计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200100036

面向中文电子病历的多粒度医疗实体识别

周晓进[1]　徐陈铭[2]　阮　彤[1]
1 华东理工大学信息科学与工程学院　上海 200237
2 华东理工大学理学院　上海 200237
(zhouxiaojin@mail.ecust.edu.cn)

摘　要　在现有的面向中文临床电子病历的命名实体识别任务中,实体标注粒度通常过细或过粗,过细的标注结果难以找到实际应用场景,而过粗的标注结果通常需要在进行复杂的处理后,才能明确实体的规范形式和语义类型,以便于后续的数据挖掘应用。为简化处理步骤,根据常见的7类粗粒度临床实体的特点,定义了用以解释粗粒度实体的9类细粒度解析实体。同时,针对多粒度实体的特点,提出了基于多任务学习和自注意力机制的多粒度临床实体识别模型,并在真实的医院电子病历库中标注了5000条包含多粒度实体的文本以验证模型的效果。实验结果表明,该模型优于主流的序列标注模型,在粗、细粒度实体识别任务中,两者的$F1$值分别达到了92.88和85.48。

关键词:电子病历;多粒度实体识别;多任务学习
中图法分类号　TP391

Multi-granularity Medical Entity Recognition for Chinese Electronic Medical Records

ZHOU Xiao-jin[1], XU Chen-ming[2] and RUAN Tong[1]
1 School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China
2 School of Science, East China University of Science and Technology, Shanghai 200237, China

Abstract　In the existing named entity recognition task for Chinese clinical electronic medical records, the granularity of annotation is usually too fine or too coarse, and it is difficult to find actual application scenarios for the too thin annotation results while the too thick annotation results usually need complex post-processing steps to clarify the standard form and the semantic type of entities, so as to facilitate subsequent data mining applications. In order to simplify post-processing steps, 9 kinds of fine-grained analytical entities are defined to explain coarse-grained entities according to characteristics of 7 common coarse-grained clinical entities. Besides, according to characteristics of multi-granularity entities, a multi granularity clinical entity recognition model based on multi-task learning and self-attention mechanism is proposed, and 5000 texts containing multi-granular entities are annotated on real hospital electronic medical records to verify the model. Experiment results show that this model outperforms the mainstream sequence labeling model. In the task of coarse and fine granularity entity recognition, their $F1$ scores reach 92.88 and 85.48, respectively.

Keywords　Electronic medical records, Multi-granularity named entity recognition, Multi-task learning

1　引言

电子病历(Electronic Medical Records,EMR)是重要的个人健康记录以及医学科研材料。人们期待利用EMR积累的数据,来辅助临床决策、医学科研以及公共卫生管理等医学课题的发展。但是,EMR中存在大量的非结构化以及半结构化数据,这些数据无法被计算机直接处理,因此阻碍了电子病历数据发挥充分的作用。其中,命名实体识别被认为是电子病历结构化的基础。

中文电子病历的命名实体识别主要关注EMR记录中的临床实体,即患者临床记录中的主诉症状、疾病诊断、检验指标、检查方法、治疗方案和身体部位等。与形式相对固定的通用领域的命名实体不同,临床医疗实体为了能够充分表现患者的情况,有着更丰富且自由的表达形式和用语方式。临床实体比专有名词更像是短语,其包含复合的结构以及多种表达形式。例如,当患者自诉"头痛"时,临床实体有"头有点疼"

到稿日期:2020-01-05　返修日期:2020-05-14　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:"精准医学研究"重大专项项目(2018YFC0910500);国家自然科学基金项目(61772201)
This work was supported by the Major Special Project of Precision Medical Research(2018YFC0910500) and National Natural Science Foundation of China(61772201).
通信作者:阮彤(ruantong@ecust.edu.cn)

“头很疼”“头痛”等多种表达形式；在医生的诊断记录中，“无肺恶性肿瘤”包含“不存在肺部恶性肿瘤”和“肿瘤观察位置为肺部”这两种语义信息。

在现有的中文临床实体数据集中，有些数据集标注实体的粒度过细，例如 CCKS2017 将“无腹壁静脉曲张”中的“腹壁”标为部位，将“静脉曲张”标为症状，但和“静脉曲张”相关的语义信息如“患者未患有”和“发生在腹壁位置”却都无法从实体中得到。另一些数据集采用了保留语义的标注方式，但是标注实体的粒度过粗，例如 He 等[1]将“无口角或肢体的抽动”标注为一个症状实体，在实际应用时需要对这种标注方式得到的实体进行更麻烦的后处理，下文将具体说明后处理的必要性和面临的困难。

