DOI:10.16781/j.0258-879x.2019.05.0497

·专题报道·

基于句子级 Lattice-长短记忆神经网络的中文电子病历命名实体识别

潘璀然[1]，王青华[1]，汤步洲[2]，姜　磊[3]，黄　勋[4]，王　理[1*]
1. 南通大学医学院医学信息学教研室，南通 226001
2. 哈尔滨工业大学（深圳）计算机科学与技术学院，深圳 518055
3. 海军军医大学（第二军医大学）长征医院风湿免疫科，上海 200433
4. 南通大学信息科学技术学院通讯工程教研室，南通 226001

[摘要] **目的**　提出一种基于 Re-entity 新分词方法的条件随机场（CRF）模型，并与双向长短记忆神经网络（BiLSTM）-CRF 和 Lattice-长短记忆神经网络（LSTM）进行比较。**方法**　比较了现有实体识别方法和模型后，针对 2018 年全国知识图谱与语义计算大会（CCKS2018）任务一“电子病历命名实体识别”，提出基于 Re-entity 的 CRF、BiLSTM-CRF、Lattice-LSTM 方法，并在不同语料库训练不同参数级别的字符向量集。分别将各方法引入神经网络模型中进行模型性能对比实验，最后分别基于句子级和篇级输入句长进行对比研究。**结果**　CRF 模型在最优特征工程的结果下引入 Re-entity 方法后性能得到提高，句子级的 Lattice-LSTM 模型在该任务上取得了 89.75% 的严格 F1-measure，优于 CCKS2018 任务一的最高结果（89.25%）。**结论**　基于 Re-entity 新分词方法的 CRF 模型可利用中文临床药物知识库有效提高电子病历中药物的识别率，Re-entity 方法可改善数据预处理阶段分词导致的错误累加，Lattice 结构可以更好地结合字符和词序列的潜在语义信息，同时句子级输入能有效提高神经网络模型的识别准确率。

[关键词] 计算机化病案系统；中文电子病历；实体识别；条件随机场；双向长短记忆神经网络；点阵长短记忆神经网络
[中图分类号] R-37　**[文献标志码]** A　**[文章编号]** 0258-879X(2019)05-0497-10

Chinese electronic medical record named entity recognition based on sentence-level Lattice-long short-term memory neural network

PAN Cui-ran[1], WANG Qing-hua[1], TANG Bu-zhou[2], JIANG Lei[3], HUANG Xun[4], WANG Li[1*]
1. Department of Medical Informatics, School of Medicine, Nantong University, Nantong 226001, Jiangsu, China
2. College of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Shenzhen 518055, Guangdong, China
3. Department of Rheumatology and Immunology, Changzheng Hospital, Naval Medical University (Second Military Medical University), Shanghai 200433, China
4. Department of Communication Engineering, School of Information Science and Technology, Nantong University, Nantong 226001, Jiangsu, China

[Abstract] **Objective**　To propose a conditional random field (CRF) model based on the new word segmentation method Re-entity, and to compare with bi-directional long short-term memory neural network (BiLSTM)-CRF and Lattice-long short-term memory neural network (LSTM). **Methods**　After analyzing the existing entity recognition methods, we proposed CRF method based on Re-entity, BiLSTM-CRF and Lattice-LSTM for the China Conference on Knowledge Graph and Semantic Computing in 2018 (CCKS2018) task one: Chinese clinical named entity recognition, and trained character vector sets at different parameter levels based on different corpora. The comparative experiments on model performance were carried out in the different neural network models for each methods. Finally, the comparative study was carried out based on different input lengths such as the sentence level and the text level. **Results**　Re-entity method can improve the performance of CRF model. Lattice-LSTM model based on sentence level achieved a strict F1-measure of 89.75% on this task, which was higher than the

[收稿日期] 2019-02-23　[接受日期] 2019-04-12
[基金项目] 国家重点研发计划(2018YFC0116902), 国家自然科学基金(81873915), 江苏省研究生科研与实践创新计划项目(KYCX17-1932). Supported by National Key Research and Development Plan (2018YFC0116902), National Natural Science Foundation of China (81873915), and Postgraduate Scientific Research and Practice Innovation Program of Jiangsu Province (KYCX17-1932).
[作者简介] 潘璀然，硕士生．E-mail: 18606422000@163.com
*通信作者(Corresponding author). Tel: 0513-85051891, E-mail: wangli@ntu.edu.cn

highest F1-measure (89.25%) on the task one of CCKS2018. **Conclusion** The CRF model based on Re-entity can effectively improve the recognition rate of traditional Chinese medicines in electronic medical records by using normalized Chinese clinical drug. Re-entity method can improve the error accumulation caused by word segmentation in data preprocessing. Lattice structure can better combine the latent semantic information of characters and word sequences. At the same time, sentence-level input can effectively improve the recognition accuracy of neural network models.

[Key words] computed medical records systems; electronic medical record; entity identification; conditional random field; bi-directional long short-term memory neural network; lattice-long short-term memory neural network

[Acad J Sec Mil Med Univ, 2019, 40(5): 497-506]

电子病历（electronic medical record，EMR）记录了患者的整个医疗过程，它包含有患者大量的诊疗信息[1]，是生物医学临床研究的重要数据来源。EMR 命名实体识别（named entity recognition，NER）作为生物医学文本挖掘的第一步，对医学知识库的构建、药物安全性检测和临床决策支持系统等的研究有着重要意义。挖掘出的临床数据不仅可以用于临床科研，还有助于为患者制定个性化的精准医疗服务[2]。

