文章编号：2095-2163(2021)03-0094-04 中图分类号：TP391.4 文献标志码：A

融合知识的中文医疗实体识别模型

刘龙航，赵铁军
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：从医疗文本中抽取知识对构建医疗辅助诊断系统等应用具有重要意义。实体识别是其中的核心步骤。现有的实体识别模型大都是基于标注数据的深度学习模型，非常依赖高质量大规模的标注数据。为了充分利用已有的医疗领域词典和预训练语言模型，本文提出了融合知识的中文医疗实体识别模型。一方面基于领域词典提取领域知识，另一方面，引入预训练语言模型 BERT 作为通用知识，然后将领域知识和通用知识融入到模型中。此外，本文引入了卷积神经网络来提高模型的上下文建模能力。本文在多个数据集上进行实验，实验结果表明，将知识融合到模型中能够有效提高中文医疗实体识别的效果。

关键词：实体识别；序列标注模型；融合知识

Chinese medical entity recognition model with knowledge fusion

LIU Longhang, ZHAO Tiejun
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Extracting knowledge from medical texts is of great significance to the construction of medical auxiliary diagnosis system and other applications. Entity recognition is an important step. Most of the existing entity recognition models are based on the deep learning model of annotation data, which rely heavily on high-quality large-scale annotation data. In order to make full use of the existing medical dictionary and pre-training language model, this paper proposes a Chinese medical entity recognition model with knowledge fusion. On one hand, domain knowledge is extracted based on domain dictionary; on the other hand, the pre-training language model BERT is used as general knowledge, and then domain knowledge and general knowledge are integrated into the model. In addition, convolution neural network is introduced to improve the context modeling ability of the model. In this paper, experiments are carried out on multiple datasets. The experimental results show that knowledge fusion can effectively improve the effect of medical entity recognition.

【Key words】 entity recognition; sequence labeling model; knowledge fusion

0 引 言

在医疗健康领域中，拥有大量疾病及药品等数据。这些数据广泛存在于在线百科和医疗网站中，其中则蕴含着丰富的医学知识。从医疗文本中抽取知识对构建医疗辅助诊断系统等应用具有重要意义。中文医疗实体识别指的是给定一篇医学文本，标注出文本中出现的医学实体，是从医学文本中获取医学知识的关键技术。对于中文医疗实体识别任务而言，采用词级别的序列标注，会引入分词错误带来的误差。通常将中文医疗实体识别任务转为字符级别的序列标注问题。

本文探索了一种融合知识的深度学习模型架构。一方面基于领域词典提取领域知识，另一方面，引入预训练语言模型 BERT 作为通用知识，然后将领域知识和通用知识融入到模型中。此外，引入了 CNN 来提高模型的上下文建模能力。实验方面，本文在多个数据集上进行实验，实验结果表明，将知识融合到模型中能够有效提高中文医疗实体识别的效果。

1 相关工作

早期的研究人员通常采用医学专家定义的规则并且基于医学领域词典对医疗实体进行自动识别[1-2]。基于医学词典及规则方法的优点是无需标注数据，缺点是维护高质量的医学词典困难，并且专家定义的规则只适合某些场景。后来机器学习模型逐渐成为了实体识别的主流方法[3-4]。基于传统机器学习方法无需人工定义规则和医学词典，具有不错的稳定性。然而，该方法的效果很大程度上取决于定义的特征模板是否考虑周全，限制了模型的泛化能力。

近年来，深度学习方法在实体识别领域取得了显著的效果[5]。Li 等人[6]将 BiLSTM-CRF 模型应

作者简介：刘龙航(1996-)，男，硕士研究生，主要研究方向：自然语言处理、信息抽取；赵铁军(1962-)，男，博士，教授，博士生导师，主要研究方向：自然语言处理、机器翻译、机器学习与人工智能。