针对不同的应用，临床实体的多种表达形式有时需要被规范化为同一个标准名称，有时临床实体又需要不同表达形式中的特定信息。例如，在统计患有“头部疼痛”的患者时，需要将“头部酸痛”“头部钝痛”“头部胀痛”统一成它们的上位词“头部疼痛”的形式；而在查询患者的头痛类型时，又需要用到“酸痛”“钝痛”“胀痛”等信息。因此，常采用通过制定规则来进行术语匹配的方式，对临床实体进行后处理。但是，医学词典常常面临着术语缺乏、一词多义、多类型的问题，并且固定的规则会导致错误的匹配结果。本文将有短语特性的临床实体称为粗粒度临床实体，如果要让机器分辨粗粒度临床实体的信息直接的方式是建立规则对其进行后处理。但是事实上，粗粒度实体的语义信息可以通过粗粒度实体内部的构成来获得，且粗粒度实体内部的中心词、部位词、等级词也能作为实体规范化的强特征。因此，本文将粗粒度实体的细粒度组成称为细粒度解析实体，将它和非短语形式的细粒度临床实体统称为细粒度实体。本文在粗粒度标注方式的基础上，进一步对其中相关信息的构成元素进行了细粒度标注，识别出病历文本中的细粒度解析实体能够减少后处理的工作量。以细粒度解析实体元组的形式表示粗粒度临床实体，有助于实现机器对粗粒度临床实体的理解和规范化。

本文在某三甲医院的大肠癌电子病历上，人工标注了 5000 条包含临床和解析实体的文本，构建了一份同时包含多粒度医疗实体的数据集。

近年来，深度学习凭借其在各类任务中的出色表现，成为了命名实体识别的主流方法。深度学习的优点在于不需要额外添加特征，可以直接通过神经网络来学习字词本身和上下文的语义信息。此外，考虑到粗细粒度实体之间存在相互补充、相互说明的关系，本文提出了融入自注意力机制的多粒度实体识别模型来同时对粗细实体进行识别。

2 相关工作

2.1 基于规则和词典的方法

生物医学领域的命名实体识别任务主要是识别领域文本中的疾病、症状、检查、基因等。早年的研究主要使用基于规则和词典的方法，这类方法依赖人工构建的启发式规则和生物医学词典来实现命名实体识别。Fukuda 等[2]利用蛋白质名称的结构以及用词特征，来识别科研文献中新提出的蛋白质实体。Friedman 等[3]基于多个通用领域的知识库，构建了三阶段文本解析器，其通过对放射科文本进行句法解析来定位实体，对实体进行标准化来规范实体形式，将实体和知识库中的术语对齐以最终确定实体。Song 等[4]提出了利用多源词典的命名实体识别方法。

但是，生物医学领域的子领域众多，命名实体的形式也是多种多样，而基于规则和词典的方法普遍泛化性差，且人工成本高。因此，越来越多的研究逐渐开始关注使用基于统计的机器学习方法来进行生物医学领域的命名实体识别。

2.2 基于统计学习的方法

基于统计学习的方法将命名实体识别任务看作序列标注任务，主要使用概率图模型为输入序列的每个字符打上其在实体中的位置标签和其对应的实体的类型标签。Zhao[5]提出了利用词相似度进行数据平滑化的隐马尔可夫方法(Hidden Markov Model，HMM)。Finkel 等[6]选取词的词形、前后词、前后缀、语义触发词等特征信息，训练最大熵马尔可夫模型(Maximum Entropy Markov Models，MEMM)，并根据实体在互联网等外部数据源中出现的频率来调整实体在任务中的受重视程度。Settles[7]针对基因和蛋白质识别任务，提出了使用启发式生物医疗特征的条件随机场方法(Conditional Random Field，CRF)。

