Computer Science and Application 计算机科学与应用, 2025, 15(7), 100-113
Published Online July 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.157185

Hans汉斯

态靶辨治语料的深度学习标注方法

蒋欣然[1,2]，连世新[1]，董富江[1]，马启国[1,3]，马　静[1,4]，张文学[1*]

[1]宁夏医科大学医学信息与工程学院，宁夏　银川
[2]宁夏医科大学中医学院，宁夏　银川
[3]宁夏回族自治区银川市灵武市白土岗乡人民政府，宁夏　银川
[4]美华建筑设计有限公司宁夏分公司，宁夏　银川

收稿日期：2025年6月18日；录用日期：2025年7月17日；发布日期：2025年7月25日

摘　要

针对中医“态靶辨治”理论语料标注中存在的术语复杂性、表达灵活性及领域知识依赖等挑战，本研究提出一种融合深度学习的两阶段标注方法体系。首先，基于BERT + BiLSTM + CRF模型实现高精度命名实体识别，准确抽取疾病、证候、态、靶药等六类核心实体(F1值达72.06%)；其次，通过Attention + BiLSTM模型完成实体关系抽取，精准捕捉“疾病－证候－靶药”间的诊疗逻辑关系(F1值达98.23%)。实验表明，该方法显著提升标注效率与一致性，为构建态靶辨治知识图谱及开发智能诊疗系统提供了可靠的技术支撑。

关键词

态靶辨治，语料库标注，深度学习，命名实体识别，实体关系抽取

Deep Learning Annotation Methods for the Corpora of State-Target Differentiation and Treatment

Xinran Jiang[1,2], Shixin Lian[1], Fujiang Dong[1], Qiguo Ma[1,3], Jing Ma[1,4], Wenxue Zhang[1*]

[1]School of Medical Information and Engineering, Ningxia Medical University, Yinchuan Ningxia
[2]School of Traditional Chinese Medicine, Ningxia Medical University, Yinchuan Ningxia
[3]The People’s Government of Baitugang Town, Lingwu City, Yinchuan City, Ningxia Hui Autonomous Region, Yinchuan Ningxia
[4]Ningxia Branch of Meihua Architectural Design Co., Ltd., Yinchuan Ningxia

Received: Jun. 18th, 2025; accepted: Jul. 17th, 2025; published: Jul. 25th, 2025

*通讯作者。

文章引用: 蒋欣然, 连世新, 董富江, 马启国, 马静, 张文学. 态靶辨治语料的深度学习标注方法[J]. 计算机科学与应用, 2025, 15(7): 100-113. DOI: 10.12677/csa.2025.157185

Abstract

To address challenges in annotating Traditional Chinese Medicine (TCM) corpora for "State-Target Differentiation and Treatment" (STDT), including terminological complexity, expressive flexibility, and domain knowledge dependency, this study proposes a dual-stage deep learning annotation framework integrating deep learning. First, a BERT + BiLSTM + CRF model achieves high-precision named entity recognition, accurately extracting six core entities, such as diseases, syndromes, states, and target herbs (with an F1-score of 72.06%). Second, an Attention + BiLSTM model completes the extraction of entity relationships, accurately capturing diagnosis and treatment logical relationships between "disease-syndrome-target herb" (with an F1-score of 98.23%). Experiments demonstrate significant improvements in annotation efficiency and consistency, providing a robust technical foundation for constructing STDT knowledge graphs and developing intelligent diagnosis and treatment systems.

Keywords

State-Target Differentiation and Treatment, Corpus Annotation, Deep Learning, Named Entity Recognition, Entity Relationship Extraction

1. 引言

仝小林院士提出的“态靶辨治(State-Target Differentiation & Treatment, STDT)”理论体系，因其融合了现代医学的精准“靶点”思维与传统中医的整体“状态”调控优势，展现出强大的临床生命力和广阔的应用前景，成为推动中医精准化、标准化发展的关键路径之一[1] [2]。该理论强调在明确疾病核心“靶点”的基础上，精准辨识患者的整体功能状态(“态”)，实现病证结合、方证相应的精准治疗。随着“态靶辨治”在临床实践和科研中的广泛应用，积累了蕴含丰富诊疗经验的医案、文献和临床报告文本数据。

然而，这些宝贵的非结构化或半结构化文本语料(“态靶辨治语料”)蕴含着复杂的医学实体(如核心靶点疾病、中医证候要素、核心症状、核心方药、核心理化指标等)及其间的逻辑关系(如“某证候要素”导致“某靶点疾病”、“某中药”靶向治疗“某疾病”、“某症状”提示“某证候要素”等)。传统的人工标注方式不仅效率低下、成本高昂，而且难以保证标注的一致性和覆盖度，严重制约了基于大规模语料的知识发现、智能辅助诊疗系统的构建以及诊疗经验的传承与推广[3]-[5]。

近年来，自然语言处理(NLP)技术，特别是深度学习模型，在生物医学文本挖掘领域取得了显著进展[6]-[9]。命名实体识别(NER)和关系抽取(RE)作为其核心任务，是构建结构化知识库的基础。然而，直接将通用领域的 NLP 技术应用于中医“态靶辨治”语料面临独特挑战：1) 术语复杂性：中医术语具有高度的专业性、歧义性和古今演变性，“态靶辨治”体系又引入了新的概念和术语组合；2) 表达灵活性：中医文本表述风格多样，句式结构灵活多变，实体边界模糊，关系表达常隐含于语境之中；3) 领域知识依赖：准确识别实体和关系深度依赖于对中医理论(尤其是“态靶”理论)和临床实践的理解；4) 标注资源稀缺：高质量、大规模、专门针对“态靶辨治”体系的标注语料库相对匮乏。