收稿日期：2020-10-28

用于中文电子病历的实体识别任务，并基于医疗领域数据训练了更丰富、更专业的词向量，进一步提高了模型性能。Lee 等人[7]将预训练语言模型 BERT[8]应用于医疗领域，基于大规模医学领域的英文语料训练得到 BioBert 模型，最终在多个英文实体识别语料上取得最优结果。基于深度学习的方法效果优于传统机器学习方法的重要原因是该方法无需人工定义特征模板，而是通过深度神经网络自动进行特征学习，从而具有更强的泛化能力。

2　知识提取

2.1　基于领域词典的领域知识提取

字级别的序列标注问题本质上是对每个字进行多分类。因此，可以利用医疗领域词典这一额外资源增强每个字的特征表示，从而提高分类的准确度。基于此，最朴素的思想就是基于医疗领域词典给每个字打标签，再对离散化的标签进行特征表示。具体来说，给定一个由 T 个汉字构成的句子 $S = < x_1, \cdots, x_T >$ 和一个额外的医疗领域词典 D，首先基于双向最大匹配算法[9]对句子 S 进行切分，将属于 D 的文本片段切分出来，并打上对应的实体类型标签，不属于 D 的汉字标记为“None”。

通过双向最大匹配算法得到打上标签的文本片段后，可以进一步对文本片段中的字打标签。考虑了每个字在其所属实体的位置信息：如果该字单独构成一个实体，那么在字标签由前缀“S”和其所属文本片段的实体类型标签构成；类似地，用标志“B”和其所属文本片段的实体类型标签指示某个实体的第一个字；用标志“E”和其所属文本片段的实体类型标签指示某个实体的最后一个字；用标志“I”和其所属文本片段的实体类型标签指示某个实体中间的字。表 1 中举例说明了这种标记方式。通过 embedding 方式对字标签进行表示得到相应的特征表示向量。

表 1　字标签例子

Tab. 1　An example of word tags

标记类别	标记内容									
句子	二	甲	双	胍	在	小	肠	吸	收	。
输出标签	B-m	I-m	I-m	E-m	O	B-b	E-b	O	O	O
基于词典的字标签	B-m	I-m	I-m	E-m	None	B-b	E-b	None	None	None
所属实体类型	medicine				body part					

2.2　基于 BERT 的通用知识提取

从大规模无标注文本中进行语言表示学习是自然语言处理的重要研究方向。BERT(Bidirectional Encoder Representations from Transformers)是一个上下文表示的语言表示模型。这是基于使用双向多层 Transformer 编码器[10]的屏蔽语言模型(masked language model)预先训练的，结合下一个句子预测任务和更大的文本语料库，可以用于学习更好的双向上下文表示。

BERT 模型有 2 个步骤，分别是：预训练和微调(finetuning)。通过预训练，BERT 从大规模无标注数据学习到的语言上下文表示向量，这些向量蕴含了自然语言的组织内在规律，本文把这种内在规律称为通用知识。序列标注任务是 token 级别的分类，对于中文而言，BERT 模型的 token 是字级别，这与本文采用字符级别的序列标注解决中文医疗实体识别问题正好吻合。基于 BERT 的通用知识提取则如图 1 所示。由图 1 可知，本文将 BERT 模型最后一层隐状态输出向量作为字的表示向量，将其视为通用知识融入到后续序列标注模型部分，丰富序列标注模型的输入信息，从而提高模型的识别能力。

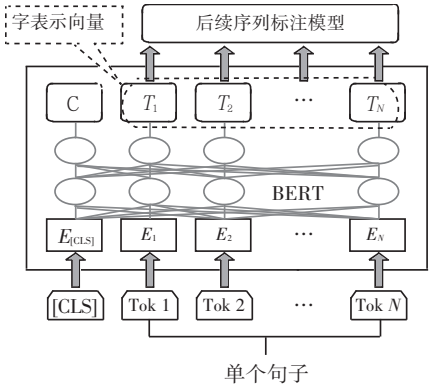

图 1　基于 BERT 的通用知识提取

Fig. 1　General knowledge extraction based on BERT

3　融合知识的实体识别模型

融合知识的实体识别模型的输入是单个句子，输出是字符级别的标注结果。模型分为 3 部分，分别是：输入编码层、上下文建模层以及条件随机场(CRF)输出层。其中，输入编码层将融合领域知识和通用知识，上下文建模层将通过 CNN 和 BiLSTM 对输入编码进行上下文建模，CRF 输出层用于解决

