计算机科学
COMPUTER SCIENCE

基于深度学习的细粒度医学知识图谱构建

王钰涵，马涪元，王英

引用本文

王钰涵，马涪元，王英. 基于深度学习的细粒度医学知识图谱构建[J]. 计算机科学，2024，51(11A)：230900157-7.

WANG Yuhan, MA Fuyuan, WANG Ying. Construction of Fine-grained Medical Knowledge Graph Based on Deep Learning [J]. Computer Science, 2024, 51(11A): 230900157-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230900157

基于深度学习的细粒度医学知识图谱构建

王钰涵[1]　马涪元[2]　王　英[3]
1 吉林大学软件学院　长春 130012
2 吉林大学人工智能学院　长春 130012
3 符号计算与知识工程教育部重点实验室(吉林大学)　长春 130012
(yuhanw23@mails.jlu.edu.cn)

摘　要　医疗知识图谱作为整合海量医疗信息的有力工具,正被广泛应用于临床决策支持系统、医疗问答系统等便民平台。目前,大规模医疗知识图谱层出不穷,但大多都将注意力放在实体数量的扩充,而忽略了实体种类的细粒度化。医疗术语具有冗长且难以理解的特点,因此构建细粒度化的知识图谱可以在很大程度上提高知识图谱便民系统的实用性,并为问答系统提供更具有针对性的诊断说明。文中针对垂直网站爬取的大规模医疗知识库,以实现医疗长文本细粒度化为目标,运用 BiLSTM 从长句子的两个方向为每个词语建模完整上下文信息,同时引入预训练模型 BERT 加强对词语上下文语义的建模,并结合 CRF 模型学习状态转移矩阵维持标签序列的一致性,高效识别长句中的实体,并通过实体对齐和属性填充构建细粒度医疗知识图谱。医疗实体细粒度化任务的对比实验表明,BERT+BiLSTM+CRF 模型的效果优于其他模型,可视化结果也说明了所提方法进行细粒度化的有效性。

关键词:知识图谱;BiLSTM;CRF;细粒度
中图分类号　TP391

Construction of Fine-grained Medical Knowledge Graph Based on Deep Learning

WANG Yuhan[1], MA Fuyuan[2] and WANG Ying[3]
1 College of Software, Jilin University, Changchun 130012, China
2 College of Artificial Intelligence, Jilin University, Changchun 130012, China
3 Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, Jilin University, Changchun 130012, China

Abstract　As a powerful tool for integrating massive medical information, medical knowledge graphs are being widely evaluated on convenient platforms such as clinical decision support systems and medical question and answer systems. At present, large-scale medical knowledge graphs are emerging one after another, but most of them focus on the supplement of the number of entities. Medical terminology is lengthy and difficult to understand. Therefore, building a fine-grained knowledge graph can make the knowledge graph convenient for the system to a large extent. practicality and provide more crown diagnostic instructions for the question and answer system. This paper targets the large-scale medical knowledge base crawled by vertical websites, with the goal of achieving fine-grained medical long texts. BiLSTM is used to model complete contextual information for each word from both directions of the long sentence. At the same time, we introduce the pre-training model BERT to enhance the modeling of word context semantics and combined with the CRF model learning status. The incremental matrix maintains the consistency of the label sequence, efficiently identifies entities in long sentences, and builds a fine-grained medical knowledge graph through entity alignment and attribute filling. Comparative experiments on the fine-grained task of medical entities demonstrate that the BERT+BiLSTM+CRF model is better than other models, and the visualization results also illustrate the fine-grained effect of this method.

Keywords　Knowledge graph, BiLSTM, CRF, Fine-grained

1　引言

医疗健康作为人们生活的中心,其信息库也日益增大,如何有效管理医疗信息成为人工智能的关注重点。2012 年谷歌提出了知识图谱[1],在整合海量数据方面展现出了巨大优势,因而知识图谱成为了管理医疗信息的有力工具。互联网上基于医疗知识图谱的智能化应用,例如原发性肝癌问答系统[2]、临床决策支持系统(CDSS)[3]、TCMKG 中医药知识平台[4]等,已开始崭露头角,支撑着人们对医疗知识的需求。

目前,电子病历知识图谱[5]、CMeKG[6]、中医药 TCM 图谱[4]等大规模的知识图谱日益增多,这些图谱资源为了方便构建医疗辅助决策系统、问答系统等知识图谱的智能化应用,

基金项目:国家自然科学基金(62272191);吉林省科技厅重点研发项目(20220201153GX)
This work was supported by the National Natural Science Foundation of China(62272191) and Science and Technology Development Program of Jilin Province(20220201153GX).
通信作者:王英(wangying2010@jlu.edu.cn)

将注意力集中于医疗知识的整体结构，因此保留其冗长的医疗信息术语。但是，长文本术语仅适合帮助专业人士进行辅助决策，对于非专业人士而言，长文本术语的可理解性较低，这也成为长文本知识图谱在便民应用上的一大阻碍。相比长文本信息，知识的细粒度化强调了医疗术语中内部实体的联系，在应用上为医疗问答系统提供了更多的问答可能性，可以更精确地满足用户在不同信息粒度上的认知需求。

针对大规模医疗数据处理，通过选择合适的实体关系抽取模型，可以在很大程度上解决医疗长文本细粒度化这一问题。随着知识图谱的发展，众多基于机器学习、深度学习的模型，例如 KNN+CRF[7]、RD-CNN-CRF[8]、双向长短记忆网络(Bi-directional Long Short-Term Memory，BiLSTM)等，被提出并应用于命名实体识别(Named Entity Recognition，NER)实现实体关系抽取。目前的实体关系抽取模型在完成 NER 任务时，更加注重同一类别下海量实体关系的整理，从而忽略了实体文本冗长且难以理解的特点。