通常，NER 的方法学研究主要分为 3 大类：（1）基于词典和规则的方法。这种方法在数据量少时效果较好且识别速度快，但是该方法对词典规模及词典覆盖率的依赖性较大且编写规则需要耗费大量的人力和物力。现在大多情况下将规则和机器学习方法结合使用。（2）传统机器学习的方法，其主流方法有条件随机场（conditional random field，CRF）、支持向量机（support vector machine，SVM）等。程健一等[3]基于 2014 i2b2/UTHealth 中的任务，提出基于 SVM 和 CRF 的双层分类模型对 EMR 去隐私化。（3）深度学习方法。深度神经网络是一种挖掘潜在有用特征的多层神经网络，每一层输出都是该语句的一种抽象表示，语言本身就是一种抽象的表达，因此，在大量训练数据的基础上生成基于向量的特征表示，利用神经网络进行 NER 是目前学者正在探索的一种方法[4]。由于中英文语言特征的差异，中文实体识别首先要对文本进行分词，分词错误则会导致在 NER 上的错误累加。因此，已有研究证明基于字符的方法在中文 NER 上优于基于词的模型[5]。不会导致分词错误累加是字符级模型的优点，但从另一方面来说也是这种方法的缺点，因为有些单词信息蕴含的语义信息可以使字符级模型在识别实体时产生歧义，如将“上腹部疼痛”识别成“上”和“腹部疼痛”。因此我们使用一种基于字和词混合网格的 Lattice-长短记忆神经网络（long short-term memory neural network，LSTM）结构[6]，这种结构能够实现对句中专有名词的识别，并将潜在的单词信息整合到基于字符的 LSTM-CRF 模型中。为了完成 2018 年全国知识图谱与语义计算大会（China Conference on Knowledge Graph and Semantic Computing，CCKS）[7]的测评任务，我们提出使用基于 Re-entity 的 CRF、双向长短记忆神经网络（bi-directional long short-term memory neural network，BiLSTM）-CRF 和 Lattice-LSTM 3 种方法，针对 EMR 中的解剖部位（body）、独立症状（osign）、症状表现（signs）、药物（drug）和手术治疗（treatment）5 类实体进行 NER。实验结果证明与词序列信息的混合模型 Lattice-LSTM 相比，字符粒度 CRF 使用更少的特征工程，并与基于字符的 BiLSTM-CRF 模型相比能更好地利用上下文中单词的潜在语义信息。

1　资料和方法

1.1　*实验语料*　本研究使用 CCKS2018 任务一“电子病历命名实体识别”发布的中文 EMR 语料，共 600 篇已标注的训练数据（train）和 400 篇未标注的测试数据。表 1 描述了 CCKS2018 测评任务发布的中文 EMR 数据规模，在 600 篇训练数据中共有 9 472 个身体解剖部位标注实体。

表 1　CCKS2018 数据规模

Tab 1　Data scale of CCKS2018

n

Data	Size/piece	Body	Osign	Signs	Drug	Treatment
Train	600	9 472	3 712	2 484	1 221	1 329
Test	400					

CCKS: China Conference on Knowledge Graph and Semantic Computing

1.2 CRF 模型 CRF 被广泛用于生物医学实体识别任务，它是一种判别式无向图模型，根据提供的特征对数据的标注进行学习，通过复杂的函数映射或决策叠加等机制，最后画出一个比较明显的边界区分实体和非实体。通俗来讲就是直接对 $P(Y|X)$ 建模，根据观测序列 $X(x_1, x_2, \cdots, x_n)$ 对标记序列 $Y(y_1, y_2, \cdots, y_n)$ 进行建模训练[8]。

广义的 CRF 是指满足 $P(Y_v|X, Y_w, w\neq v) = P(Y_v|X, Y_w, w\sim v)$ 的马尔科夫随机场。条件概率 $P(Y|X)$ 的定义为

$$P(Y|X)=\frac{1}{Z}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1}, y_i, x, i)+\sum_{i,l}\mu_l s_l(y_i, x, i)\right)$$

其中，

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

t_k 为 i 处的转移特征，对应权重 λ_k，每个 y_i 都有 k 个特征，转移特征针对的是前后标签之间的限定。s_l 为 i 处的状态特征，对应权重 μ_l，每个 y_i 都有 l 个特征，$Z(x)$ 是规范化因子。

为准确提取 EMR 中实体，不仅需要根据语料选择适合的学习方法，还要给出特征集，丰富的特征工程可以有效提高模型的学习能力和识别准确率。本研究基于对中文 EMR 文本的分析选取以下特征集：

（1）字符特征。中文字符本身可以作为一种基本特征，反映中文字符的基本信息[9]。为避免分词错误导致错误累加，我们使用字符级的分字结果作为语言符号特征引入。

（2）词性特征。通过分析 EMR 自由文本可以发现很多实体由多个名词嵌套组成，例如“剑突下压痛”由解剖位置“剑突下”和症状表现“压痛”组成。疾病名、症状等名词可能出现在动词等词性后面，因此将词性标注作为一类特征。

（3）词典特征。因为 EMR 的专业性比较强，为减少实体的错切分率，引入专业词典具有重要意义。在英文 NER 领域，MeSH、UMLS、SNOMED-CT、RxNORM 等通用词典发挥了重要作用。由于中文领域缺乏公开且完整的医学词典，因此，我们从网络和书本中整理并构建了 EMR 基本元素（item）、解剖位置（body）、症状（symptom）、中文临床药物标准知识库（normalized Chinese clinical drug，NCCD）[10] 4 个词典。词典规模如表 2 所示。例如 NCCD 包括 CONCEPT-NAME、CONCEPT-ID、CONCEPT-CLASS-ID 等属性，本研究整理了 CONCEPT-NAME 药名属性共 28 008 个作为词典特征加入。

表 2 词典规模

Tab 2 Dictionary scale

Type	Size
Item	1 985
Body	812
Symptom	1 137
NCCD[a]	28 008

[a]: CONCEPT-NAME; NCCD: Normalized Chinese clinical drug

（4）三元特征。在 CRF++ 工具的特征模板中采用 %x [row, col] 的格式，row 表示与当前位置相对应的行数，col 表示当前位置相对应的列数[11]。本研究使用的特征模板见表 3。