标签依赖问题,并输出最终的序列标注结果。对此拟展开研究分述如下。

3.1 输入编码层

给定一个由 T 个汉字构成的句子 $S = < x_1, \cdots, x_T >$,模型的输入编码层是要将离散型语言符号映射为低维稠密向量 $\boldsymbol{v}_i \in \mathbb{R}^{d_v}$,其中 d_v 表示最终每个字表示成的向量维度。为了考察融合领域知识和通用知识的效果,本文将输入编码层设计成如下 4 种方式:

(1)字嵌入:对于 S 中的第 t 个字 x_t,通过对嵌入矩阵 $\boldsymbol{W} \in \mathbb{R}^{V \times d_e}$ 的查找,可得到对应的字向量 $\boldsymbol{e}_t \in \mathbb{R}^{d_e}$,其中 d_e 表示字嵌入的维度,V 表示字表大小。将 $\boldsymbol{e}_t$ 视为最终的输入编码层向量表示 $\boldsymbol{v}_t$。

(2)字嵌入结合领域知识:对于 S 中的第 t 个字 x_t,通过基于领域词典提取领域知识方法可以将其表示为一个特征向量 $\boldsymbol{k}_t \in \mathbb{R}^{d_k}$,其中 d_f 表示词典特征向量的维度。然后将字嵌入和领域知识进行结合,具体采用向量拼接的方式,即将 $\boldsymbol{k}_i$ 与 $\boldsymbol{e}_i$ 进行拼接得到最终的输入编码层向量表示 $\boldsymbol{v}_t \in \mathbb{R}^{d_e+d_k}$,计算公式如下:

$$\boldsymbol{v}_t = \boldsymbol{e}_t \oplus \boldsymbol{k}_t, \tag{1}$$

(3)BERT:对于 S 中的第 t 个字 x_t,基于 BERT 的通用知识提取,可以得到 x_t 对应的上下文特征表示向量 $\boldsymbol{b}_t \in \mathbb{R}^{d_b}$,其中 d_b 表示 BERT 模型隐藏层的维度。将 $\boldsymbol{b}_t$ 视为最终的输入编码层向量表示 $\boldsymbol{v}_t$。

(4)BERT 结合领域知识:考虑将领域知识和通用知识同时融入模型,同样采用向量拼接的方式,即将 BERT 的字向量表示 $\boldsymbol{b}_t$ 与词典特征向量 $\boldsymbol{k}_t$ 进行拼接得到最终的输入编码层向量表示 $\boldsymbol{v}_t \in \mathbb{R}^{d_b+d_k}$,即:

$$\boldsymbol{v}_t = \boldsymbol{b}_t \oplus \boldsymbol{k}_t. \tag{2}$$

3.2 上下文建模层

输入编码层的输出为 $V = < v_1, \cdots, v_T >$,在输入编码的基础上,模型将进一步进行上下文建模。本文采用如下 2 种上下文建模方式:

(1)BiLSTM[11]:BiLSTM 用于学习每一个字的上下文信息,对于每个时间步 t,BiLSTM 输出相应的隐状态向量 $\boldsymbol{h}_t$,即:

$$\boldsymbol{h}_t = BiLSTM(V, t), \forall t \in [1, \cdots, T], \tag{3}$$

(2)CNN+BiLSTM:CNN[12] 用于学习每一个字的局部窗口上下文信息 $C = < c_1, \cdots, c_T >$,对于每个时间步 t,CNN 输出相应的向量 $\boldsymbol{c}_t$,即:

$$\boldsymbol{c}_t = CNN(v_{t-\frac{m-1}{2}}, \cdots, v_{t-\frac{m+1}{2}}), \forall t \in [1, \cdots, T], \tag{4}$$