针对上述挑战，为了高效、准确地从海量“态靶辨治”语料中提取结构化知识，本研究提出并探讨一种面向该领域的深度学习标注方法。本论文的核心工作包含两个紧密衔接的部分：1) 态靶辨治命名实

体识别：针对中医“态靶”体系中核心实体的特点，我们构建了基于 BERT 预训练语言模型、双向长短期记忆网络(BiLSTM)和条件随机场(CRF)的序列标注模型(BERT + BiLSTM + CRF)。BERT 模型强大的上下文语义理解能力为识别复杂的中医实体提供了基础，BiLSTM 有效捕捉长距离依赖特征，CRF 层则保证了输出标签序列的全局最优性。2) 态靶辨治实体关系抽取：在识别出的实体基础上，为了揭示“态”与“靶”之间、“药”与“病/证”之间的核心逻辑关系，我们设计了基于注意力机制(Attention)增强的双向长短期记忆网络(Attention + BiLSTM)模型。该模型能聚焦于句子中对关系判断最具信息量的关键词语，有效提升对隐含、复杂语义关系的捕捉能力。

本研究的目标在于，通过构建一套融合预训练语言模型、序列建模、注意力机制等深度学习技术的标注方法体系，显著提升“态靶辨治”语料中关键医学实体及其关系自动标注的精度与效率，为构建“态靶辨治”知识图谱、开发智能辅助诊疗工具、深化中医精准化研究奠定坚实的技术基础。这不仅是中医药信息学领域的重要探索，也是深度学习技术赋能传统医学知识工程的有益实践。

2. 基于 BERT + BiLSTM + CRF 的态靶辨治命名实体识别

2.1. BERT + BiLSTM + CRF 模型

态靶辨治中医药文本特点对命名实体识别任务造成了一定挑战。首先，态靶辨治文本所涉及的实体类别较为丰富，包括疾病、证候、态、靶方、靶药、药理作用等；与此同时，文本中包含长度较长的实体，如病因病机“肺脾亏虚致运化失职”等；以及难以界定和标记的嵌套实体，如方剂“葛根芩连汤”本身就是一个靶方实体，又包含“葛根”这一中药实体等，这些实体边界不清将会对分词等步骤产生影响，进而影响命名实体识别任务的进程。

BERT + BiLSTM + CRF 的模型结构主要分为三层，这种组合模型能够充分利用 BERT 的语义理解能力、BiLSTM 的序列建模能力和 CRF 的标签依赖处理能力，提高了命名实体识别的准确性，其模型结构图如图 1 所示。

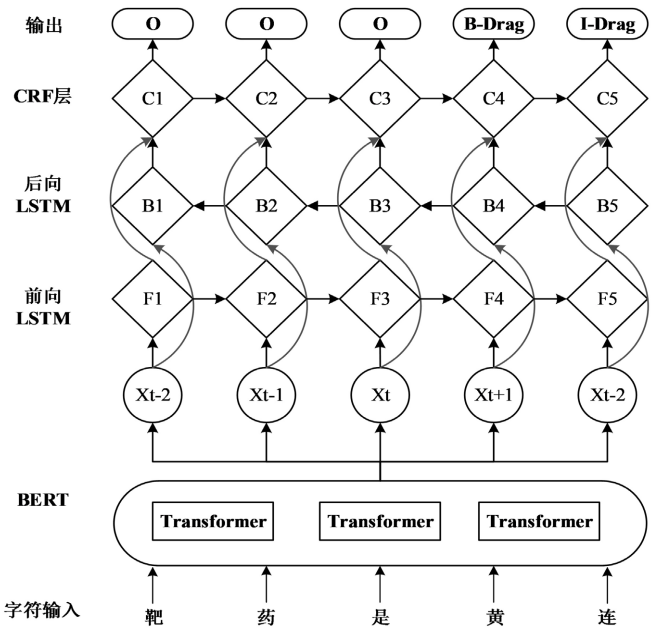

Figure 1. Model structure diagram of BERT + BiLSTM + CRF
图 1. BERT + BiLSTM + CRF 模型结构图

1) 词嵌入层

BERT + BiLSTM + CRF 模型的词嵌入层是使用了 BERT 预训练模型。它是一种基于 Transformer 架构的预训练语言模型。Transformer 架构主要由多头注意力(Multi-Head Attention)机制和前馈神经网络(Feed-Forward Neural Network)组成。在多头注意力机制中，它可以同时关注输入序列的不同位置信息，计算出每个位置的表示。例如，对于一个句子中的每个单词，它会综合考虑句子中其他单词的信息来更新该单词的表示，这种方式能够捕捉到丰富的语义和语法信息。它还通过大规模无监督语料进行预训练，预训练任务主要包括掩码语言模型(Masked Language Model, MLM)和下一句预测(Next Sentence Prediction, NSP)。在 MLM 任务中，模型会随机掩盖输入句子中的一些单词，然后预测这些被掩盖单词的原始内容；NSP 任务则是判断两个句子是否是连续的句子。模型通过这种方式训练，会准确地学习到句子的上下文语义关系，提高词汇理解能力，同时也减少了对特定顺序的依赖。BERT 的模型结构图如图 2 所示。

