model decreased in most groups. However, if the fine-grained setting was used, the NER performance of the proposed model consistently increased.

For the medium-grained setting in which only part of the properties (“entity-type properties” as defined in Section 4.3) was used, the information of the properties was not enough for the proposed model. Therefore, in this case, the class name was necessary.

For the fine-grained setting in which both entity-type properties and attribute-type properties were used, the information of the properties was enough. In this case, the redundant information brought by the class name was not necessary and relatively harmful.

6. Conclusions

In this study, we proposed a novel attention layer called the “ontology attention layer” to improve NER performance for medical texts by incorporating ontology. It enhances the NER performance of a BERT-based language model for clinical text by leveraging an ontology to improve the recognition of domain-specific entities. In the experiments, the proposed method enhanced three different BERT-based language models’ ability to accurately identify and classify entities, resulting in an improved F1 score for the medical NER task when utilizing a tailored ontology.

We further studied the effects to the performance by the design details of the ontology, which were also studied in several experiments. The results suggest that to maximize the performance of the proposed model, the granularity of the ontology should be tailored, and the name of the properties of different classes should be as distinct as possible.

Our research findings highlight that integrating an ontology can significantly improve the NER performance of a language model when implemented through a well-designed attention layer. This suggests that introducing the ontology reasoning model in entity recognition represents a practical approach to enhance the performance of BERT-based NER models. In the practical construction of a knowledge base, an ontology is typically crafted prior to processing unstructured corpora. In such scenarios, our proposed method would be valuable during the NER phase of corpus processing, requiring little additional manual effort.

Given that the design of the ontology significantly influences the effectiveness of the proposed method, our primary focus in future work is to explore an automated method for tailoring an ontology to suit the proposed approach. An ontology contains connections between entity names and their alternative representations may also be helpful but it is out of the scope of this work. Additionally, we aim to investigate similar approaches on a larger and more complex ontology. The combination of text-based embedding and graph-based embedding would prove beneficial in such scenarios, with the challenge lying in developing an effective method to integrate them seamlessly. Furthermore, the proposed

method may find utility in other domain-specific NLP tasks, such as relation extraction or question answering. Our plan is to delve into applications for these tasks in the future.

Author Contributions: Conceptualization, Y.K.; Methodology, Y.Z., X.H. and Y.K.; Software, Y.Z. and Y.K.; Validation, Y.K.; Formal analysis, Y.Z.; Investigation, Y.Z. and X.H.; Resources, C.X.; Data curation, Y.K.; Writing—original draft preparation, Y.Z.; Writing—review and editing, Y.K. and C.X.; Visualization, Y.Z. and Y.K.; Supervision, C.X. and Y.K.; Project administration, C.X.; Funding acquisition, X.H. and C.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by National College Students Innovation and Entrepreneurship Training Program, China (grant number: 202210500030).

Informed Consent Statement: Not applicable.

Data Availability Statement: Code is available at https://github.com/yuanzhiKe/NER_KG_Att_2023, accessed on 25 December 2023. Related data can be obtained from the following sites: CMeEE: <https://github.com/CBLUEbenchmark/CBLUE>, accessed on 25 December 2023, NCBI disease: https://huggingface.co/datasets/ncbi_disease, accessed on 25 December 2023.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Study on Post-Order Traversal and the Combination of Pre-Order and Post-Order Traversal

Table A1 shows the results of our study on the choice among pre-order DFS traversal, post-order DFS traversal, and the combination of them for the proposed method. All of the groups used the BERT-based model as the base language model. The serialization results are as shown in Table A2. The macro F1 scores on the NCBI disease dataset were used as the metrics to evaluate the effectiveness.

In the experiments, the pre-order DFS traversal performed slightly better than post-order DFS. We believe it is due to the fact that the serialization results by the pre-order DFS are closer to the natural order of corresponding words in a natural language. Furthermore, the combination of the sequences from two different orders adversely affected the effectiveness of the proposed method. It is because that such a combination leads to inconsistencies in the relative positions between two nodes in the encoded sequence.