传统的统计学习方法仍然需要通过人工方式来进行特征选择，成本较高，并且因为传统统计学习模型对文本的语义编码的能力较差，模型效果严重依赖训练数据，所以其难以应对罕见词和未登录词。因此，近年来使用深度学习模型来进行命名实体识别成为了研究的主流。

2.3 基于深度学习的方法

Huang 等[8]提出了用于预测序列标签的后接条件随机场的双向长短期记忆网络(Bi-LSTM-CRF)。Griden[9]首先在生物医学领域使用 Bi-LSTM-CRF 模型进行命名实体识别。Dang 等[10]在 Bi-LSTM-CRF 模型的基础上，利用语言学信息对词向量进行微调。Liu 等[11]在 Bi-LSTM-CRF 模型上后接了一个多通道卷积神经网络，并将词的词法和形态学特征也作为实体识别的信息。

除 Bi-LSTM-CRF 模型外，也有相关研究将其他深度学习模型应用到生物医疗领域。Giorgi 等[12]在大规模有噪声语料上训练模型，并将模型迁移到小规模可靠的语料上。Qiu 等[13]利用残差膨胀卷积模型来快速高效地进行医疗领域的命名实体识别。Wang 等[14]提出 3 种在深度模型中加入词典特征的方式，将有用的外部知识纳入深度模型中。

2.4 多任务学习

多任务学习同时对多个任务进行建模。神经网络一般通过共享权重的方式，让底层网络同时兼顾不同的任务，以提取更具有代表性的低级特征；同时，各个任务也拥有各自的权重来提取与自身任务相关的特征。研究表明，多任务学习的效果优于单任务学习，其能够极大地提升模型的泛化性能，尤其是在多个任务直接相关性较大的情况下。Luong 等[15]于 2015 年首次把多任务学习应用到序列到序列的任务中，取得了不错的效果。

3 细粒度解析实体介绍

本文将中文临床实体分为粗粒度临床实体和细粒度临床实体这两类。定义粗粒度临床实体为有可组合性和可扩展性的临床实体，如疾病、症状等；细粒度临床实体的定义有着相对固定的形式和范围，如检验指标名、药品名等。其中，粗粒度临床实体又可以根据其短语结构进行分解，因此本文将粗粒度临床实体的分解结果称为细粒度解析实体。

本文参考卫生部发布的电子病历数据组与数据元标准，以及现有的工作[16]，并结合在真实病历文本上的实际标注经验，总结了自诉症状、体格检查、检查项目、疾病诊断、疾病史、手术名、手术史这 7 类粗粒度临床实体的结构特征。若按照细粒度的语义对实体进行分类，则可分为表示否定、表示肯定、表示可能、表示颜色等几十种类型。考虑到过多的实体类型会给标注工作和模型带来很大的负担，因此本文将细粒度解析实体按功能分类，定义了 9 类细粒度解析实体，其中起到了相似功能的词被归为同一类。图 1 给出了一个结合粗细粒度标注的实例。

此外，不具有明显复合结构的细粒度临床实体为检验指标、检验结果、药品名、药品用量、药品用法。细粒度解析实体具体如表 1 所列。

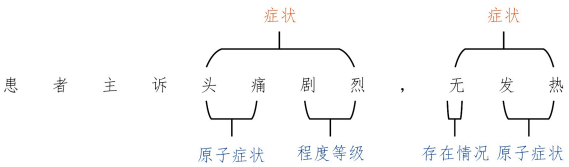


图 1 标注样例

Fig. 1 Example of annotation

表 1 细粒度实体定义

Table 1 Definition of fine-grained entity

细粒度实体	定义	例子
存在情况	表示某种事物是否存在，或是否可能存在的词	[难以]入睡
情景限定	表示某种特定情景的词语	[餐后]腹痛
程度等级	表示症状或疾病严重程度或医学分级的词语	[二级]高血压
身体部位	表示身体部位的词语	[上腹部]疼痛
原子症状	拆分后不再是原症状的症状词	反复[咳嗽]
原子疾病	拆分后不再是原疾病的疾病词	2 型[糖尿病]
时间点	病历中某一时刻的时间	[刻下]乏力
时间段	病历中描述持续一定时长的时间	低烧持续[一周]