为了解决长文本知识图谱带来的弊端，本文将重点关注对实体和关系种类的扩充，以深度学习为基础，针对医疗知识图谱中的长句子序列，使用 BiLSTM 模型，利用了 LSTM 门控单元序列化建模长句中相邻词语语义的相互影响，同时从两个方向捕获句子中单词的完整上下文信息以获取自身表示。基于 BiLSTM 获取的蕴含语义信息的词表示，本文结合条件随机场(CRF)层进行实体标签预测，条件随机场通过学习一个状态转移矩阵可以确保预测的标签形成有效的实体序列，并遵守标签转换的约束，增强实体识别的能力。通过提取知识图谱中长句子中的实体和关系，并基于实体识别对长实体进行再拆分，实现文本细粒度化，构建更加实用的大规模细粒度医疗知识图谱。本文在瑞金医疗数据集上对比了 RNN，BERT，RNN+CRF，LSTM+CRF，BiLSTM+CRF 等相关模型的性能，说明了本文方法 BERT+BiLSTM+CRF 在实体识别任务上的有效性，同时对经过细粒度化的医疗信息知识图谱进行了可视化展示。

2 相关工作

为了给医学领域提供更有效的知识支撑，医疗知识图谱的细粒度化是必要的。通过命名实体识别模型实现细粒度操作，从一个长文本中抽取出多个三元组知识，对医疗语句中的词级语义进行表示，细化文本内涵、增强医疗术语的可理解性。

目前，国内外许多的医疗知识图谱被广泛应用，例如 DBpedia[9]，Freebase[10]，Yago[11]，CN- DBpedia[12] 以及 OpenKG[13]等。NER 任务作为构建知识图谱至关重要的一环，越来越多的实体关系抽取模型被研究挖掘，文献[14]将这些模型总结为 3 类，即基于规则和词典的方法、基于传统机器学习的方法和基于深度学习的方法，根据该种分类方法，对已有的 NER 任务模型进行深入研究和总结。

2.1 基于规则和词典的方法

命名实体识别，本质上是由 Rau[15] 结合启发式通过人工方式编写规则，实现的一种从财经新闻中提取公司名称的自动化算法演变而来。在早期技术不成熟的阶段，主要采取编写规则[15-18]实现实体抽取，相比人工抽取的方法，编写规则提高了 NER 任务的效率和准确率，但规则的灵活性和在多样化数据集上的可迁移性不高。

2.2 基于传统机器学习的方法

为了提高模型可迁移性，引入机器学习目前最常用的方法包括隐马尔可夫模型(Hidden Markov Model，HMM)和条件随机场。例如 Ponomareva 等[19]提出了 HMM 模型的生物医学 NER 系统，提供了关于实体边界的附加信息；Xu 等[20]提出了一种半监督迭代模式学习方法，用于抽取疾病风险关系，建立生物医疗知识库；Sui 等[21]通过多特征 CRF 为化学物质-疾病 NER 模型的特征确定提供参考。同时，文献[22]发现单一机器学习方法存在一定的局限性，于是有学者对混合实体抽取方法进行了探索。Liu 等[7]提出了 KNN+CRF 模型，用于捕获推文中的细粒度信息；Li 等[23]提出了机器学习与规则结合的方法对医学实体进行抽取，在中文电子病历 NER 任务上达到了不错的效果。将机器学习加入实体关系抽取任务的想法，优化了任务性能，但是该类方法存在特征提取误差传播问题，为了解决这一缺陷，引入了深度学习技术强大的特征表达能力。

2.3 基于深度学习的方法

目前，最常用于实体关系抽取模型的深度学习方法为卷积神经网络(Convolutional Neural Network，CNN)和循环神经网络(Recurrent Neural Network，RNN)。其中基于 CNN 的模型如下：Feng 等[24]基于 CNN 提出的强调实体关系分类的模型 CNN-RL、Ji 等[25]提出的引入句子级注意力的实体关系抽取模型 APCNNs 以及 Wu 等[26]解决了中文边界模糊性问题的 BERT-CNN 模型等。同时，也有众多利用 RNN 的记忆性提出的实体关系抽取模型，例如 Ding 等[27]提出的结合注意力机制的 BiLSTM 抽取中文生物医学实体关系的模型；Ukov-gregori 等[28]提出的通过多层独立 BiLSTM，实现并行 RNN 模型，大大减少了参数量；Gao 等[29]提出的加入关系发现词算法的 BiGRU-2ATT 模型，以上模型都有效提升了 NER 任务的效果。文献[14]还将基于深度学习的关系抽取方法分为流水线方法和实体关系联合抽取方法两类。但流水线方法[30-32]在分布执行的过程中难以避免传播误差，因此实体识别与关系抽取相融合的联合抽取模型应用更为广泛。例如 Katiyar 等[33]首次将注意力机制与 BiLSTM 结合实现联合抽取，并解决了已有模型依赖于复杂特征的缺点；Miwa 等[34]提出 LSTM-RNN 堆叠的 SPTree，通过共享参数实现联合抽取；Xiao 等[35]提出了一种粗粒度的联合抽取模型 Ch-MEL，提高了 SemEval-2010 数据集上 NER 任务的效果。