表 3 特征模板

Tab 3 Feature template

Type	Template
Unigram	U00: %x [−2,0]
Unigram	U01: %x [−1,0]
Unigram	U02: %x [0,0]
Unigram	U03: %x [1,0]
Unigram	…
Unigram	U13: %x [1,1]
Unigram	U14: %x [2,1]
Unigram	U15: %x [−2,1]/%x [−1,1]
Unigram	…
Unigram	U18: %x [1,1]/%x [2,1]
Unigram	U20: %x [−2,1]/%x [−1,1]/%x [0,1]
Unigram	U22: %x [0,1]/%x [1,1]/%x [2,1]
# Bigram	B

U00～U14 为一元特征，本研究选择当前字符的前后 2 个字符作为上下文特征，窗口为 5，U15～U18 是将相邻的 2 个一元特征分别进行组合合并为二元特征，U20～U22 为将当前字符的前后 3 个字符进行组合形成三元特征，当训练模型时会自动生成相邻特征的自由组合。

1.3 Re-entity 方法 由于现有分词方法比较适合通用数据，对 EMR 这种专业性强的文本的分词效果较差，因此为避免分词错误带来的错误累加提出了 Re-entity 方法。先训练出基于以上特征的 CRF 模型，然后将该模型用于 CCKS2017 的诊疗经过中共 2 205 条未标注的 EMR 文本，预测完成后抽取识别

的实体，最后将这些实体与已标注数据集中的实体混合整理成词典，利用该词典对训练数据再进行分词，保证已知命名实体在文本分词时不被切割，提高分词准确率，然后基于以上结果重新训练模型。

1.4　BiLSTM-CRF　Hochreiter 和 Schmidhuber[12] 于 1997 年提出的 LSTM 最初是为了解决循环神经网络（recurrent neural network，RNN）训练伴随的梯度缓慢和梯度爆炸，为保持信息完整引入了记忆细胞[13]，记录历史上下文信息。近年 LSTM 方法被广泛应用于自然语言处理领域，目前 NER 的主流模型是 BiLSTM-CRF，结构如图 1 所示。BiLSTM 是将每个序列呈现前向和后向 2 个单独的隐藏状态，分别捕获过去和未来的信息，然后连接 2 个隐藏状态合成最终输出。该模型在多项任务中取得了好成绩。Lample 等[14]提出了 2 种神经网络方法，一种是基于 BiLSTM 和 CRF，另一种是受移位归约解析器（shift-reduce parser）启发提出的基于转换（transition）的方法构建和标记分段。模型的 2 个词信息分别来源于从监督语料库中学习到的基于字符的词和从未标注数据中学习到的无监督词。该模型分别在 4 种语言下获得了目前 NER 的最好性能。我们使用 BiLSTM 模型自动从训练数据中学习特征，并在最后的输出层中使用 CRF 替换 softmax 函数进行分类决策，CRF 中的转移特征会考虑输出标注之间的关联性和合理性。

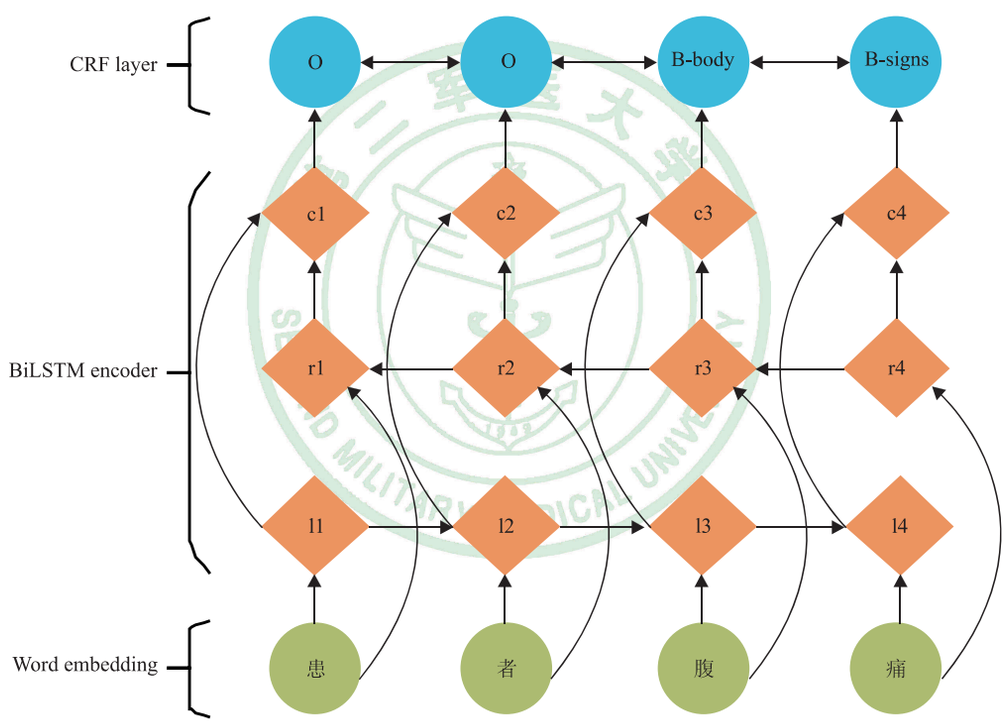


图 1　BiLSTM-CRF 结构图

Fig 1　BiLSTM-CRF structure

BiLSTM: Bi-directional long short-term memory neural network; CRF: Conditional random field; O: Placeholder; B: Beginning of entity; l: Forward hidden layer; r: Backward hidden layer; c: Final out of hidden layer; 1, 2, 3, 4: Character sequence. The arrows indicate model running direction