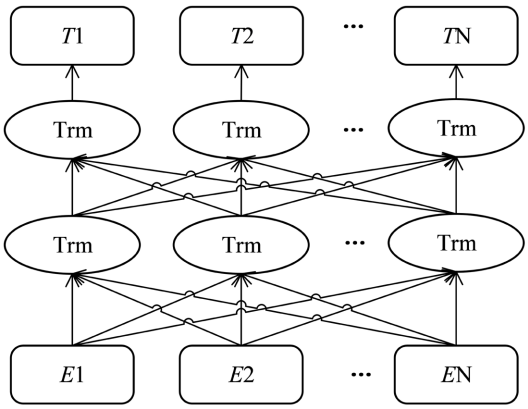

Figure 2. Model structure diagram of BERT
图 2. BERT 模型结构图

2) 上下文编码层

BERT + BILSTM + CRF 模型的上下文编码器层采用的是双向长短期记忆网络(BiLSTM)。LSTM 是一种特殊的循环神经网络(Recurrent Neural Network, RNN)，它可以有效处理长序列数据中的长期依赖问题。LSTM 模型是由 t 时刻的输入 t、细胞状态 C_t、临时细胞状态 $\tilde{C}_t$、隐藏层状态 h_t、遗忘门 f_t 和输出门 O_t 组成。遗忘门决定了上一时刻的细胞状态有多少信息被遗忘，输入门决定了当前输入有多少信息被更新到细胞状态，输出门则控制细胞状态有多少信息作为当前时刻的输出。LSTM 内部结构图如图 3 所示。

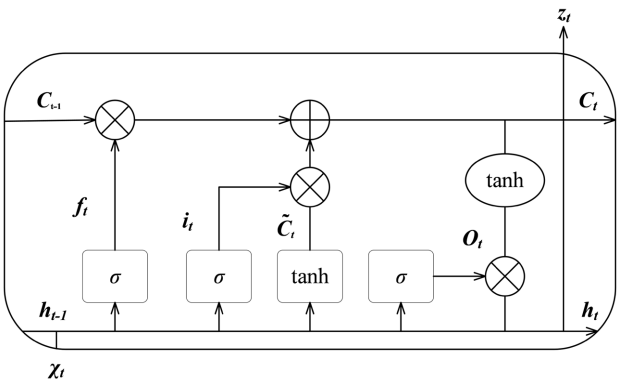

Figure 3. Model structure diagram of LSTM
图 3. LSTM 模型图

3) 解码器层

解码层由条件随机场(Conditional Random Field, CRF)组成，其目的是解决输出标签之间的相关性，从而获得文本的全局最优标注序列。在 CRF 标签解码层中，该模型通过动态规划算法来计算输出实体序列的最优路径。具体模型根据当前标签取值和前一个标签取值，计算输出实体序列中每个位置上各个标签取值的得分，并选择得分最高的标签作为当前位置上的输出实体标签。在计算过程中，转移概率矩阵约束了相邻标签之间的转移概率，从而保证了输出实体序列的合法性，CRF 的模型结构如图 4 所示。

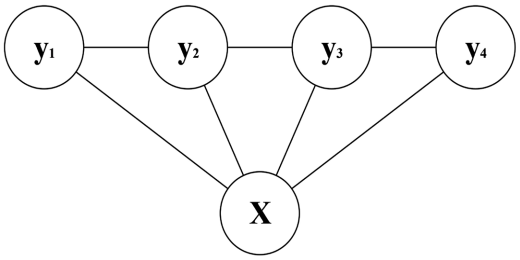

Figure 4. Model structure diagram of CRF
图 4. CRF 模型结构图

假设输入序列为 $X=(x_1,x_2,x_3\cdots,x_n)$，预测的标注序列为 $Y=(y_1,y_2,y_3\cdots,y_n)$。编码层输出的得分矩阵 P 的大小为 $n\times k$，其中 n 为输入序列的长度，k 为不同类型的标签数，$P_{i,yi}$ 表示在给定输入序列 X 的条件下，第 i 个字符标注为 y_i 的概率得分，状态转移得分矩阵 A 表示不同标签之间转移的概率得分，$A_{yi,yi+1}$ 表示从标注 y_i 转移到标注 y_{i+1} 的转移概率得分。

在给定序列的条件下，可以获取相应标注序列的得分 $s(X,y)$。其计算公式如式(1)所示：

$$s(X,y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \tag{1}$$

预测概率为 $p(y|X)$。计算公式如(2)所示：

$$p(y|X)=\frac{e^{s(X,y)}}{\sum_{y\in Y_X}e^{s(X,y)}} \tag{2}$$

解码层的目标是获得最优标注序列。使用动态规划的思想，逐个计算每个位置上标注序列的最大概率，并记录每个位置的最优路径。最终通过回溯，得到整个句子的最优标注序列。凭借这一特性，CRF 可以充分利用上下文信息，为每个标签的预测提供更全面、准确的依据，进而提升模型在序列标注任务中的表现。

2.2. 态靶辨治命名实体标注

1) 数据来源与预处理

由于实验的语料标注对实体识别的准确性有较大影响，所以在筛选文献方面需要符合专业、准确、有效等特点。实验数据来源于 CNKI 期刊数据库，检索表达式：主题：(“态靶”or“态靶辨治”or“态靶辨证”or“态靶因果”or“态靶结合”)。选取的文献需符合中文核心期刊来源、创始人撰写的经典文献、态靶辨治与糖尿病相结合等三个特点，最终筛选 40 篇文本进行预处理，去除特殊字符与标点符号等非文本结构数据，通过正则表达式去除无用的特殊字符、标点符号等数据，只留下有用的字符、常用的中文标点符号等文本，对于一些过长的句子，按照优先级对其进行分割，对处理后的文本进行汇总，整理后作为本次实验的待标注语料，如图 5 所示。