目前，神经网络作为深度学习的核心，被广泛用于实现 NER 任务，然而大多模型都忽略了知识细粒度问题。因此，本文将机器学习与神经网络结合，利用 BERT+BiLSTM+CRF 模型，以实现医疗文本细粒度化为目标完成命名实体识别，从而构建细粒度的医疗知识图谱。

3 基于 BERT+BiLSTM+CRF 的联合抽取模型

命名实体识别(Named Entity Recognition，NER)可以根据实体在不同的句子中的表示以及上下文关系构造相应的三元组，为后续构建知识图谱提供支撑。本文提出了 BERT+BiLSTM+CRF 联合抽取模型，用于实现细粒度三元组的抽取，模型的整体结构如图 1 所示，模型可以通过词嵌入层、BiLSTM 层和 CRF 层很好地完成任务。

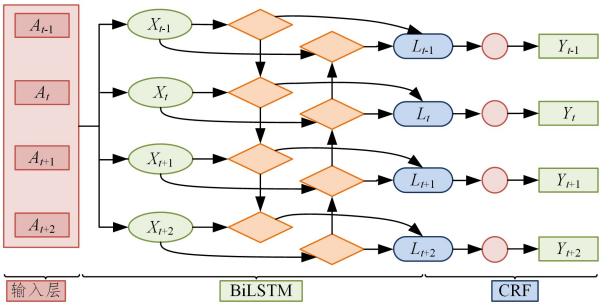

图 1　BERT＋BiLSTM＋CRF 模型的整体结构

Fig. 1　Overall structure of BERT＋BiLSTM＋CRF

3.1　输入层

由于计算机无法理解人类的语言，当一个句子传入模型时，需要词向量层或者引入预训练模型将词汇映射为计算机能够理解的向量格式，帮助计算机完成实体识别任务。从数学的角度分析，其本质就是将多维的词语空间，投射至多个低纬度的向量空间，至此可以提高模型的训练效果。

为了防止一词多义影响医疗文本的上下文语义，在 BERT＋BiLSTM＋CRF 模型中选用 BERT 预训练模型，其本质是一个多层的 Transformer 编码器，它通过多个 Transformer 层处理输入数据，并在预训练过程中学习上下文表示，将医疗长文本转换为包含上下文信息的高维实数矩阵，得到 NER 任务中词汇所需的自身表示，并将其作为 BiLSTM 层的输入信息，衔接自然语言与训练模型。

3.2　BiLSTM 层

上下文信息在 NER 任务中十分重要，通过目标词语相邻的语义信息可以准确判断实体的起始和结束边界，确定实体的类型并区分同名实体。在医疗领域中，专业术语大多为长文本且上下文具有强烈的依赖性，任一相邻实体间关联的偏差都可能导致整个诊断出现失误。为了充分建模上下文信息，本文选择 BiLSTM 对知识图谱中的长句子进行编码，以同时捕获目标单词在两个方向上的上下文信息。

3.2.1　长短记忆网络(LSTM)

1997 年，Hochreiter 等[36]提出了长短记忆网络(Long Short-term Memory，LSTM)，它基于 RNN 可以实现上下文联系的性质，通过引入门控实现选择性记忆解决梯度爆炸问题。长短记忆网络的循环单元主要为 3 个：遗忘门(f_t)、输入门(i_t)和输出门(O_t)，其结构如图 2 所示。

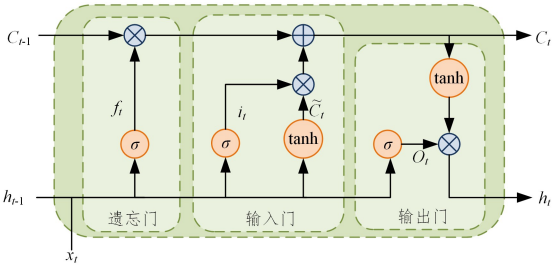

图 2　LSTM 循环单元结构图

Fig. 2　Recurrent unit structure diagram of LSTM

其中遗忘门的主要作用为基于当前时刻的输入x_t和上一个时刻的隐藏层活性值h_{t-1}，计算决定遗忘的信息量。在 t 时刻，遗忘门定义为：

$$f_t=\sigma(W_f[h_{t-1},x_t]+b_f) \tag{1}$$

LSTM 的遗忘门通过使用 *Sigmoid* 函数，将f_t映射至区间[0，1]。针对 NER 任务，f_t等于 1 时，代表当前词语与上文关联性强，因此保留上文信息的影响。

输入门的作用与遗忘门相反，它决定了候选状态$\tilde{C}_t$的保留程度，在 NER 任务中输入门用于筛选当前输入的词语信息中重要的部分。在 t 时刻i_t被定义为：

$$i_t=\sigma(W_i[h_{t-1},x_t]+b_i) \tag{2}$$

$$\tilde{C}_t=\tanh(W_c[h_{t-1},x_t]+b_c) \tag{3}$$

输出门的作用为控制在 t 时刻的细胞状态C_t保留到隐藏层h_t的信息量，O_t大小决定信息的保留程度，其中O_t的计算式如下：

$$O_t=\sigma(W_o[h_{t-1},x_t]+b_o) \tag{4}$$

LSTM 循环单元的最终输出值h_t由细胞状态C_t和候选状态$\tilde{C}_t$共同决定，而C_t由遗忘门、输入门和中间值候选状态共同决定。在 NER 任务中，该输出代表了词语对下文的影响程度，其中内部状态C_t与输出值h_t的计算式如下：

$$C_t=f_t\odot C_{t-1}+i_t\tilde{C}_t \tag{5}$$

$$h_t=O_t\odot\tanh(C_t) \tag{6}$$

综上所述，LSTM 网络设置记忆单元 C，建立更长距离的上文依赖性，但是 LSTM 虽然保留了上文记忆，但忽略了医疗文本下文的关联性。因此搭建双向长短期记忆神经网络(BiLSTM)，实现对前后信息的双向记忆，用于捕获单词在前后两个方向上的上下文信息以及单词之间的依赖关系，使得模型能够理解每个单词的上下文环境，通过考虑过去和未来的单词信息实现更精准的预测。