其中,m 为 CNN 卷积核的窗口大小。然后,将 CNN 得到的局部上下文表示向量与字表示向量进行拼接,得到 $R = < r_1, \cdots, r_T >$,再输入到 BiLSTM 进行全局编码,因此有:

$$\boldsymbol{r}_t = \boldsymbol{c}_t \oplus \boldsymbol{v}_t, \tag{5}$$

$$\boldsymbol{h}_t = BiLSTM(R, t), \forall t \in [1, \cdots, T]. \tag{6}$$

3.3 条件随机场层

对于字符级别的序列标注任务,通常来说考虑相邻标签的依赖性有助于提高模型的识别能力。例如,开始标签“B”后面应该跟中间标签“I”或结束标签“E”,I 标签后面不能跟 B 标签或 S 标签。因此,研究中没有只使用的 $\boldsymbol{h}_t$ 来进行标签分类决策,而是使用条件随机场(CRF)来联合建模标签序列。CRF 层是一个将状态转移矩阵作为参数的线性链式无向图模型。通过该模型,可以利用前一个标签和后一个标签的信息来预测当前标签。

4 实验

4.1 数据集与评价指标

本实验采用 2 个数据集,分别是:CCKS 2019 评测—面向中文电子病历的医疗实体识别数据集[13] 和天池平台中文糖尿病标注数据集[14](A Labeled Chinese Dataset for Diabetes)。上述两个数据集都是按照文档级进行构建的,需要将文档级样本切分为句子级样本,切分后的数据集详细情况见表 2。

对应医疗实体识别任务,本文选择最常用的评价指标,即所有实体类型上的微平均(micro-average)F_1 值。

表 2 医学实体识别数据集

Tab. 2 Medical entity recognition dataset

数据集	CCKS2019 医疗实体识别数据集		中文糖尿病标注数据集	
	文档级	句子级	文档级	句子级
训练集	800	7 512	320	55 522
验证集	200	2 031	73	12 712
测试集	379	3 083	100	16 241
总计	1 379	12 626	493	84 475

4.2 实验设置

本文采用 Pytorch 框架[15] 进行模型实现。具体来说,对于输入编码层部分,字嵌入维度为 128,字标签的嵌入向量维度均为 128,BERT 采用 Google 官方基于中文维基百科训练的 $\text{BERT}_{\text{base}}$ 模型;上下文建模层部分,BiLSTM 的隐状态维度是 128,CNN 采取多种窗口大小的卷积核,分别是 3,5,7,每种卷积核特征数为 100。

4.3　实验结果与分析

通过对不同的输入编码以及不同的上下文建模进行组合，可以得到多个模型，将这些模型应用于实验数据集进行训练和预测。实验结果见表 3。由表 3 可以看出，采用 BERT 结合字标签的词典特征作为输入编码，使用 CNN+BiLSTM 作为上下文建模层时，在 2 个数据集的实验效果达到最好。下面将单独分析不同输入编码方式以及不同上下文建模方式的效果。

(1)领域知识：在其他条件相同的情况下，融入词典特征要比不融入词典特征的效果好。实验结果表明基于领域词典提取的领域知识能够有效提高模型的性能。这种领域知识取决于上下文和领域词典，不受其他句子或统计信息的影响。因此，在某种程度上可以提供与监督学习数据驱动不同的信息。

(2)通用知识：在其他条件相同的情况下，使用 BERT 的实验效果要明显优于不使用 BERT。这表明将 BERT 输出的字向量作为通用知识融入到模型能有效提高医疗实体识别的效果，研究认为这是因为 BERT 模型蕴含了自然语言构成的内在规律，这种规律是一种通用知识，能够提高模型的泛化能力。