1.5　Lattice-LSTM　基于字符的模型可以避免词模型在分词阶段的错误累加，但有时也会导致信息丢失，因为有些上下文中的词序列蕴含的语义信息可辅助模型性能的提高。而 Lattice 结构可以利用字符和词序列信息、门控结构选择最相关的字符和单词以获得更好的 NER 结果。该模型整体思路是利用 Lattice-LSTM 表示输入句子中的词汇词（lexicon word），然后将潜在的词信息融合到基于字符粒度的 LSTM-CRF 模型中。

如图 2 所示，通过输入句子与词典 D 进行匹配构造单词-字符的 Lattice 结构。词典 D 是由大规模经过分词后的中文文本 Gigaword（*https://catalog.ldc.upenn.edu/LDC2011T13*）使用 Word2vec 训练后得到的。例如“疼痛”“腹部”“腹部疼

痛”等存在于词典 D 中的单词可用于消除上下文中潜在的命名实体歧义。按照顺序一个一个字符组合还能组成“无腹部”一词，而该词不在 Lattice 词格中，因此在进行匹配时就可以避免这种歧义实体的发生。

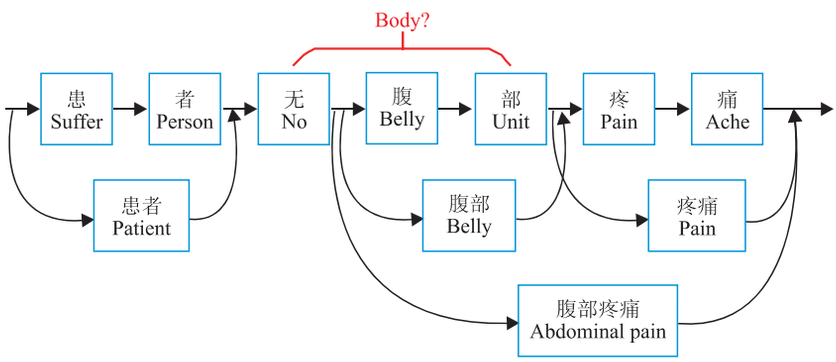


图 2　单词－字符 Lattice

Fig 2　Word-character Lattice

The arrows indicate character combination

1.5.1　Lattice-LSTM 整体结构　如图 3 所示，使用基于字符的 LSTM-CRF 作为主干模型，模型中的“红色细胞（ ）”是句中的潜在词序列，与主干 LSTM 模型中相应的字符连接。例如“痛”这个字，它的潜在词汇有“腹痛”和“腹部疼痛”，因此当计算“痛”的向量时除了考虑“痛”字以外还应考虑“腹痛”和“腹部疼痛”。

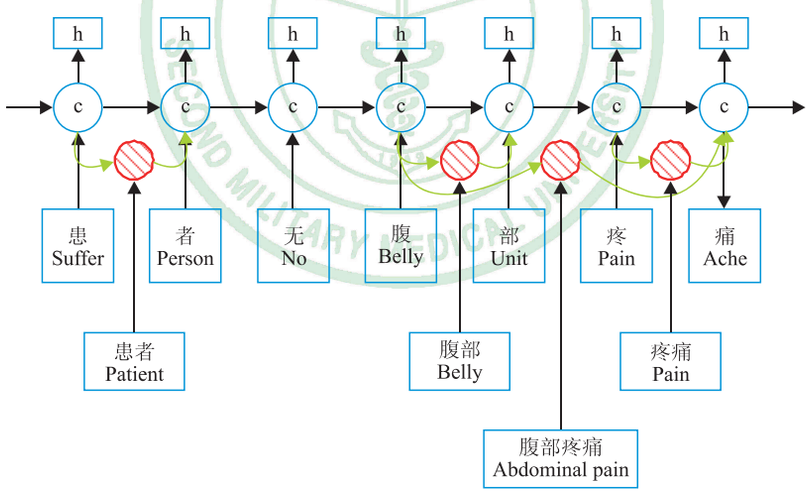


图 3　Lattice-LSTM 结构图

Fig 3　Lattice-LSTM structure

LSTM: Long short-term memory neural network; c: Character vector; h: Hidden layer vector; : Potential word information. The black arrows indicate model running direction, and the green arrows indicate dynamical route information from different paths to each character

1.5.2　Lattice 模型　使用英文领域效果最好的 NER 模型 LSTM-CRF 作为主模型[15]，在此基础上集成了词序列信息和用于控制信息流的附加门。输入的句子 S 可以表示为 $S=c_1, c_2, c_3, \cdots, c_m$（$c_j$ 表示句中第 j 个字符，共有 m 个字符），也可以表示为 $S=w_1, w_2, w_3, \cdots, w_n$（$w_i$ 表示句中第 i 个单词结果，共有 n 个单词）。$t(i,k)$ 表示句中第 i 个单词的第 k 个字符的索引 j。