3.2.2　双向长短记忆网络(BiLSTM)

双向长短记忆网络(Bi-directional Long Short-Term Memory，BiLSTM)是一种运用两个 LSTM 分别处理正向和反向序列的模型。为了在 NER 任务中，保留医疗长文本对上下文信息的强依赖性，我们使用双向模型思想，针对单向 LSTM 模型只能学习上文信息的问题，提出了引入双向构建模型的想法。如图 3 所示，BiLSTM 网络结构模型所包含的 2 个 LSTM 是相互独立，通过将双向的序列输入到 2 个 LSTM 中进行特征提取，最终将 2 个特征提取结果进行拼接并输出为结果特征向量，实现对医疗文本上下文的双向信息提取，大大提高了 LSTM 模型在 NER 任务中的效果。

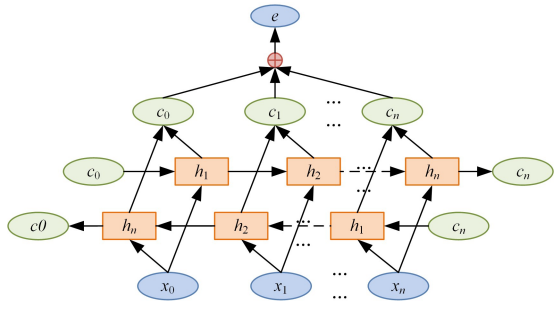

图 3　BiLSTM 模型的结构图

Fig. 3　Structure diagram of BiLSTM

3.3　CRF 层

BiLSTM 层只考虑上下文词语的相互依赖，而忽略了标签间的联系，仅用 BiLSTM 完成 NER 任务，就会出现标注连接不合理的问题，因此引入条件随机场(Conditional Random Fields，简称 CRF)完善模型。CRF 提供了序列标注的独特

功能，以 BiLSTM 层的输出概率为输入，使用一个全局正则化评分技术为每一个单词赋予标签，以反映序列单词之间的依赖关系，这确保了预测标签形成有效的实体序列，并遵守标签转换的约束。CRF 模块有助于捕获实体的整体结构和连贯性，增强模型生成一致且有意义的实体预测的能力。

CRF 的概念由 Lafferty 等[37]提出，假定存在一个由一系列变量组成的整体，当其中一个变量被标注后，该整体就成为了一个随机场。而如果在一个随机场中，某变量的赋值只与相邻的变量有关，该整体就构成了马尔可夫随机场（Markov Random Field，简称 MRF）。CRF 作为马尔可夫随机场的特例，它假设马尔可夫随机场中只有 X 和 Y 两种变量，其概率分布 $P(Y|X)$ 是条件随机场，可以定义为：

$$P(Y|X)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k \boldsymbol{t}_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l \boldsymbol{e}_l(y_i,x,i)\right) \tag{7}$$

$$Z(x)=\sum_{y}\exp\left(\sum_{i,k}\lambda_k \boldsymbol{t}_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right) \tag{8}$$

其中，$\boldsymbol{t}_k$ 代表转移矩阵；$\boldsymbol{e}_l$ 则是状态矩阵；λ_k 和 μ_l 是特征函数对应的权值；$Z(x)$ 代表所有路径和，即一种规范化因子。本文模型尝试利用 CRF 学习 NER 任务中标签序列的相互约束，从而提高上下文标签的有效性。因此，用 $P(Y|X)$ 中的 X 代表词，Y 则代表标签，对状态矩阵和转移矩阵进行应用定义。

将状态矩阵 $\boldsymbol{e}_l(y_i,x,i)$ 定义为第 i 个词 x 的标签为 y_i 的概率，转移矩阵 $\boldsymbol{t}_k(y_{i-1},y_i,x,i)$ 则表示在第 i 个位置上，标签为 y_i 的词 x 的邻位标签为 y_{i-1} 的概率，即相邻标签之间的转移概率。通过定义不难看出，状态矩阵就是 BiLSTM 层的输出，而转移矩阵就是 CRF 学习标签序列约束的渠道。根据转移矩阵和状态矩阵的定义，可以将 NER 任务中的 $P(Y|X)$ 表示为：

$$P(Y|X)=\frac{\exp(Score(X,Y))}{\sum_{\widetilde{Y}\in Y_x}\exp(Score(X,\widetilde{Y}))} \tag{9}$$

$$Score(X,Y)=\sum_{i=1}^{n}\boldsymbol{e}_{i,y_i}+\sum_{i=0}^{n}\boldsymbol{t}_{y_i,y_{i-1}} \tag{10}$$