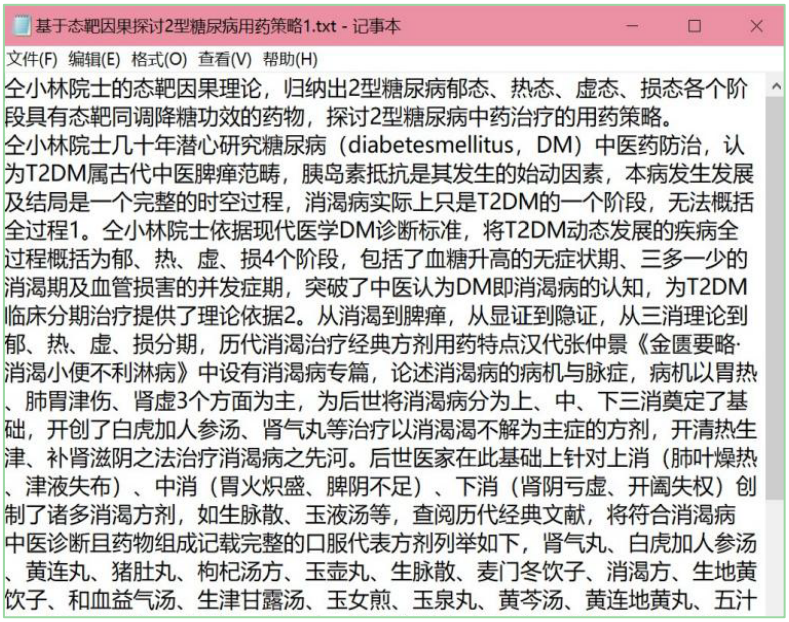

Figure 5. NER corpus to be labeled
图 5. NER 待标注语料

2) 标注语料实体类别

根据人工预标注和专家评判，确定了态靶辨治标注语料的实体类别有疾病、证候、态、靶方、靶药、靶点。

3) 语料标注结果

使用态靶辨治命名实体标注工具对待标注文本进行标注，得到格式为 BIE 的输出结果，由于本命名实体识别模型需求的输入格式为 BIO，所以利用 Python 语句对标注工具输出的结果进行格式转化，最终得到 BIO 格式的文件，将其整理在同一个 TXT 文档中以备训练，具体标注结果如下图 6 所示。

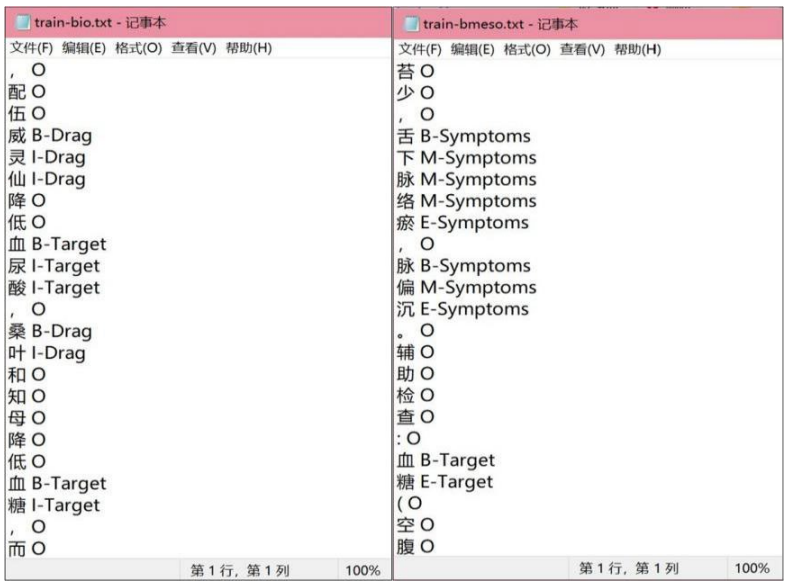

Figure 6. BIO annotation results of named entity recognition for STDT
图 6. 态靶辨治命名实体识别 BIO 标注结果

2.3. 命名实体识别的实验结果与分析

1) 实验配置与参数

本命名实体识别实验与分析，确定实验参数如下表 1 所示。

Table 1. Environment configuration table of named entity recognition
表 1. 命名实体识别环境配置表

项目	配置
操作系统	Ubuntu 20.04
CPU	Intel(R) Xeon(R) Platinum 8474C
GPU	RTX 4090D (24 GB)
Python	3.8.0
Pytorch	1.11.0
内存	80 GB

如下表 2 所示，batch_size 表示批量大小，即每次训练输入网络的样本量；max_seq_length 表示最大序列的长度；learning_rate 表示学习率；batch_size 和 epochs 参数是通过不同实验进行对比，选择最好的效果得出最优参数。

Table 2. Experimental parameter table of named entity recognition
表 2. 命名实体识别实验参数表

参数	数值
max_seq_len	512
epochs	10
train_batch_size	12
dev_batch_size	12
bert_learning_rate	3e−5
crf_learning_rate	3e−3