模型的输入是字符序列及其与词典 D 中单词匹配的所有字符子序列。如图 4 所示，c_j^c 的计算考虑到了词格中的单词序列 $w_{b,e}^d$。

$$x_{b,e}^w = e^w(w_{b,e}^d)$$

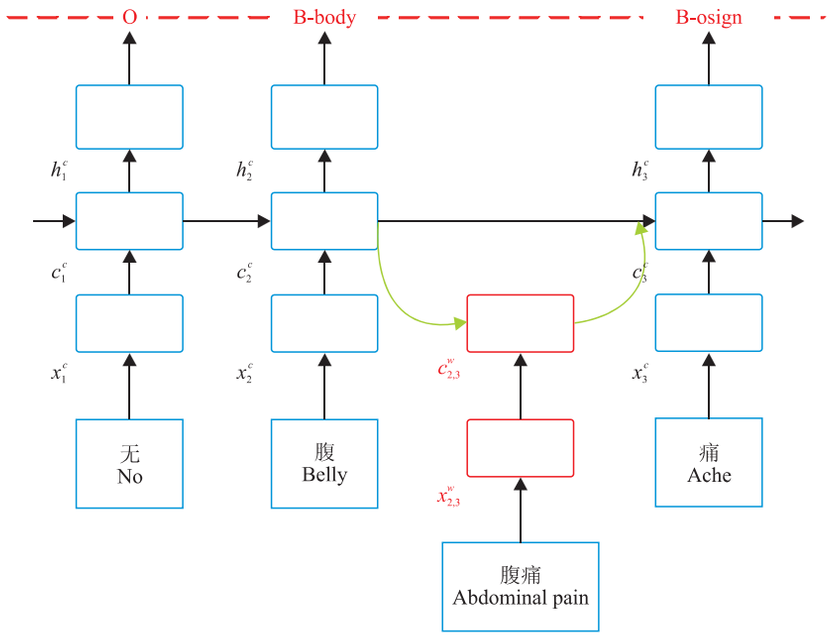


图 4　Lattice 模型

Fig 4　Lattice model

O: Placeholder; B: Beginning of entity. x_j^c: Character input vector; c_j^c: Character cell vector; h_j^c: Hidden vector; $x_{2,3}^w$: Word embedding of abdominal pain; $c_{2,3}^w$: Recurrent state of $x_{2,3}^w$ from the beginning of the sequence. The arrows indicate information sources

其中 e^w 是词向量查找表，$c_{b,e}^w$ 表示每一个 $x_{b,e}^w$ 的递归状态，w 表示词，下标 b,e 分别表示词 w 在序列中的开始和结束位置。

$$\begin{bmatrix} i_{b,e}^w \\ f_{b,e}^w \\ \tilde{c}_{b,e}^w \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(W^{wT} \begin{bmatrix} x_{b,e}^w \\ h_b^c \end{bmatrix} + b^w \right)$$

$$c_{b,e}^w = f_{b,e}^w \odot c_b^c + i_{b,e}^w \odot \tilde{c}_{b,e}^w$$

其中 $i_{b,e}^w$ 和 $f_{b,e}^w$ 是 LSTM 的输入和遗忘门集合。W^{wT}和 b^w 是模型参数，σ 表示 sigmoid 函数，$tanh$ 是输出的激活函数。求出词序列的 $c_{b,e}^w$ 之后，每个隐含层的 c_j^c 的计算即受到多路径信息流的影响，例如“痛”这个字符的 c_3^c 的计算即会受 x_3^c 和 $c_{2,3}^w$ 及上一个隐含层输出的影响，使用附加门控制 $c_{b,e}^w$ 到 $c_{b,e}^c$ 的信息流，即根据当前字符和词汇信息计算输入到字符的词汇信息的权重。

$$i_{b,e}^c = \sigma \left(W^{lT} \begin{bmatrix} x_e^c \\ c_{b,e}^w \end{bmatrix} + b^l \right)$$

然后利用如下公式计算当前位置索引为 j 的字符向量融合了潜在单词后的更新状态。

$$c_j^c = \sum_{b \in \{b' \mid w_{b',j}^d \in D\}} \alpha_{b,j}^c c_{b,j}^w + \alpha_j^c \tilde{c}_j^c$$

最后在模型最上层加一层 CRF 保证最终标注的合理性。

1.6　评估标准　评估方法使用精确率、召回率及 F1-measure 作为评测指标，并根据系统预测结果与人工标注的金标准结果的比较分为严格指标和松弛指标，本研究仅讨论严格指标。参赛系统的预测标注集合记为 $S=\{s_1, s_2, \cdots, s_m\}$，其中 m 是预测序列集 S 的字符（s）数，人工标注的金标准结果集合记为 $G=\{g_1, g_2, \cdots, g_n\}$，$n$ 表示金标准集合 G 的字符（g）数。集合元素为一个个实体，表示为四元组 $\langle d, pos_b, pos_e, c \rangle$，其中 d 表示文档，pos_b 和 pos_e 分别对应实体提及在文档 d 中的起止下标，c 表示实体提及所属预定义类别。

严格指标：我们定义 $s_i \in S$ 与 $g_j \in G$ 严格等价，当且仅当

（1）$s_i.d = g_j.d$

（2）$s_i.pos_b = g_j.pos_b$

（3）$s_i.pos_e = g_j.pos_e$

（4）$s_i.c = g_j.c$

基于以上等价关系，我们定义集合 S 与 G 的严格交集为 $\cap_s$。由此得到严格指标

$$P_s = \frac{|S \cap_s G|}{|S|}$$

$$R_s = \frac{|S \cap_s G|}{|G|}$$

$$F_{1s}=\frac{2PR}{P+R}$$

2 结 果

2.1 CRF 特征扩展实验结果与分析 CRF 模型实验首先以模型 ①（CRF-onlychar）作为 baseline，baseline 模型只使用了原始文本做字符特征在 CRF＋＋ 0.58 工具中进行 NER。模型 ②（CRF-POS-Dic）在 baseline 模型的基础上加入词性标注（part-of-speech，POS）和词典 Item、Body、Symptom 作为特征扩展。模型 ③（CRF-POS-Dic-NCCD）则在模型 ② 的基础上加入词性标注和 3 个词典的基础上引入 NCCD。模型 ④（CRF-POS-Dic-NCCD-Re-entity）是在模型 ③ 的基础上加入 Re-entity 分词方法降低分词错误率。以上 4 个特征扩展模型的实验结果如图 5 所示。