通过式(9)、式(10)可以看出，$\exp(Score(X,Y))$ 表示文本真实的标签序列，CRF 通过概率归一化将得分转化为概率，同时不断更新转移矩阵、学习约束条件，使模型的标注结果不断接近真实序列，进而提高标签序列的有效性，建模句子中单词的上下文依赖关系。由于条件概率 $P(Y|X)$ 在模型训练过程中不断增加，无法作为衡量模型效果的损失函数进行应用，因此本文定义 CRF 的特殊损失函数为：

$$Loss=-\log P(Y|X) \tag{11}$$

4 医疗知识图谱的构建

构建知识图谱包括许多步骤，具体的流程如图 4 所示。下文将从实体细粒度化这一角度出发，围绕数据集、模型评估、医疗知识存储等方面展开描述。

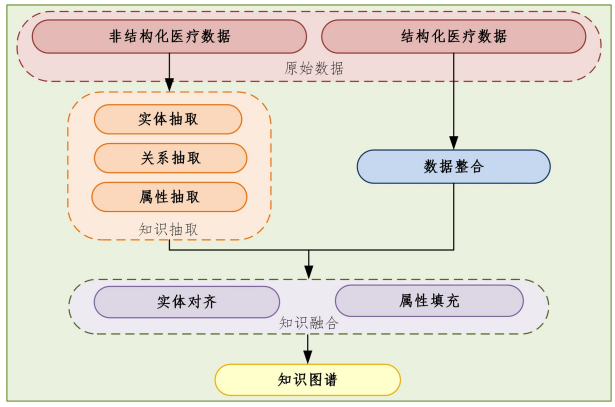

图 4 知识图谱构建流程图

Fig. 4 Flow chart of building knowledge graph

医疗知识图谱构建过程中，最核心的部分是根据评估指标对实体关系抽取模型进行优化，通过分析 F1 值调整模型参数，从而达到较好的模型效果，完善 BERT＋BiLSTM＋CRF 模型的知识抽取过程并构造医疗知识图谱。最终，将医疗知识的三元组形式存入图数据库 Neo4j 中，并充分运用 Neo4j 数据库的可视化功能，实现医疗知识图谱的展示。

4.1 数据集与模型评估

4.1.1 数据集

本文选取瑞金医院糖尿病数据集在 NER 任务中对模型进行训练，数据集主要包括 254 个训练样本，109 个测试样本，其中实体类别分为 15 种。针对序列标注问题，常见的有 BIO，BIOE，BIOES 等多种标注方案，针对该数据集，BIO 标签的标注效果已经足够优秀，故选用 BIO 标注作为本文数据集的标注方式（见图 5）。其中，标注“B-”代表实体的起始字符，标注“I-”代表实体的中间字符，标注“O-”则代表非实体字符，对标注的数据集文本进行举例。

心	血	管	疾	病	合并	糖	尿	病	口	服	降	糖	药	物
B-Disease	I-Disease	I-Disease	I-Disease	I-Disease	O	O	B-Disease	I-Disease	I-Disease	B-Method	I-Method	I-Drug	I-Drug	I-Drug	I-Drug

图 5 BIO 标注示例图

Fig. 5 Example diagram of BIO annotation

4.1.2 不同模型对比实验

上文讲述了 LSTM 和 BiLSTM 的区别与联系，下文将对比 RNN，LSTM 和 BiLSTM 的性能，并在 3 个模型的基础上分别加入 BERT 预训练和 CRF 约束层构建多个相关模型进行指标评估，展现各个模型在命名实体识别任务上效果的差异。

1）评估指标

本文采用评估指标——精确率（precision）、召回率（recall）和 F1 值模型的效果进行评价，指标的计算式如下：

$$P=\frac{TP}{TP+FP} \tag{12}$$

$$R=\frac{TP}{TP+FN} \tag{13}$$

$$F=\frac{2\times P\times R}{P+R} \tag{14}$$

其中，TP 表示实际标签为指定标签且预测正确的样本数，FN 表示实际标签为指定标签但漏报的样本数，FP 则表示实际标签并不是指定标签，但被预测为指定标签的样本数。根据以上说明，可以理解各项指标的定义：精确率表示真正预测正确的样本在预测为正样本中的占比；召回率表示真正预测正确的样本在实际正确的样本中的占比；F1 值则用于平衡

精确率和召回率，是两者的一种调和平均值。

针对 BiLSTM＋CRF 模型进行评估，各个标签的指标如表 1 所列，其中平均 F1 值到达了 0.86。由此可以看出，模型对医疗知识进行实体关系抽取展现出了较好的效果。

表 1　评估指标

Table 1　Evaluation indexes

	Precision	*Recall*	*F*1
Amount	0.92	0.91	0.92
Anatomy	0.87	0.87	0.87
Disease	0.94	0.94	0.94
Drug	0.93	0.90	0.91
Duration	0.83	0.74	0.79
Frequency	0.91	0.80	0.85
Level	0.86	0.88	0.87
Method	0.80	0.88	0.84
Operation Reason	0.97	0.94	0.95
SideEff	0.90	0.75	0.82
Symptom	0.82	0.81	0.81
Test	0.93	0.92	0.93
Test_VaLue	0.86	0.84	0.85
Treatment	0.92	0.90	0.91

2）对比实验

本文选取 BERT＋BiLSTM＋CRF 模型完成 NER 任务，为了更直观地体现模型优势，在相同的实验环境下对比了多个模型的召回率、准确率和 F1 值。通过瑞金医疗数据库训练 3 个模型，指标对比如表 2 所列。可以明显看出，BiLSTM 在 3 项指标内都由于 RNN 和 LSTM 证明了双向提取上下文信息的有效性，同时 3 个模型引入 CRF 性能都有所提升，因此 CRF 选取概率最高标签的能力可以很好地适用于 NER 任务。对比 BERT，BERT＋CRF 和 BERT＋BiLSTM＋CRF 模型可知，BERT 自身已经可以很好地适用于 NER 任务，引入 BERT 更能增强模型性能。对比所用模型的指标，本文模型相比性能最好的模型，F1 值高出 0.02，该对比实验也为模型选取提供了更加可靠的依据。