4 多粒度实体识别模型

针对两种不同粒度的实体，我们提出了基于多任务学习并融合注意力机制的多粒度实体识别模型，该模型由 4 个部分组成，分别是嵌入层、Bi-LSTM 层、Self-Attention 层和 CRF 层。该模型首先通过共享嵌入层和 Bi-LSTM 层来学习公共的上下文特征，然后将实体传入两个平行的 Self-Attention 层分别进行过滤，以提取粒度相关的特征，最后将提取的粒度相关的特征和公共特征相加，再通过两个平行的 CRF 层来得到预测的粗粒度标签序列和细粒度标签序列。具体的网络模型如图 2 所示。

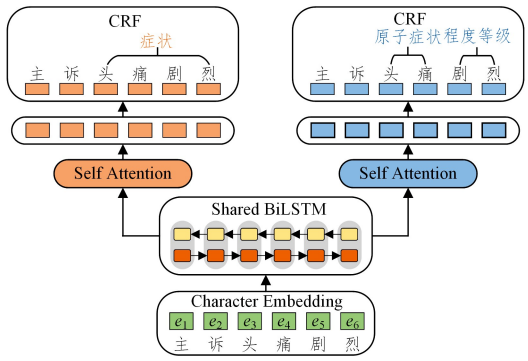


图 2 模型架构图

Fig. 2 Model architecture diagram

4.1 嵌入层

嵌入层的作用是将文字转换成对应的字向量形式，初始字向量可以通过使用文本语料训练得到或随机初始化得到，本文采用 word2vec[17-18] 来学习初始字向量。具体而言，输入固定长度为 n 的电子病历文字序列 $X=[x_1,x_2,\cdots,x_n]$ 到嵌入层，将序列中的每个汉字 x_i 转换成对应的字向量 $\boldsymbol{e}_i\in\mathbb{R}^n$，最终将文字序列 X 转换成矩阵 $\boldsymbol{E}=[\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n]$ 的形式，并将 $\boldsymbol{E}$ 传入 Bi-LSTM 层。

4.2 Bi-LSTM 层

LSTM 是一种特殊的循环神经网络（Recurrent Neural Network，RNN），它继承了 RNN 在处理序列问题时的“记忆”优势，也就是网络在时刻 t 的输出与前 $t-1$ 时刻的输入相关。同时，LSTM 通过引入门控机制缓解了传统 RNN 容易梯度消失或爆炸的问题。LSTM 的具体计算过程如下：

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_i\boldsymbol{e}_t+\boldsymbol{U}_i\boldsymbol{h}_{t-1}+\boldsymbol{b}_i) \tag{1}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f\boldsymbol{e}_t+\boldsymbol{U}_f\boldsymbol{h}_{t-1}+\boldsymbol{b}_f) \tag{2}$$

$$\tilde{\boldsymbol{c}}_t=tanh(\boldsymbol{W}_c\boldsymbol{e}_t+\boldsymbol{U}_c\boldsymbol{h}_{t-1}+\boldsymbol{b}_c) \tag{3}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\tilde{\boldsymbol{c}}_{t-1} \tag{4}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_o\boldsymbol{e}_t+\boldsymbol{U}_o\boldsymbol{h}_{t-1}+\boldsymbol{b}_o) \tag{5}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_t) \tag{6}$$

其中，$\boldsymbol{e}_t,\boldsymbol{h}_t\in\mathbb{R}^d$ 分别是网络在时刻 t 的输入和输出向量；$\boldsymbol{i}_t,\boldsymbol{f}_t,\boldsymbol{o}_t\in\mathbb{R}^d$ 分别是时刻 t 的输入门向量、遗忘门向量、输出门向量；$\tilde{\boldsymbol{c}}_t,\boldsymbol{c}_t$ 是计算的中间结果；$\boldsymbol{W}\in\mathbb{R}^{d\times n},\boldsymbol{U}\in\mathbb{R}^{d\times d},\boldsymbol{b}\in\mathbb{R}^d$ 是可训练的参数，不同的下标表示参数对应不同的门；σ，tanh，$\odot$ 分别表示 sigmoid 函数、双曲正切函数、逐元素相乘。特别地，当 $t=1$ 时，$\boldsymbol{h}_0,\tilde{\boldsymbol{c}}_0=0$。