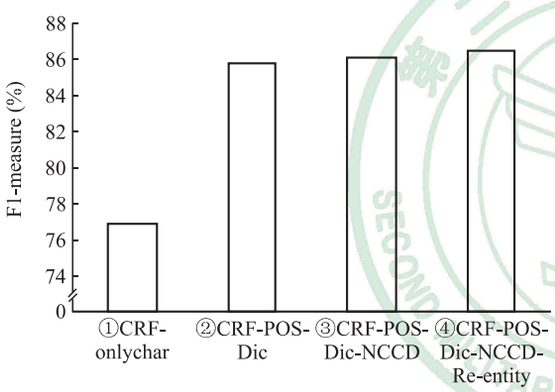


图 5 CRF 特征扩展实验结果

Fig 5 Experimental results of CRF feature extension

CRF: Conditional random field; POS: Part-of-speech; Dic: Item＋body＋symptom (dictionary); NCCD: Normalized Chinese clinical drug

模型 ① 中仅使用字符特征进行 CRF 模型训练，得到 76.96% 的 F1-measure。模型 ② 中添加词性标注和词典特征之后，F1-measure 提高到 85.80%。模型 ③ 在模型 ② 的基础上引入 NCCD 提高临床药物药品名的识别率，F1-measure 提升到 86.13%。最后在模型 ④ 中加入了可改善词性标注分词错误的 Re-entity 方法，在 CRF 模型中获得了最高的 F1-measure（86.51%）。训练集和测试集中有很多词典中没有的实体，因此单纯使用词典匹配的方法会过度依赖词典规模和词典覆盖率，CRF 模型作为机器学习模型能够学习识别出词典中没有的实体，同时也能根据实体上下文环境对实体作出比较准确的分类。因此以 CRF 模型为基础融合词典、词性标注等特征和 Re-entity 方法可以弥补各单独 NER 模型的不足。

但 CRF 这种传统机器学习方法也有明显的缺点，即过度依赖训练集的规模和质量及有效特征的选择。因此本研究进行了后续神经网络模型的训练。

2.2 BiLSTM-CRF 模型加入预训练向量的实验结果 本研究使用向量训练工具 Word2vec 基于大规模医学专业文本训练了 2 个字符向量集，准备了 2 个训练语料库，第 1 个是电子版的内科学、外科学、妇产科学和儿科学 4 本医学专业书（以下简称 nwfe_emd），第 2 个语料是 2017 年 CCKS 竞赛发布的未标注 EMR 数据（以下简称 unlabel_emd）。用 Python 进行数据清洗，去掉与专业内容无关的部分，然后进行字符粒度分割，分割后作为模型输入在 Word2vec 中进行预训练分别得到 nwfe_emd 和 unlabel_emd 2 个字符向量集。字符分割后训练的向量集规模如表 4 所示。

表 4 语料库规模

Tab 4 Corpus size

Data set	Number of character per piece	Number of vectors Per piece
Nwfe_emd	7 257 743	2 834
Unlabel_emd	7 150 418	1 869

首先验证向量维度对模型效果的影响，本研究基于 unlabel_emd 语料库分别训练了维度为 100、200、300 和 400 的向量集。在 BiLSTM-CRF 模型中基于 epoch 在 40、60 和 80 的参数设置下分别进行不同维度向量的对比实验，结果如图 6 所示。

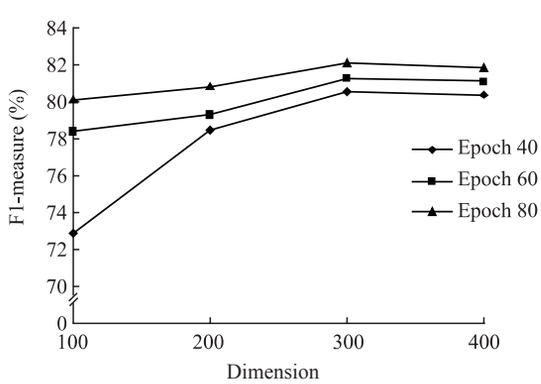


图 6 不同向量维度下实验结果图

Fig 6 Experimental results at different embedding dimensions

由图 6 可知，随着向量维度的增加，3 组 epoch 在不同维度的向量实验中表现出相同的趋势。在一定范围内，F1-measure 随着向量维度的增加呈上升趋势，在维度为 300 时表现出最好的结果，然而在维度为 400 时 F1-measure 相比维度 300 下降，表明并不是向量维度越高模型效果越好，需要根据具体任务要求进行实验，依据具体性能和实验所需时间等方面的需求选择合适的向量维度，因此本实验选择维度 300 向量作为该模型的 baseline 参数进行后续实验。

Lattice-LSTM 模型是在字符粒度神经网络模型 LSTM 模型的基础上结合了词的信息，并且没有经过分词，也避免了分词错误导致的错误累加。

2.3　Lattice-LSTM 中引入预训练向量的实验结果　Lattice-LSTM 模型提供了预训练字符向量集和词向量集，字符向量 gigaword_chn.all.a2b.uni.ite50.vec 是基于大规模标准分词后的中文语料库 Gigaword 使用 Word2vec 工具训练的向量集合，向量集规模为 704 400 个字符和词，包括 5 700 个单字符向量、29 150 个双字符向量和 278 100 个三字符向量。词向量 ctb.50d.vec 是基于 CTB 6.0（Chinese Treebank 6.0）语料库训练得到的。该模型实验在保持词向量 ctb.50d.vec 不变的前提下分别使用 gigaword_chn.all.a2b.uni.ite50.vec、nwfe_emd 和 unlabel_emd 3 个字符向量集进行对比实验，实验结果如表 5 所示。