表 2　评估指标对比表

Table 2　Comparison of evaluation indicators

	Precision	Recall	F1
RNN	0.24	0.37	0.28
LSTM	0.35	0.48	0.40
BiLSTM	0.63	0.71	0.67
BERT	0.77	0.78	0.77
RNN＋CRF	0.76	0.66	0.71
LSTM＋CRF	0.82	0.74	0.78
BiLSTM＋CRF	0.89	0.86	0.87
BERT＋CRF	0.83	0.79	0.79
BERT＋BiLSTM＋CRF	0.90	0.89	0.89

4.1.3　超参数分析

对于命名实体识别模型而言，参数的变化时刻影响着实验，BERT＋BiLSTM＋CRF 模型用到的相关参数如表 3 所列。

表 3　参数列表

Table 3　Parameters

参数名	参数含义
epoch	迭代次数
lr	学习率
batch_size	批处理数目

表 3 中，epoch 表示模型训练过程中迭代学习的次数，本模型不同 epoch 的指标对比图如图 6 所示，可见设置 epoch＝50，在不浪费学习资源的同时达到最优效果；lr 即 Learning Rate，表示模型训练的迭代步长，其大小决定了模型是否能收敛于最优解；batch_size 则表示每轮训练在训练集中抽取的样本个数。下文将主要针对 lr 展开模型的对比优化实验。

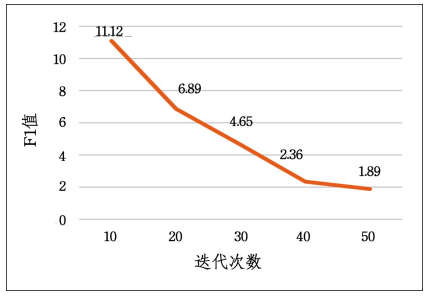

图 6　epoch 调节对比图

Fig. 6　Comparison chart of adjusting epoch

学习率作为影响模型效果的重要参数，用于控制迭代的步长。当学习率过小时，训练模型的效率降低，浪费时间；学习率过大又会导致学习效果差，重要的信息会被遗漏。因此通过对比不同的学习率，并找到最适宜的数值是优化模型过程中非常必要的一步。本文选取了 4 个不同的学习率，F1 值对比如图 7 所示，可以看出 $lr=10^{-3}$ 时模型效果最好。

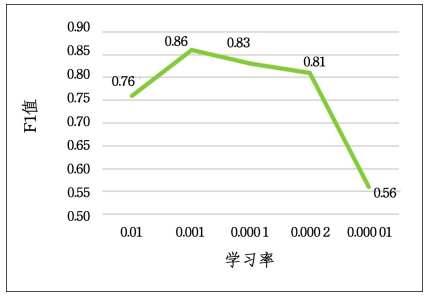

图 7　学习率调节对比图

Fig. 7　Comparison chart of adjusting learning rate

4.2　医疗知识图谱可视化

本文主要针对网络上爬取到的大规模医疗知识库，运用训练好的 BERT＋BiLSTM＋CRF 模型，对长文本节点进行实体抽取，实现细粒度知识图谱的构建。将文本以三元组格式存储至 Neo4j 数据库中并实现医疗知识图谱的可视化。

4.2.1　Neo4j 知识存储

目前，知识图谱的存储方式众多，Neo4j 良好的兼容性使其在众多数据库中脱颖而出，该数据库提供 Cypher 查询语言，它类似 SQL 语言却能实现更复杂的功能，我们可以通过编写查询、更新语句对图形数据关系进行探索和调整。基于其强大的功能，本文选取 Neo4j 进行医疗知识存储。

Neo4j 是目前最流行的图存储数据库，对于知识图谱而言，节点和边作为图的基本元素，可用于表示实体和关系。又由于医疗知识的数量庞大，关系错综复杂，因此选用三元组的形式，存储实体关系信息，并保存至 Neo4j 数据库中。三元组表示关系的格式可以举例为：〈放射性肺炎，并发症，肺气肿〉〈支气管炎，症状，喘息〉〈百日咳，检查，血常规〉〈苯中毒，宜吃，鸡蛋〉等。

4.2.2　知识图谱规模分析与可视化

本文构建的知识图谱中，包含 10 种医疗实体类别：检查

项目、医疗科室、疾病名称、药物、食物、在售药品、疾病症状、病因、病变部位和持续时间等。其中各类别实体数量如表4所列，形成了13.8万数量级的大规模知识库。同时，我们抽取了13种实体关系类别：所属科室、常用药品、宜吃食物、药品在售药品、所需检查、忌吃食物、推荐药品、推荐食谱、症状、并发疾病、疾病诱因、发病部位、疾病持续时间等。其中各类别关系数量如表5所列，总计达到30万关系量级。

根据上述规模分析可知图谱规模庞大，为了方便展示成果，运用Cypher查询语句在Neo4j中查询以疾病“二硫化碳中毒”为中心的知识图谱关系网络，并选取部分有代表性的节点进行展示，如图8所示。

表4 实体类型

Table 4 Entity types

实体名称	实体含义	实体数量	举例
Check	检查项目	3353	脑血流显像
Department	医疗科目	54	整形美容科
Disease	疾病名称	8807	血栓闭塞性脉管炎
Drug	药物	3828	京万红痔疮膏
Food	食物	4870	竹笋炖羊肉
Producer	在售药品	17210	紫荆花制药血尿胶囊
Symptom	疾病症状	5989	传导性耳鸣
Cause	病因	9901	感染
Anatomy	病变部位	83259	中性粒细胞
Duration	持续时间	835	1～2周
Total	总计	138106	—