(3)上下文建模层：在其他条件相同的情况下，CNN+BiLSTM 要优于 BiLSTM。这表明加入 CNN 能够提高模型上下文建模能力。研究认为这是因为 CNN 通过多窗口卷积能够捕获局部上下文信息，尤其是对于字级别的序列标注任务而言，这种局部上下文信息类似于字的组合信息，将这种信息和 BiLSTM 的全局上下文信息结合，提高了模型上下文建模能力。

表 3　医学实体识别实验结果

Tab. 3　Experimental results of medical entity recognition

上下文建模层	使用 BERT	融入词典特征	$F_1score/\%$	
			CCKS2019	中文糖尿病
BiLSTM	否	否	78.74	76.09
		是	81.85	77.46
	是	否	81.92	78.32
		是	83.06	79.23
CNN+BiLSTM	否	否	79.27	77.10
		是	82.10	77.86
	是	否	82.21	78.57
		是	**83.25**	**79.40**

5　结束语

针对中文医疗实体识别问题，本文提出了融合知识的实体识别模型，包括利用了词典提取领域知识和利用 BERT 预训练模型提取通用知识，并且在上下文建模方面引入了 CNN 来提取局部窗口上下文信息。实验结果表明，CNN 能够提高上下文的建模能力，基于词典的领域知识和基于 BERT 的通用知识都能提高模型效果。

参考文献

[1] 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.

[2] WU S T, LIU Hongfang, LI Dingcheng, et al. Unified medical language system term occurrences in clinical notes: A large-scale corpus analysis[J]. Journal of the American Medical Informatics Association, 2012, 19(e1): e149-e156.

[3] 叶枫, 陈莺莺, 周根贵, 等. 电子病历中命名实体的智能识别[J]. 中国生物医学工程学报, 2011, 30(2):256-262.

[4] 王世昆, 李绍滋, 陈彤生. 基于条件随机场的中医命名实体识别[J]. 厦门大学学报(自然科学版), 2009,48(3): 359-364.

[5] JAGANNATHA A N, YU Hong. Bidirectional RNN for medical event detection in electronic health records[C]//Proceedings of The 2016 Conference of The North American Chapter of The Association For Computational Linguistics: Human Language Technologies. San Diego, California:ACL, 2016, 2016 : 473-482.

[6] LI Z, ZHANG Q, LIU Y, et al. Recurrent neural networks with specialized word embedding for chinese clinical named entity recognition[C]//CEUR Workshop Proceedings 2017. [S.l.]: dblp, 2017,1976:55-60.

[7] LEE J, YOON W, KIM S, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining[J]. arXiv preprint arXiv:1901.08746v2, 2019.

[8] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceeding of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies(Long and Short Papers). Minneapolis, Minnesota: ACL,2019:4171-4186.

[9] GAI Rongli, GAO Fei, DUAN Liming, et al. Bidirectional maximal matching word segmentation algorithm with rules[J]. Advanced Materials Research, 2014,926-930: 3368-3372.

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

[11] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8):1735-1780.

[12] KIM Y . Convolutional neural networks for sentence classification [J]. arXiv preprint arXiv:1408.5882, 2014.

[13] 医渡云. CCKS 2019 评测任务一面向中文电子病历的命名实体识别数据集[DB/OL] .[2019-08-05]. http://openkg.cn/dataset/yidu-s4k.

[14] 阿里云. 中文糖尿病标注数据集[DB/OL] .[2019]. https://tianchi.aliyun.com/dataset/dataDetail?dataId=22288.

[15] PASZKE A, GROSS S, MASSA F, et al. PyTorch: An imperative style, high-performance deep learning library[M]//WALLACH H, LAROCHELLE H, BEYGELZIMER A, et al. Advances in Neural Information Processing Systems. Harju Maakond Tallin Estonia: Curran Associates, Inc., 2019:8024-8035.