Table A1. Comparison of pre-order DFS traversal, post-order DFS traversal, and the combination of them for the proposed method. In this experiment, the base language model was BERT-base. “Bi-order DFS” here is the group that used the concatenated two sequences generated by pre-order traversal and post-order traversal, respectively. Macro F1 scores are reported.

Dataset	Pre-Order DFS	Post-Order DFS	Bi-Order DFS
NCBI disease	66.27%	65.80%	64.90%

Table A2. Serialization results of pre-order DFS traversal and post-order DFS traversal for ontology described in Table 1 in the experiments.

Traversal Method	Serialization Result
Pre-order DFS	anatomical location anatomical property anatomy part name disease cause clinical manifestations diagnostic criteria disease name drug contraindication dosage drug component drug name indications hospital department department name departmental functions range of action service object medical equipment equipment name operational requirements symptoms suitable for use microbe biological characteristics microbial name microbiology functions pathogenicity symptom symptom property symptom signs symptom timing

Table A2. Cont.

Traversal Method	Serialization Result
Post-order DFS	disease name cause clinical manifestations diagnostic criteria disease symptom signs symptom property symptom timing symptom equipment name symptoms suitable for use operational requirements medical equipment part name anatomy anatomical property anatomical location department name departmental functions service object range of action hospital department microbial name microbiology functions biological characteristics pathogenicity microbe drug name drug component indications contraindication dosage drug

References

- Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural Architectures for Named Entity Recognition. In Proceedings of the North American Chapter of the Association for Computational Linguistics, San Diego, CA, USA, 12–17 June 2016. [\[CrossRef\]](#)
- Yang, T.; He, Y.; Yang, N. Named Entity Recognition of Medical Text Based on the Deep Neural Network. *J. Healthc. Eng.* **2022**, *2022*, 3990563. [\[CrossRef\]](#) [\[PubMed\]](#)
- Li, L.; Hou, L. Named Entity Recognition in Chinese Electronic Medical Records Based on the Model of Bidirectional Long Short-Term Memory with a Conditional Random Field Layer. *Stud. Health Technol. Inform.* **2019**, *264*, 1524–1525. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wang, S.; Sun, X.; Li, X.; Ouyang, R.; Wu, F.; Zhang, T.; Li, J.; Wang, G. GPT-NER: Named Entity Recognition via Large Language Models. *arXiv* **2023**, arXiv:2304.10428. [\[CrossRef\]](#)
- Jiang, H.; Zhang, D.; Cao, T.; Yin, B.; Zhao, T. Named Entity Recognition with Small Strongly Labeled and Large Weakly Labeled Data. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 22–27 May 2021. [\[CrossRef\]](#)
- Li, J.; Ding, H.; Shang, J.; McAuley, J.; Feng, Z. Weakly Supervised Named Entity Tagging with Learnable Logical Rules. *arXiv* **2021**, arXiv:2107.02282. [\[CrossRef\]](#)
- Li, Y.; Shetty, P.; Liu, L.; Zhang, C.; Song, L. BERTifying the Hidden Markov Model for Multi-Source Weakly Supervised Named Entity Recognition. *arXiv* **2021**, arXiv:2105.12848. [\[CrossRef\]](#)
- Aly, R.; Vlachos, A.; McDonald, R. Leveraging Type Descriptions for Zero-shot Named Entity Recognition and Classification. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 22–27 May 2021.
- Wu, S.; Song, X.; Feng, Z. MECT: Multi-Metadata Embedding based Cross-Transformer for Chinese Named Entity Recognition. *arXiv* **2021**, arXiv:2107.05418. [\[CrossRef\]](#)
- Wang, X.; Jiang, Y.; Bach, N.; Wang, T.; Huang, Z.; Huang, F.; Tu, K. Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 22–27 May 2021. [\[CrossRef\]](#)
- Wei, X.; Wang, S.; Zhang, D.; Bhatia, P.; Arnold, A. Knowledge Enhanced Pretrained Language Models: A Comprehensive Survey. *arXiv* **2021**, arXiv:2110.08455. [\[CrossRef\]](#)
- Zhou, B.; Cai, X.; Zhang, Y.; Yuan, X. An End-to-End Progressive Multi-Task Learning Framework for Medical Named Entity Recognition and Normalization. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 22–27 May 2021. [\[CrossRef\]](#)
- Ji, Z.; Xia, T.; Han, M.; Xiao, J. A Neural Transition-based Joint Model for Disease Named Entity Recognition and Normalization. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 22–27 May 2021. [\[CrossRef\]](#)
- Wang, Y.; Shindo, H.; Matsumoto, Y.; Watanabe, T. Nested Named Entity Recognition via Explicitly Excluding the Influence of the Best Path. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 22–27 May 2021. [\[CrossRef\]](#)
- Li, F.; Wang, Z.; Hui, S.C.; Liao, L.; Song, D.; Xu, J.; He, G.; Jia, M. Modularized Interaction Network for Named Entity Recognition. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 22–27 May 2021. [\[CrossRef\]](#)
- Wang, Y.; Yu, B.; Zhu, H.; Liu, T.; Yu, N.; Sun, L. Discontinuous Named Entity Recognition as Maximal Clique Discovery. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 22–27 May 2021. [\[CrossRef\]](#)
- Li, F.; Lin, Z.; Zhang, M.; Ji, D. A Span-Based Model for Joint Overlapped and Discontinuous Named Entity Recognition. *arXiv* **2021**, arXiv:2106.14373. [\[CrossRef\]](#)
- Luong, T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. *arXiv* **2015**, arXiv:1508.04025. [\[CrossRef\]](#)

19. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E.H. Hierarchical Attention Networks for Document Classification. In Proceedings of the North American Chapter of the Association for Computational Linguistics, San Diego, CA, USA, 12–17 June 2016. [[CrossRef](#)]
20. Zhang, J.; Lin, Z.L.; Brandt, J.; Shen, X.; Sclaroff, S. Top-Down Neural Attention by Excitation Backprop. *Int. J. Comput. Vis.* **2016**, *126*, 1084–1102. [[CrossRef](#)]
21. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y. Convolutional Sequence to Sequence Learning. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017. [[CrossRef](#)]
22. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the NIPS, Long Beach, CA, USA, 4–9 December 2017. [[CrossRef](#)]
23. Eldele, E.; Chen, Z.; Liu, C.; Wu, M.; Kwoh, C.; Li, X.; Guan, C. An Attention-Based Deep Learning Approach for Sleep Stage Classification With Single-Channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2021**, *29*, 809–818. [[CrossRef](#)] [[PubMed](#)]
24. Li, X.; Xiong, H.; Wang, H.; Rao, Y.; Liu, L.; Huan, J. DELTA: DEep Learning Transfer using Feature Map with Attention for Convolutional Networks. *arXiv* **2019**, arXiv:1901.09229. [[CrossRef](#)].
25. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019. [[CrossRef](#)]
26. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Wang, S.; Hu, G. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, Online, 16–20 November 2020; pp. 657–668.
27. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z.; Wang, S.; Hu, G. Pre-Training With Whole Word Masking for Chinese BERT. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *29*, 3504–3514. [[CrossRef](#)]
28. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
29. Glorot, X.; Bordes, A.; Weston, J.; Bengio, Y. A semantic matching energy function for learning with multi-relational data. *Mach. Learn.* **2013**, *94*, 233–259.
30. Dettmers, T.; Minervini, P.; Stenetorp, P.; Riedel, S. Convolutional 2D Knowledge Graph Embeddings. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
31. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
32. Vashishth, S.; Sanyal, S.; Nitin, V.; Talukdar, P. Composition-based Multi-Relational Graph Convolutional Networks. In Proceedings of the International Conference on Learning Representations, Online, 26–30 April 2020.
33. Zhang, N.; Chen, M.; Bi, Z.; Liang, X.; Li, L.; Shang, X.; Yin, K.; Tan, C.; Xu, J.; Huang, F.; et al. CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022. [[CrossRef](#)]
34. Doğan, R.I.; Leaman, R.; Lu, Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Inform.* **2014**, *47*, 1–10. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.