表5 关系类型

Table 5 Relationship types

关系名称	关系含义	关系数量	举例
belongs_to	所属科室	8844	〈妇科，属于，妇产科〉
common_drug	常见药品	14599	〈胃肠道功能紊乱，常用，香砂养胃丸〉
do_eat	适宜食物	22198	〈大叶性肺炎，宜吃，熟栗子〉
drugs_of	在售药品	17315	〈明目地黄丸，在售，北京御生堂明目地黄丸〉
need_check	所需检查	38994	〈肺泡蛋白质沉积症，所需检查，肺活检〉
no_eat	不适宜食物	22247	〈成人呼吸窘迫综合征，忌吃，啤酒〉
recommand_drug	建议药品	59488	〈喘息样支气管炎，推荐用药，小青龙颗粒〉
reason	疾病诱因	32359	〈百日咳，诱因，百日咳杆菌〉
disease_anatomy	发病部位	83259	〈肺念珠菌病，发病细胞，肺泡〉
disease_duration	疾病持续时间	835	〈中央晕轮状脉络膜萎缩，持续，1～3个月〉
Total	总计	326508	—

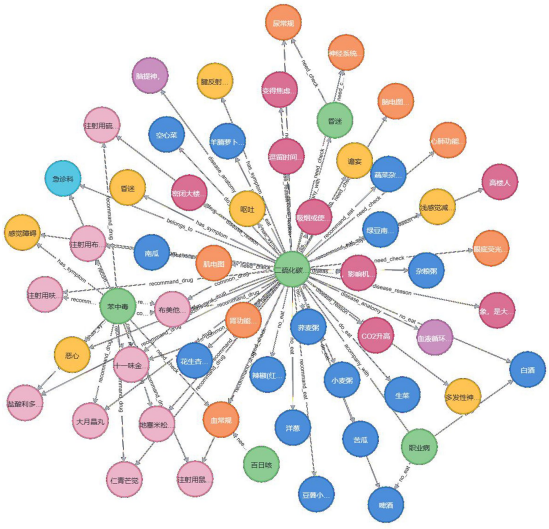

图8 部分知识图谱可视化

Fig. 8 Visualization of partial knowledge graph

结束语 从医疗知识图谱庞大的规模中不难看出，其在现实生活中的应用性很强，医疗检索、医疗问答系统等智能化应用的普及，也让本文的研究有了更高的价值。构建细粒度的知识图谱使得长文本的医疗术语得到了更加细致的划分，并且使长文本中实体的内在关联有了直观的展现，提高了可理解性。本文选取BERT+BiLSTM+CRF联合抽取模型完成了医疗实体细粒度的目标，并最终在大规模医疗知识图谱上完成可视化。

经过相关指标评估，对比财经、新闻、生物医学等其他领域的命名实体识别模型指标可知，BERT+BiLSTM+CRF在医疗领域并未达到最优效果，且由于医疗术语的专业性，目前医疗数据集的种类匮乏，不足以支撑完成更加精准且便于应用的细粒度知识图谱，因此本文仅细分出10种实体类别和13种实体关系，相比其他领域的细粒度任务而言比较匮乏。在未来的工作中会尝试探索覆盖面更广、种类更繁多的医疗知识库，同时计划引入新兴的深度学习相关模型，以提高BERT+BiLSTM+CRF模型的效果，从而提供更好的知识图谱智能化应用体验。

参 考 文 献

[1] SINGHAL A. Introducing the knowledge graph: things, not strings, May 2012[OL]. http://googleblog.blogspot.ie/2012/05/introducing-knowledgegraph-things-not.html, 2012.

[2] MINGYU C, QINGQING L, ZHIHAO Y, et al. A Question Answering System for Primary Liver Cancer Based on Knowledge-Graph [J]. Journal of Chinese Information Processing, 2019, 33(6): 88-93.

[3] ZHAO C, JIANG J, GUAN Y, et al. EMR-based medical knowledge representation and inference via Markov random fields and distributed representation learning[J]. Artificial Intelligence in Medicine, 2018, 87: 49-59.

[4] ZHENG Z, LIU Y, ZHANG Y, et al. TCMKG: A deep learning based traditional Chinese medicine knowledge graph platform [C]// 2020 IEEE International Conference on Knowledge Graph (ICKG). IEEE, 2020: 560-564.

[5] ROTMENSCH M, HALPERN Y, TLIMAT A, et al. Learning a health knowledge graph from electronic medical records [J]. Scientific Reports, 2017, 7(1): 1-11.

[6] ODMAA B, YUNFEI Y, ZHIFANG S, et al. Preliminary Study on the Construction of Chinese Medical Knowledge Graph [J]. Journal of Chinese Information Processing, 2019, 33(10): 1-9.

[7] LIU X, WEI F, ZHANG S, et al. Named entity recognition for tweets[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2013, 4(1): 1-15.