Bi-LSTM 层由一个正向 LSTM 和一个逆向 LSTM 组成。在 t 时刻，向量 $\boldsymbol{e}_t$ 自左向右地传入正向 LSTM，输出记为 $\overrightarrow{\boldsymbol{h}_t}$；逆向 LSTM 自右向左地处理向量 $\boldsymbol{e}_t$，输出记为 $\overleftarrow{\boldsymbol{h}_t}$。Bi-LSTM 在时刻 t 的输出，为正向 LSTM 和逆向 LSTM 输出的拼接 $[\overrightarrow{\boldsymbol{h}_t};\overleftarrow{\boldsymbol{h}_t}]$。

4.3 Self-Attention 层

Self-Attention 最早由谷歌翻译团队[19]于 2017 年提出。注意力机制的核心思想是在翻译任务中，给模型输入源语句和目标语句，让模型能够学习出目标语句中的每个字或词对源语句中的每个字或词的关注度。而自注意力实际上就是把

目标语句替换为源语句，让模型学习句子内部中的字词对其他字词的关注度。本文采用 Self-Attention 的原因是单纯地在模型中采用权重共享只能提取公共特征，且会损失不同粒度实体的独特性。因为这两类实体都共用了完整的公共特征，所以引入 Self-Attention 机制可以让不同粒度的实体学习到对公共特征的关注度，从而体现差异性。形式上，Self-Attention 的具体计算过程如下：

$$attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)V \tag{7}$$

其中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 分别代表查询向量矩阵、键值向量矩阵和值向量矩阵，且 $\boldsymbol{Q}\in\mathbb{R}^{m\times d_k}$，$\boldsymbol{K}\in\mathbb{R}^{m\times d_k}$，$V\in\mathbb{R}^{m\times d_v}$，$m$ 表示输入语句的长度，d_k 和 d_v 分别表示查询向量和值向量的维度，$\sqrt{d_k}$ 起到一个缩放作用。在模型中，我们令 $\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V}=Bi\text{-}LSTM$ 的输出。

在此基础上，为了增加 Self-Attention 的表示能力，我们还使用了多头注意力机制，具体计算过程如下：

$$\boldsymbol{head}_i=attention(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V)\boldsymbol{W}^O \tag{8}$$

$$multiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=concat(\boldsymbol{head}_1,\boldsymbol{head}_2,\cdots,\boldsymbol{head}_h) \tag{9}$$

其中，$\boldsymbol{W}_i^Q\in\mathbb{R}^{d_{\text{model}}\times d_k}$，$\boldsymbol{W}_i^K\in\mathbb{R}^{d_{\text{model}}\times d_k}$，$\boldsymbol{W}_i^V\in\mathbb{R}^{d_{\text{model}}\times d_v}$，$\boldsymbol{W}^O\in\mathbb{R}^{hd_v\times d_{\text{model}}}$，$d_{\text{model}}$ 表示输入的特征向量的维度，h 表示相互平行的自注意力层的个数，并且 $d_k=d_v=d_{\text{model}}/h$。

4.4 Skip-Connect 机制

残差连接最早出现在 He 等[20]于 2015 年提出的残差网络中，原始的残差单元由一个卷积层、批处理归一化层和非线性激活函数 ReLU 层堆叠而成，本文中残差单元就是由采用多头注意力的 Self-Attention 层组成的。残差连接的作用是加速深度网络的训练，并且在很大程度上缓解了由于网络层数过深导致的梯度消失和梯度爆炸问题。具体计算过程如下：

$$skipConnect(\boldsymbol{x})=\boldsymbol{x}+multiHead(\boldsymbol{x},\boldsymbol{x},\boldsymbol{x}) \tag{10}$$

其中，$\boldsymbol{x}$ 表示经过 Bi-LSTM 层输出的隐藏向量。

4.5 CRF 层

在一个标签序列中，标签之间存在着相互依赖、相互约束的关系，如 I(Inside)标签应该跟在 B(Begin)标签或 I 标签之后、一组实体标签的实体类型应该尽可能相同等。CRF 能通过转移矩阵刻画来输出标签之间的依赖关系，并得到全局最优的标签序列。