表 5　Lattice-LSTM 基于不同字符向量实验结果

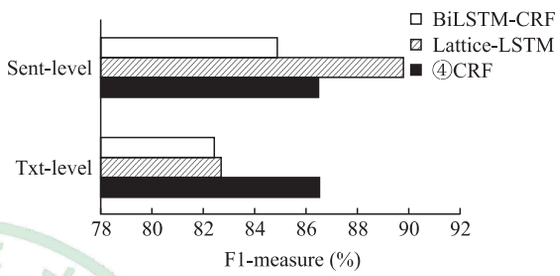
Tab 5　Different character embeddings experiment results of Lattice-LSTM

Dataset	F1-measure (%)
Gigaword_chn.all.a2b.uni.ite50.vec	89.75
Nwfe_emd	89.12
Unlabel_emd	89.70

LSTM: Long short-term memory neural network

从表 5 可知，基于大规模经过标准分词的中文语料库训练的向量集 gigaword_chn.all.a2b.uni.ite50.vec 在 Lattice-LSTM 模型的训练中获得最好效果，F1-measure 为 89.75%。本研究的字符向量是基于字符分割的语料库训练的，F1-measure 稍低于使用 gigaword_chn.all.a2b.uni.ite50.vec 的模型。字符分割后的语料在训练空间向量时相比分词无法充分利用上下文的词信息，训练出的向量会丢失词与上下文之间的语义信息。因此选择使用 gigaword_chn.all.a2b.uni.ite50.vec 向量模型作为后续对比实验的 baseline 模型。

2.4　不同输入长度对比实验结果　分别对 CRF、BiLSTM-CRF 和 Lattice-LSTM 3 种模型开展训练数据输入句子长度变化的实验，txt-level 是以篇为单位，即 1 篇 EMR 作为 1 个输入；sent-level 是以句号“。”为分隔符，即以句号“。”结尾的一个句子为模型的输出。结果如图 7 所示。



	Txt-level	Sent-level
BiLSTM-CRF	82.42%	84.93%
Lattice-LSTM	82.72%	89.75%
④CRF	86.51%	86.50%

图 7　句子级和篇级 F1-measure 比较

Fig 7　Comparison of F1-measure between sent-level and txt-level

BiLSTM: Bi-directional long short-term memory neural network; CRF: Conditional random field; LSTM: Long short-term memory neural network; ④CRF: CRF-POS-Dic-NCCD-Re-entity; POS: Part-of-speech; Dic: Item＋body＋symptom (dictionary); NCCD: Normalized Chinese clinical drug; Sent-level: Sentence level; Txt-level: Text level

由图 7 可见，CRF 对输入句子长度不敏感，句子级和篇级在该模型中的 F1-measure 分别为 86.50% 和 86.51%。BiLSTM-CRF 和 Lattice-LSTM 表现出相同的趋势，句子级的结果均较篇级好。在 BiLSTM-CRF 中句子级的 F1-measure 比篇级高 2.51%，而在 Lattice-LSTM 中句子级比篇级高 7.03%，也证明在基于字符的模型中加入潜在单词信息的有效性。在该模型中随着输入句子长度的增加，Lattice 词格中的词组也在递增，Lattice 的精度会下降[6]，因此句子级的 Lattice-LSTM NER 模型在本实验中得到了 89.75% 的最高值。

2.5　模型实验结果对比　表 6 列出了 CRF、BiLSTM 和 Lattice-LSTM 3 个模型框架中的全部实验结果。前 4 个模型是 CRF 模型框架下的特征

扩展实验结果，模型⑤和⑥是 BiLSTM-CRF 模型框架下的不同输入长度的实验结果，模型 ⑦ 和 ⑧ 是 Lattice-LSTM 模型中的不同输入长度对比实验结果。

表 6　不同模型的实验结果

Tab 6　Results of different models

Model	F1-measure (%)
① CRF-onlychar	76.96
② CRF-POS-Dic	85.80
③ CRF-POS-Dic -NCCD	86.13
④ CRF-POS-Dic-NCCD-Re-entity	86.51
⑤ BiLSTM-CRF (txt-level)	82.42
⑥ BiLSTM-CRF (sent-level)	84.93
⑦ Lattice-LSTM (txt-level)	82.72
⑧ Lattice-LSTM (sent-level)	89.75

CRF: Conditional random field; POS: Part-of-speech; Dic: Item＋body＋symptom (dictionary); NCCD: Normalized Chinese clinical drug; BiLSTM: Bi-directional long short-term memory neural network; Sent-level: Sentence level; LSTM: Long short-term memory neural network; Txt-level: Text level

由表 6 可知，当添加词性标注、词典等特征后，F1-measure 比只用字符特征提高了 8.84%，item、body、symptom 等专业词典的加入可在数据预处理时提高分词准确性及模型的识别准确率。NCCD 是 Wang 等[10]针对国内药物参考国际通用的 Rxnorm 模型建立的 NCCD，本研究在模型 ② 的基础上抽取出药名部分作为词典特征加入，可有效提高药物的 F1-measure。在以上特征的基础上加入 Re-entity 方法，F1-measure 从 86.13% 提高到 86.51%。

在 BiLSTM-CRF 模型中，模型的复杂度会随着模型训练而增加，此时模型在训练集上的训练误差逐渐减小，但当模型复杂到一定程度且数据量不够大时，模型在训练集以外的数据集上的误差反而开始增大，发生过拟合现象[16]。本实验训练数据只有 600 篇，数据量较小，因此该模型在本次任务中只得到 84.93% 的 F1-measure。本研究中，CRF 模型和 BiLSTM-CRF 模型均是基于字符级别的，虽然改善了词级别模型的分词错误，但也漏掉了很多潜在有用的词序列语义信息。Lattice-LSTM 则是在字符级模型的基础上融合了文本中潜在的单词信息，在本任务中获得 89.75% 的 F1-measure，本研究的结果高于 CCKS2018 竞赛任务一“电子病历命名实体识别”所有参赛团队的最优值 89.25%（*https://biendata.com/competition/CCKS2018_1/leaderboard/*）。