[8] QIU J, ZHOU Y, WANG Q, et al. Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field[J]. IEEE Transactions on Nano-

Bioscience，2019，12：306-315.
[9] LEHMANN J，ISELE R，JAKOB M，et al. DBpedia：a large-scale，multilingual knowledge base extracted from wikipedia [J]. Semantic Web，2015，6(2)：167-195.
[10] BOLLACKER K D，COOK R P，TUFTS P. Freebase：a shared database of structured general human knowledge[C]// Proceedings of the 22nd AAAI Conference on Artificial Intelligence，Vancouver. Menlo Park：AAAI，2007：1962-1963.
[11] SUCHANEK F M，KASNECI G，WEIKUM G. Yago：a large ontology from Wikipedia and WordNet[J]. Journal of Web Semantics，2008，6(3)：203-217.
[12] XU B，LIANG J，XIE C，et al. CN-DBpedia2：an extraction and verification framework for enriching Chinese encyclopedia knowledge base[J]. Data Intelligence，2019，1(3)：271-288.
[13] CHEN H J，HU N，QI G L，et al. OpenKG chain：a blockchain infrastructure for open knowledge graphs[J]. Data Intelligence，2021，3(2)：205-227.
[14] JIXIANG Z，XIANGSEN Z，CHANGXU W，et al. Survey of Knowledge Graph Construction Techniques [J]. Computer Engineering，2022，48(3)：23-37.
[15] RAU L F. Extracting company names from text [C] // Proceedings the Seventh IEEE Conference on Artificial Intelligence Application. IEEE Computer Society，1991：29-32.
[16] SCHUTZ A，BUITELAAR P. RelExt：a tool for relation extraction from text in ontology extension[C]// Proceedings of the 4th International Semantic Web Conference. Berlin，Germany：Springer，2005：593-606.
[17] QUIMBAYA A P，MÚNERA A S，RIVERA R A G，et al. Named entity recognition over electronic health records through a combined dictionary-based approach[J]. Procedia Computer Science，2016，100：55-61.
[18] WANG H，ZHANG W，ZENG Q，et al. Extracting important information from Chinese Operation Notes with natural language processing methods[J]. Journal of Biomedical Informatics，2014，48：130-136.
[19] PONOMAREVA N，PLA F，MOLINAA，et al. Biomedical named entity recognition：apoorknowledge HMM-based approach[C]// Natural Language Processing and Information Systems：12th International Conference on Applications of Natural Language to Information Systems(NLDB 2007). Paris，France，Springer Berlin Heidelberg，2007：382-387.
[20] XU R，LI L，WANG Q Q. dRiskKB：a large-scale disease-disease risk relationship knowledge base constructed from biomedical text[J]. BMC Bioinformatics，2014，15(1)：1-13.
[21] SUI M S，CUI L. Extracting chemical and disease named entities with multiple-feature CRF model[J]. New Technology of Library and Information Service，2016(10)：91-97.
[22] FAN YY，LI Z M. Research and application progress of Chinese medical knowledge garph[J]. Journal of Frontiers of Computer Science and Technology，2022，16(10)：2219-2233.
[23] LI W，ZHAO D Z，LI B，et al. Combining CRF and rule based medical named entity recognition[J]. Application Research of Computers，2015，32(4)：1082-1086.
[24] FENG J，HUANG M，ZHAO L，et al. Reinforcement learning for relation classification from noisy data[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence，New Orleans. Menlo Park：AAAI，2018：5779-5786.
[25] JI G，LIU K，HE S，et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions[C]// Proceedings of the 31st AAAI Conference on Artificial Intelligence，San Francisco. Menlo Park：AAAI，2017：3060-3066.
[26] WU X P，ZHANG Q，ZHAO F，et al. Entity relation extraction method for guidelines of cardiovascular disease based on bidirectional encoder representation from transformers[J]. Journal of Computer Applications，2021，41(1)：145-149.
[27] DING Z Y，YANG Z H，LUO L，et al. A Chinese biomedical entity relationship extraction system based on deep learning[J]. Journal of Chinese Information Processing，2021，35(5)：70-76.
[28] UKOV-GREGORI A，BACHRACH Y，COOPE S. Named entity recognition with parallel recurrent neural networks[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2：Short Papers). 2018：69-74.
[29] GAO F，YANG J X，GU J G. Extraction of diagnosis and treatment relationship based on fusion relation discovery words and deep learning[J]. Computer Applications and Software，2021，38(12)：168-173.
[30] ZENG D，LIU K，LAI S，et al. Relation classification via convolutional deep neural network[C]// 25th International Conference on Computational Linguistics：Technical Papers(COLING 2014). 2014：2335-2344.
[31] YAN X，DUAN Y X，ZHANG Z H. Entity relationship extraction fusing self-attention mechanism and CNN[J]. Computer Engineering & Science，2020，42(11)：2059-2066.
[32] ZHANG Y，GAO D L，GONG D W，et al. Attention graph long short term memory neural network for relation extraction[J]. CAAI Transactions on Intelligent Systems，2021，16(3)：518-527.
[33] KATIYAR A，CARDIE C. Going out on a limb：Joint extraction of entity mentions and relations without dependency trees[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics(Volume 1：Long Papers). 2017：917-928.
[34] XIAO J，ZHOU Z. Chapter-level entity relationship extraction method based on joint learning[C]// 2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics(IHMSC). IEEE，2020，1：75-78.
[35] MIWA M，BANSAL M. End-to-end relation extraction using lstms on sequences and tree structures[J]. arXiv：1601. 00770，2016.
[36] HOCHREITER S，SCHMIDHUBER J. Long short-term memory[J]. Neural Computation，1997，9(8)：1735-1780.
[37] LAFFERTY J，MCCALLUM A，PEREIRA F C N. Conditional random fields：Probabilistic models for segmenting and labeling sequence data[J]. 2001.

WANG Yuhan，born in 2001，postgraduate. Her main research interests include machine learning and deep learning.

WANG Ying，born in 1981，Ph.D，professor，Ph.D supervisor，is a member of CCF(No. 183695). Her main research interests include machine learning，social networks，data mining，and search engines.