CRF 层基于文字序列 $[x]_1^T$ 给标签序列 $[i]_1^T$ 打分，根据分数来评价标签序列的优劣。标签序列的得分由发射分数和转移分数两部分组成。定义 $\boldsymbol{M}([x]_1^T)$ 为文字序列 $[x]_1^T$ 的发射概率矩阵，$[\boldsymbol{M}([x]_1^T)]_{[i]_t,t}$ 是第 t 个字预测为标签 $[i]_t$ 的概率，发射分数是序列 $[x]_1^T$ 的每个字预测为 $[i]_1^T$ 中对应标签的概率和；$\boldsymbol{A}$ 为标签转移概率矩阵，$\boldsymbol{A}_{i,j}$ 表示标签 i 转移到标签 j 的概率，转移分数是标签序列 $[i]_1^T$ 中相邻标签的转移概率和。发射概率由残差连接后得到，转移概率是 *CRF* 层训练更新的参数。总的标签打分公式如下：

$$S([x]_1^T,[i]_1^T)=\sum_{t=1}^{T}(\boldsymbol{A}_{[i]_{t-1},[i]_t}+[\boldsymbol{M}([x]_1^T)]_{[i]_t,t}) \tag{11}$$

CRF 层的优化目标是，真实标签序列的得分占总体的比重越大越好。令 $[y]_1^T$ 是真实标签序列，$[j]_1^T$ 取遍所有可能的标签序列，真实标签序列占总体的比重公式如下：

$$p([y]_1^T|[x]_1^T)=\frac{e^{S([x]_1^T,[y]_1^T)}}{\sum_j e^{S([x]_1^T,[j]_1^T)}} \tag{12}$$

损失函数一般用优化目标的负对数似然估计，具体公式如下：

$$Loss=-S([x]_1^T,[y]_1^T)+log(\sum_j e^{s([x]_1^T,[j]_1^T)}) \tag{13}$$

4.6 多任务学习

传统的 NER 模型都是单输入单输出，针对两种粒度的实体，传统的做法一般是用不同的模型分别训练这两类实体，彼此之间不交叉。本文使用图 3 所示的网络结构同时对粗细粒度实体进行多任务学习，最后将两个 CRF 层的损失和作为全局损失一起训练，具体公式如下：

$$Loss=\alpha\cdot Loss_C+\beta\cdot Loss_G \tag{14}$$

其中，下标 C,G 分别表示粗粒度和细粒度实体。

5 实验设计

5.1 数据集介绍

本文构建的数据集的原始语料来自某三甲医院的大肠癌电子病历，本文在电子病历的基础上进行了标注。本文数据集共有 12743 个粗粒度实体和 32478 个细粒度实体。在实验中，本文按照训练集和测试集 7∶3 的原则分配了数据，数据分配的具体分布如表 2 和表 3 所列。

表 2 粗粒度数据分布

Table 2 Distribution of statistics of coarse-grained entity

粗粒度实体	训练集	测试集
自诉症状	2490	1220
体格检查	2538	1196
实验室检查	1301	644
疾病	629	295
疾病史	112	38
手术名	39	15
手术史	331	125
用药方案	1228	542
总计	8668	4075

表 3 细粒度数据分布

Table 3 Distribution of statistics of fine-grained entity

细粒度实体	训练集	测试集
检查指标	3158	1367
检查结果	5392	2451
药品名	1228	542
药品用量	5906	187
药品用法	499	127
存在情况	1426	656
情景限定	191	77
程度等级	587	269
身体部位	2710	1256
原子症状	600	265
原子疾病	1418	675
时间点	643	309
时间段	379	160
总计	24137	8341

5.2 实验设置

在实验前根据数据集中的句长统计，将最大句长 n 设置

为 300，并对句长大于 n 的句子进行分析，发现其大多是检查报告形式的文本。对于这些文本，采用分号进行切分并添加到数据集中。实验采用 word2vec[17-18] 的 skip-gram 算法在数据集上进行预训练，最终得到了 128 维的字向量，并将其作为嵌入层的初始化向量。在训练阶段，在双向 LSTM 后采用 dropout 以防止过拟合，后文将对多头注意力中的参数 h 的设置进行具体介绍。实验采用 Adam 小批量梯度下降算法，损失函数中的超参数设置为 $\alpha=\beta=1$，其余参数如表 4 所列。实验采用实体级别的准确率、召回率以及准确率和召回率的调和平均数 $F1$ 值作为评价指标。