2) 评估指标

评价指标使用精确率 P、召回率 R 和 F1-score。精确率指分类器正确识别为某个类别的实体数占分类器总共识别为该类别的实体数的比例。精确率越高，代表分类器在正确预测实体类别方面的能力越强。计算公式如下(3)所示：

$$P = \frac{TP}{TP + FP} \tag{3}$$

其中，TP 表示真正例，即分类器正确预测为某个类别的实体数；FP 表示假正例，即分类器错误地将非某个类别的实体预测为该类别的实体数。召回率指分类器正确识别为某个类别的实体数占该类别实体总数的比例。召回率越高，代表分类器在识别实体方面的能力越强。计算公式如下(4)所示：

$$R = \frac{TP}{TP + FN} \tag{4}$$

其中，FN 表示假反例，即分类器错误地将某个类别的实体预测为非该类别的实体数。$F1$ 值是准确率和召

回率的调和平均数，用来综合评价分类器的性能 $F1$ 值越高，代表分类器在预测实体类别方面的能力越强。

其计算公式如下(5)所示：

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{5}$$

3) 比较方法 Random + BiLSTM + CRF

为进一步探究融入 BERT 预训练模型的实际效用，运用 PyTorch 框架内随机生成的词嵌入，将其作为词嵌入层引入 BiLSTM + CRF 模型，并维持与 BERT + BiLSTM + CRF 模型一致的参数设定开展实验。

4) 实验结果与分析

在历经 10 个 epoch 的训练后，模型的损失值逐步趋于平稳。在此过程中，我们同步留存了训练阶段 F1 值表现最优的模型，模型的实验结果如下表 3 所示。本文方法具体的实体类别的实验结果如表 4 所示。

Table 3. Experimental results of named entity recognition in STDT
表 3. 态靶辨治命名实体识别的实验结果

模型	*P* (%)	*R* (%)	F1 (%)
Random + BiLSTM + CRF	73.35	50.08	59.85
BERT + BiLSTM + CRF (本文方法)	77.46	67.36	72.06

Table 4. Entity category results of BERT + BiLSTM + CRF model
表 4. BERT + BiLSTM + CRF 模型识别的实体类别结果

实体类别	*P* (%)	*R* (%)	F1 (%)
Drag	91.36	93.54	92.43
Symptoms	68.42	42.62	52.53
Disease	60.00	50.97	55.11
Target	70.71	51.03	59.28
State	90.91	97.56	94.12
Prescription	71.43	93.75	81.08

从实验数据可以观察到，相较于 Random + BiLSTM + CRF 模型，BERT + BiLSTM + CRF 模型在精确率上提升了 4.11%，召回率提升了 17.28%，F1 值提升了 12.21%。从实验结果可以证明，BERT 模型融入 BiLSTM + CRF 架构后，极大地优化了模型整体性能，提升模型精准度、强化对目标信息的提取能力等。

3. 基于 Attention + BiLSTM 的态靶辨治实体关系抽取

实体关系抽取是中医文本自然语言处理的重要任务之一，目的在于明确被识别出实体间的语义关系，一般实体关系抽取的结果由三元组的形式呈现，例如识别出“葛根芩连汤”与“糖尿病”之间的“治疗”关系。通过对语义关系的明确，可以了解到态靶辨治文本中“疾病”与“态”、“态”与“靶药”、“靶药”与“靶点”等重要关系，有利于加深对该理论中诊疗机制的理解，并辅助后续态靶辨治领域知识图谱的构建。

3.1. Attention + BiLSTM 模型

注意力机制(Attention Mechanism)是深度学习中一种重要的技术，最初在机器翻译任务中被提出，随后在自然语言处理、计算机视觉等领域被广泛应用。其核心在于通过动态分配权重使模型能够聚焦于输入数据中的重要部分，以此提升模型的性能。在自然语言处理任务中，注意力机制通常用于捕捉文本中

不同词或句子之间的依赖关系。

基于 Attention + BiLSTM 的关系抽取模型主要由 BiLSTM 层和注意力层以及输出层构成。BiLSTM 层用于对输入的文本序列进行特征提取，能够同时考虑序列的正向和反向信息。注意力层则是在 BiLSTM 的输出基础上，动态地分配权重，聚焦于对关系抽取更重要的部分。输出层一般是一个分类层，用于根据前面提取和加权后的特征来判断实体之间的关系类别。其模型结构图如图 7 所示。

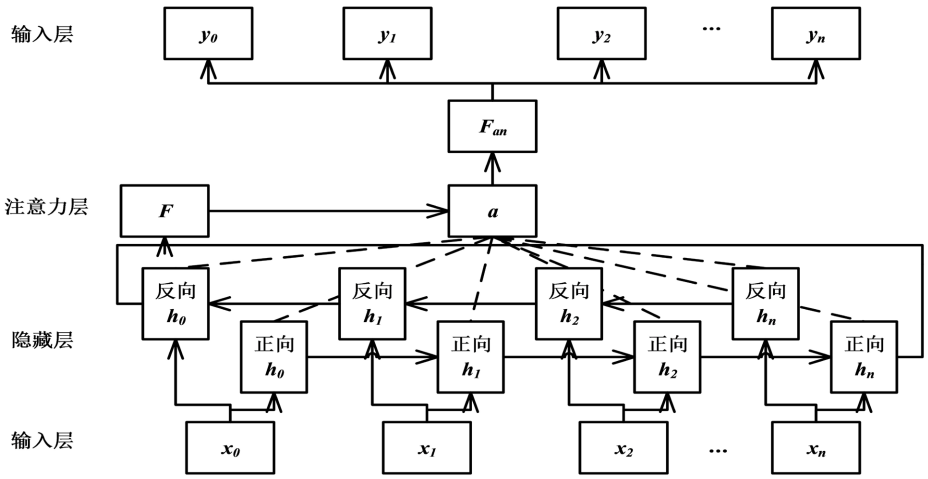

Figure 7. Model structure diagram of Attention + BiLSTM
图 7. Attention + BiLSTM 模型结构图