表 4 超参数设置

Table 4 Setting up hyperparameters

参数名	参数值
词向量维度	128
LSTM 隐层节点数	128
Self-Attention 输出维度	256
dropout	0.2
学习率	0.001
batch_size	256

5.3 实验结果

5.3.1 对比实验

为了验证本文提出的模型的有效性，我们对比了现有的一些主流 NER 模型，包括由 Huang 等[8] 提出的经典序列标注模型(Bi-LSTM)、Ma 等[21] 提出的将卷积网络用于提取字符特征的模型、Zheng 等[22] 提出的结合注意力机制的模型、Wang 等[14] 提出的将字典信息融入的模型和 Qiu 等[13] 提出的采用残差膨胀卷积的模型。

由表 5 可知，除了在粗粒度实体识别中的 Precision 和在细粒度实体识别中的 Recall 比 Qiu 等[13] 提出的模型的值略低以外，本文提出的模型在这两类实体中的其他指标均高于其他模型的其他指标，并且在综合性能 $F1$ 值上均取得了最好的效果，分别达到了 92.88 和 85.48。

表 5 与现有模型的性能对比

Table 5 Experimental comparison results with current models

模型	粗粒度实体			细粒度实体		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*
文献[8]中的模型	92.48	89.87	91.12	84.54	84.31	84.43
文献[21]中的模型	93.23	89.22	91.18	84.80	84.09	84.44
文献[22]中的模型	93.36	91.17	92.25	84.82	82.59	83.69
文献[14]中的模型	93.59	90.43	91.98	85.20	83.51	84.35
文献[13]中的模型	**94.36**	88.99	91.60	79.73	**85.87**	82.69
本文模型	93.22	**92.56**	**92.88**	**85.62**	85.34	**85.48**

5.3.2 多头注意力参数 h 对模型的影响

多头注意力中的 h 表示相互平行的自注意力层的个数，各个自注意力层相当于不同的子空间。假设每个自注意力层中的各个向量维度相同，那么随着 h 的增加，模型能够捕获到更多的来自不同字空间的注意力信息。为了研究 h 对模型的影响，我们固定 d_{model} 的维度为 256，随着 h 的变化，虽然自注意力层中的向量 $d_k=d_v=d_{\text{model}}/h$ 也会随之而变化，但总的输出向量的维度保持不变。表 6 列出了在不同 h 的情况下该模型的性能。

表 6 不同 h 下模型的性能对比

Table 6 Experimental comparison results under different h

h	粗粒度实体			细粒度实体		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*
1	92.16	91.51	91.83	85.48	83.23	84.34
2	92.43	91.88	92.15	85.66	83.54	84.58
4	92.86	91.23	92.03	85.62	85.34	**85.48**
8	92.52	91.68	92.10	**85.85**	84.46	85.15
16	**93.22**	**92.56**	**92.88**	84.71	**85.88**	85.29

表 6 所列的实验结果表明，对于粗粒度实体来说，当 $h=16$ 时，模型在所有指标上得到了最好的结果；对于细粒度实体来说，当 $h=4$ 时，模型在综合性能 $F1$ 值上取得了最好的结果。这也从侧面反映出，由于粗粒度实体所蕴含的语义更多，需要通过更多的低维子空间来捕获信息。而细粒度实体因为本身蕴含的语义没有那么丰富，所以更高的 h 并没有带来更好的效果。综上，我们使用 $h=16$ 的粗粒度模型和 $h=4$ 的细粒度模型的集成作为最终的模型。

5.3.3 消融实验

为了证明本文提出的模型中的自注意力和残差连接的有效性，表 7 列出了模型分别在去掉自注意力时或去掉残差连接时的性能。

表 7 消融实验的结果

Table 7 Results of ablation experiments

模型	粗粒度实体			细粒度实体		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*
-Self-Attention	92.55	91.03	91.78	84.70	85.23	85.02
- Skip Connect	92.82	90.43	91.61	84.72	83.40	84.05
All	**93.22**	**92.56**	**92.88**	**85.62**	**85.34**	**85.48**

由表 7 可知，在分别去掉自注意力层、残差连接后，粗细粒度的各项指标均有所降低，综合指标 $F1$ 值分别下降了 1.1%，1.27%和 0.46%，1.4%，由此证明了自注意力层和残差连接的有效性。

结束语 本文针对中文电子病历中的实体存在的粗细粒度现象，构建了一个包含粗细粒度医疗实体的数据集，并提出了一个采用自注意力的多任务学习模型来进行多粒度的临床实体识别。该模型在不使用外部资源的情况下，通过共享编码层来提取两种粒度实体的公共特征，并通过自注意力机制来学习两种粒度实体对公共特征的相关度，从而得到对两种粒度的实体更好的识别效果。在我们自己标注的数据集上的实验结果也表明了该方法的有效性，后续工作将研究如何利用外部知识来提升模型的效果。

参 考 文 献

[1] HE B, DONG B, GUANY, et al. Building a comprehensive syntactic and semantic corpus of Chinese clinical texts[J]. Journal of Biomedical Informatics, 2017, 69: 203-217.

[2] FUKUDA K, TSUNODA T, TAMURA A, et al. Toward information extraction: identifying protein names from biological papers[C]// Pac Sympbiocomput. 1998: 707-718.

[3] FRIEDMAN C, ALDERSON P O, AUSTIN J H M, et al. A general natural-language text processor for clinical radiology

[J]. Journal of the American Medical Informatics Association, 1994, 1(2): 161-174.

[4] SONG M, YU H, HANW S. Developing a hybrid dictionary-based bio-entity recognition technique[J]. BMC Medical Informatics and Decision Making, 2015, 15(1): S9.

[5] ZHAO S. Named entity recognition in biomedical texts using an HMM model[C]// Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. Association for Computational Linguistics, 2004: 84-87.

[6] FINKEL J R, DINGARE S, NGUYEN H, et al. Exploiting context for biomedical entity recognition: from syntax to the web [C]// Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP). 2004: 91-94.

[7] SETTLES B. Biomedical named entity recognition using conditional random fields and rich feature sets[C]// Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP). 2004: 107-110.

[8] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv: 1508.01991, 2015.

[9] GRIDACH M. Character-level neural network for biomedical named entity recognition[J]. Journal of Biomedical Informatics, 2017, 70: 85-91.

[10] DANG T H, LE H Q, NGUYEN T M, et al. D3NER: biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information [J]. Bioinformatics, 2018, 34(20): 3539-3548.

[11] LIU J, CHEN S, HE Z, et al. Learning BLSTM-CRF with Multi-channel Attribute Embedding for Medical Information Extraction[C]// CCF International Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2018: 196-208.

[12] GIORGI J M, BADER G D. Transfer learning for biomedical named entity recognition with neural networks[J]. Bioinformatics, 2018, 34(23): 4087-4094.

[13] QIU J, WANG Q, ZHOU Y, et al. Fast and Accurate Recognition of Chinese Clinical Named Entities with Residual Dilated Convolutions[C] // 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2018: 935-942.

[14] WANG Q, ZHOU Y, RUAN T, et al. Incorporating dictionaries into deep neural networks for the chinese clinical named entity recognition[J]. Journal of Biomedical Informatics, 2019, 92: 103-133.

[15] LUONG M T, LE Q V, SUTSKEVER I, et al. Multi-task sequence to sequence learning[J]. arXiv: 1511.06114, 2015.

[16] ZENG L, GAO D Q, RUAN T, et al. Analysis and marking of symptom composition based on CRF[J]. Journal of East China University of Science and Technology (Natural Science Edition), 2018(2): 277-282.

[17] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv: 1301.3781, 2013.

[18] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Advances in Neural Information Processing Systems. 2013: 3111-3119.

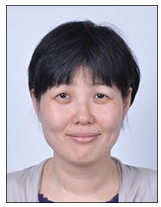
[19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[20] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[21] MA X, HOVY E. End-to-end sequence labeling via bi-directional lstm-cnns-crf[J]. arXiv: 1603.01354, 2016.

[22] ZHENG G, MUKHERJEE S, DONG X L, et al. OpenTag: Open attribute value extraction from product profiles [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018: 1049-1058.

ZHOU Xiao-jin, born in 1996, postgraduate, is a student member of China Computer Federation. His main research interests include natural language processingand information extraction.



RUAN Tong, born in 1973, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include text extraction, knowledge graph and data quality assessment.