1) 注意力机制

注意力机制(Attention Mechanism)是一种模仿人类视觉注意力的方法。在处理信息时，人类会选择性地聚焦于部分关键信息，而忽略其他无关部分。注意力机制在深度学习模型中起到类似的作用，它能够让模型在处理大量信息时，动态地聚焦于对当前任务更重要的信息部分。注意力层的核心是计算注意力权重。设 BiLSTM 层的输出为 $H=(h_1,h_2,\cdots,h_n)$，注意力机制会根据这些输出计算注意力概率分布 $a=(a_1,a_2,\cdots,a_n)$。通常通过一个可学习的函数来计算，如下公式(6)所示：

$$a_i=\frac{\exp(e_i)}{\sum_{j=1}^{n}\exp(e_j)} \tag{6}$$

其中，e_i 是根据 $e_i=w^T\tan nh(Wh_i+b)$ 是计算得到的一个得分，用于衡量 h_i 的重要性。计算出注意力概率分布后，利用这些权重对 BiLSTM 的输出进行加权求和，得到最终的特征表示如下公式(7)所示：

$$\mathrm{F}_{\mathrm{att}}=\sum_{i=1}^{n}a_ih_i \tag{7}$$

这个最终特征表示聚焦于输入序列中对关系抽取更关键的部分，能够有效利用序列中的重要信息。

2) 输入层

基于注意力机制的 BiLSTM 关系抽取模型的输出层采用了 $Soft\max$ 函数。将注意力层得到的最终特征 F_{att} 输入到输出层，经过 $Soft\max$ 函数计算分类标签的概率分布，如下公式(8)所示：

$$\mathrm{y}=Soft\max(W_{out}F_{att}+b_{out}) \tag{8}$$

其中，W_{out} 和 b_{out} 分别是输出层的权重和偏置。这个概率分布表示输入文本中实体之间属于各种关系类别的可能性。在训练阶段，模型会根据预测的概率分布与真实类别分布求取交叉熵损失。通过反向传播算法，根据损失来更新模型的参数，包括双向 LSTM 层和注意力层的参数，以提高模型的关系抽取性能。

3.2. 态靶辨治实体关系标注

1) 标注语料实体关系

根据本研究对态靶辨治关系推理的需求总结出下列实体关系进行关系抽取的训练与实验，具体详见表 5。

Table 5. Annotation list of named entity relationship
表 5. 实体关系标注列表

头实体	关系	尾实体	示例
证候	诊断	疾病	口渴多饮诊断糖尿病
靶点	辅助诊断	疾病	血糖辅助诊断糖尿病
靶方	治疗	疾病	葛根芩连汤治疗糖尿病
靶药	组成	靶方	黄连、黄芩、葛根等组成葛根芩连汤
靶方	调态	态	葛根芩连汤调节糖尿病热态
靶药	打靶	靶点	黄连降血糖功效

2) 实体关系标注结果

使用态靶辨治命名实体标注工具对待标注文本进行实体关系的标注，得到的标注结果输出如下图 8 所示。

Figure 8. Pre-annotation of relationship for STDT
图 8. 态靶辨治关系预标注

根据本研究实体关系抽取的模型所需要的输入格式，利用 Python 语句对输出的结果进行格式转化，最终得到适合格式的数据，将其整理在同一个 txt 文档中以备训练，具体实体关系标注结果如下图 9 所示。

Figure 9. Annotation results of the entity relationship for STDT
图 9. 态靶辨治实体关系标注结果

3.3. 实体关系抽取的实验结果与分析

1) 实验参数

基于 Attention + BiLSTM 关系抽取模型的训练参数中 batch_size 指的是每次训练输入网络的样本量，max_seq_len 指的是最大序列长度，learning_rate 表示学习率，epochs 表示设定的训练次数，具体如表 6 所示。

Table 6. Experimental parameter of entity relationship extraction
表 6. 实体关系抽取实验参数

参数名称	参数值
max_seq_len	256
epochs	5
train_batch_size	24
dev_batch_size	12
learning_rate	3e−5
adam_epsilon	1e−8

2) 评估指标

本次实验选用精确率、召回率及 F1 值作为关键评价指标。精确率是指在所有被预测为正例的样本中，真正的正例所占的比例。召回率是指在所有实际为正例的样本中，被正确预测为正例的样本所占的比例。F1 值是精确率和召回率的调和平均数，它综合考虑了精确率和召回率，用于衡量模型的整体性能。对应的计算公式分别为(3)~(5)。

3) 实验结果与分析

将全部的关系抽取数据集按照训练集和验证集 8:2 的比例划分，其实验结果详细如下表 7 所示。

Table 7. Experimental results of entity relationship extraction
表 7. 实体关系抽取实验结果

模型	*P* (%)	*R* (%)	F1 (%)
Attention + BiLSTM	98.48	98.15	98.23

由上述结果可见，本研究在态靶辨治实体关系抽取任务中，运用注意力机制联合双向长短期记忆网络模型取得了良好的效果。

4) 注意力机制可视化结果与分析

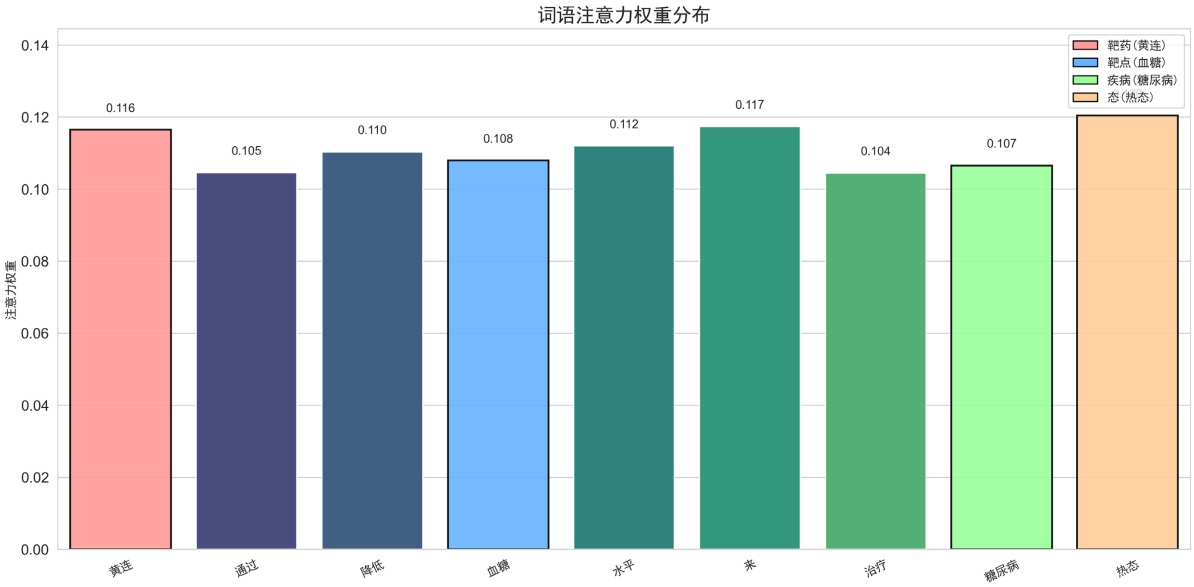

Figure 10. Distribution of word attention weights
图 10. 词语注意力权重分布

注意力机制可视化结果表明如图 10~13 所示，在预测“黄连”与“血糖”之间的“靶药”关系时，深度学习模型展现出符合中医诊疗逻辑的语义聚焦特性。融合注意力机制的深度学习模型不仅能准确提取中医文本中的实体关系，还能通过可解释的权重分配模拟人类专家的诊疗推理过程，这对构建可信赖的中医智能辅助系统具有重要价值。

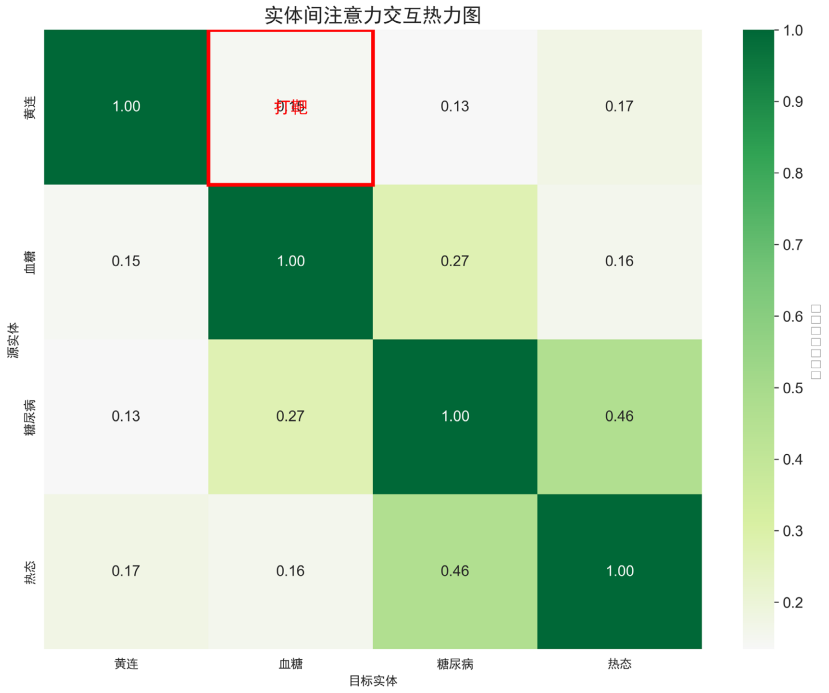

Figure 11. Heat map of the entity relationship attention
图 11. 实体关系注意力热力图

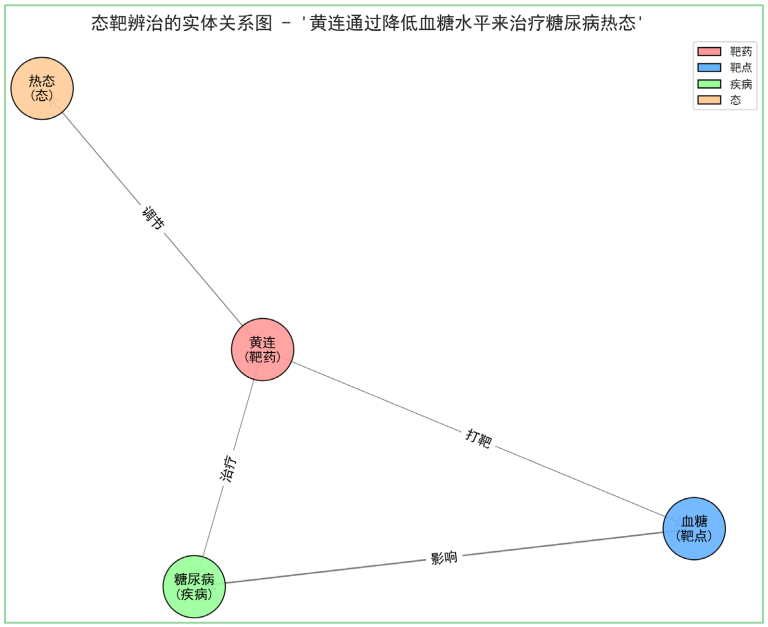

Figure 12. Visualization of the relationship between entity attention weights and “target relationship”
图 12. 实体注意力权重与“打靶”关系可视化

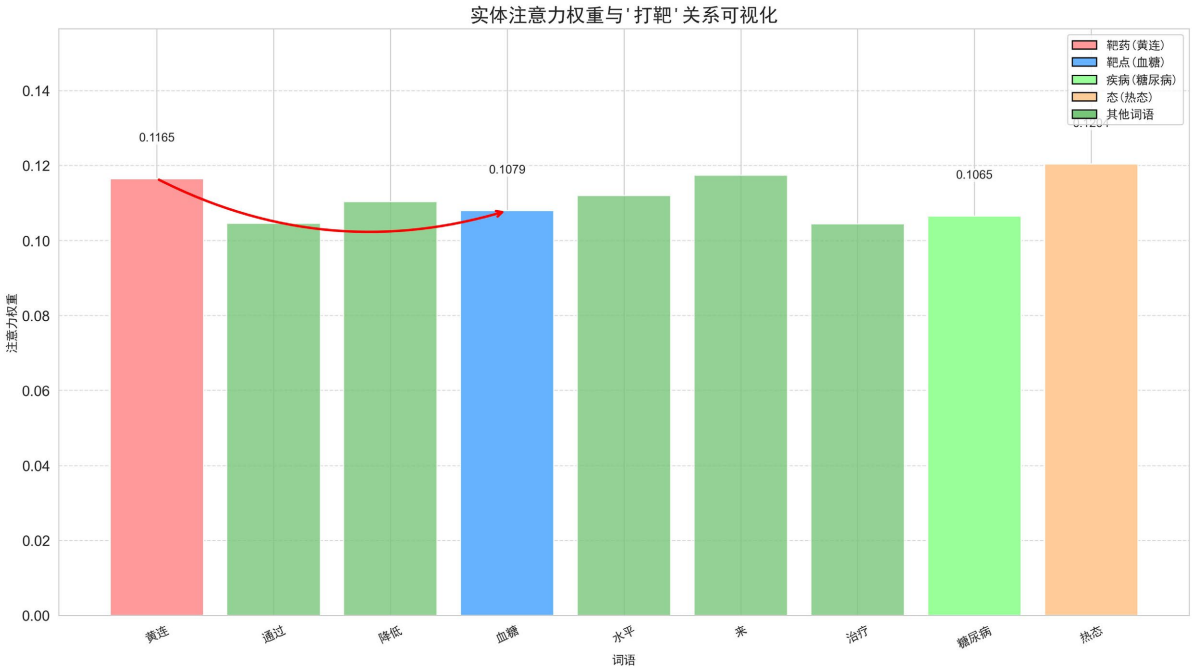

Figure 13. Analysis of detailed attention weights
图 13. 详细注意力权重分析

4. 结束语

本研究针对态靶辨治语料标注的领域特殊性，创新性地融合预训练语言模型、序列标注与注意力机制，构建了高效、可扩展的深度学习标注体系。实验验证了该方法在复杂中医实体识别及隐含关系抽取上的优越性，有效解决了传统人工标注效率低、一致性差的问题。未来工作将扩大标注语料规模，优化模型对古今术语演变的适应性，并推动成果在中医辅助诊疗平台中的应用。本研究为中医药知识工程与人工智能的深度融合提供了新范式。

基金项目

宁夏自然科学基金：面向仝小林院士态靶辨治的中医知识图谱构建研究(2023AAC03165)，态靶辨治不确定性知识的语料库构建研究(2024AAC03214)。

参考文献

[1] 仝小林. 态靶医学——中医未来发展之路[J]. 中国中西医结合杂志, 2021, 41(1): 16-18.

[2] Gou, X.W., Gao, Z.Z., Yang, Y.Y., Li, Q.W., Chen, K.Y., Lei, Y., Song, B., Zhao, L.H. and Tong, X.L. (2021) State-Target Strategy: A Bridge for the Integration of Chinese and Western Medicine. *Journal of Traditional Chinese Medicine*, **41**, 1-5.

[3] 原旎, 卢克治, 袁玉虎, 等. 基于深度表示的中医病历症状表型命名实体抽取研究[J]. 世界科学技术-中医药现代化, 2018, 20(3): 355-362.

[4] 谢云霏, 贾李蓉. 中医药知识图谱构建技术及应用的研究进展[J]. 中国医药导报, 2024, 21(20): 62-66.

[5] 胡嘉元, 邱瑞瑾, 孙杨, 等. 自然语言处理及其在医学领域的应用. 中国循证医学杂志, 2024, 24(10): 1205-1211.

[6] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., *et al.* (2019) BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*, **36**, 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

[7] Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., *et al.* (2019) Deep Learning in Clinical Natural Language Processing: A Methodical Review. *Journal of the American Medical Informatics Association*, **27**, 457-470. https://doi.org/10.1093/jamia/ocz200

[8] Liu, P., Guo, Y., Wang, F. and Li, G. (2022) Chinese Named Entity Recognition: The State of the Art. *Neurocomputing*, **473**, 37-53. https://doi.org/10.1016/j.neucom.2021.10.101

[9] Liu, Z., Luo, C., Zheng, Z., Li, Y., Fu, D., Yu, X., *et al.* (2021) TCMNER and PubMed: A Novel Chinese Character-Level-Based Model and a Dataset for TCM Named Entity Recognition. *Journal of Healthcare Engineering*, **2021**, Article ID: 3544281. https://doi.org/10.1155/2021/3544281