3　讨　论

本研究在总结英文领域 NER 的研究进展及详细了解中文 EMR 文本特点的基础上进行中文 EMR NER 研究，在对比现有实验方法后选择 CRF、BiLSTM-CRF 和 Lattice-LSTM 3 个模型框架。首先基于 CRF 进行对比实验：将 EMR 数据处理成 CRF＋＋ 0.58 工具包所需数据格式之后进行模型训练，选择只使用字符特征的模型作为 baseline 模型；然后将自主整理构建的 EMR 基本元素（item）、解剖部位（body）、NCCD 等医学专业词典，作为词典特征引入 baseline 模型进行特征扩展实验；最后在选择最佳特征工程的基础上加入 Re-entity 方法，改善词粒度模型在分词过程中因医学专业性导致的分词错误。其次是神经网络模型的实验研究，神经网络模型可缓解 CRF 模型过度依赖训练集规模和质量及有效特征工程的弊端。在神经网络模型中，本研究利用大规模未标注文本进行了无监督学习，基于不同语料获取了不同参数级别的向量。首先验证了字符向量维度对 BiLSTM-CRF 模型性能的影响，结果显示，该模型在向量维度为 300 时效果最好。在 Lattice-LSTM 模型的向量实验中，使用 Gigaword 语料库训练出的字符向量在 NER 模型中的效果优于基于字符分割语料库构建的向量，因为 Gigaword 语料库训练出的向量集充分利用了词序列的语义信息。

中文 EMR NER 研究仍处于发展阶段，未来仍有很多工作需要开展：（1）本研究实验数据仅有 600 篇已标注训练数据和 400 篇测试数据，数据量小且病种结构单一。在后续工作中我们将扩充数据量并参考英文医学标准库，构建更全面的中文专业词典以辅助系统性能的提高。（2）选择更高质量的语料库进行字符粒度和词粒度向量集的训练，引入 CRF 模型[17]和神经网络模型中进行模型优化，并在已完成 NER 模型的基础上进行 EMR 中药物到 NCCD 的映射工作。

综上所述，本研究在传统 CRF 模型的基础上，提出了一种基于 Re-entity 新分词方法的 CRF 模型，该方法可减少分词在医学专业文本上的错

误，并融合一系列的字符、词性标注、自构建医学专业词典等特征，利用 NCCD 提高药物的识别率，从而整体提高 CRF 模型的中文 EMR 中命名实体的整体识别准确性。在 BiLSTM-CRF 模型中对比了不同维度向量对系统性能的影响，维度为 300 的向量获得最好表现。在 Lattice-LSTM 模型的不同字符向量集的对比实验中，经过金标准分词的大规模中文语料库——Gigaword 语料库训练出的向量能更好地利用上下文中词序列蕴含的潜在语义信息，提高模型识别效果。在字符级神经网络模型的基础上融入单词序列信息构建 Lattice 词格结构并结合 CRF，该模型在独立于分词的情况下能利用文本中潜在的单词信息，其识别准确率较单纯基于字符粒度或基于词粒度的模型均更具优越性。在以上模型对比实验的基础上，对比验证不同输入长度对模型效果的影响，结果表明 CRF 模型对输入句长不敏感，神经网络模型在较短句长下的效果优于较长的篇级输入模型，Lattice-LSTM 在句子级的句子输入实验中获得最高的 F1-measure。

[参 考 文 献]

[1] 杨锦锋,于秋滨,关毅,蒋志鹏. 电子病历命名实体识别和实体关系抽取研究综述[J]. 自动化学报,2014,40:1537-1562.

[2] 叶枫,陈莺莺,周根贵,李昊旻,李莹. 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,2011,30:256-262.

[3] 程健一,关毅,何彬. 基于 SVM 和 CRF 双层分类器的英文电子病历去隐私化[J]. 智能计算机与应用,2016,6:17-19,24.

[4] 张海楠,伍大勇,刘悦,程学旗. 基于深度神经网络的中文命名实体识别[J]. 中文信息学报,2017,31:28-35.

[5] LI H, HAGIWARA M, LI Q, JI H. Comparison of the impact of word segmentation on name tagging for Chinese and Japanese[C/OL]//Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik: LREC, 2014: 2532-2536. [2019-01-28]. *http://www.lrec-conf.org/proceedings/lrec2014/index.html*.

[6] ZHANG Y, YANG J. Chinese NER using Lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers). Melbourne: ACL, 2018: 1554-1564.

[7] 中国中文信息学会语言与知识计算专业委员会. 全国知识图谱与语义计算大会[C/OL]. 天津: CCKS, 2018. [2019-01-28]. *http://www.ccks2018.cn/?page_id=16.%20doi:%20http://www.ccks2018.cn/?page_id=16*.

[8] PAN Y F, HOU X, LIU C L. Text localization in natural scene images based on conditional random field[C/OL]//10th International Conference on Document Analysis and Recognition. Catalonia: ICDR2009, 2009: 6-10. doi: 10.1109/ICDAR.2009.97.

[9] 张祥伟,李智. 基于多特征融合的中文电子病历命名实体识别[J]. 软件导刊,2017,16:128-131.

[10] WANG L, ZHANG Y, JIANG M, WANG J, DONG J, LIU Y, et al. Toward a normalized clinical drug knowledge base in China-applying the RxNorm model to Chinese clinical drugs[J]. J Am Med Inform Assoc, 2018, 25: 809-818.

[11] 曾冠明. 基于条件随机场的中文命名实体识别研究[D]. 北京:北京邮电大学,2009.

[12] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Comput, 1997, 9: 1735-1780.

[13] 李洋,董红斌. 基于 CNN 和 BiLSTM 网络特征融合的文本情感分析[J]. 计算机应用研究,2018,38:3075-3080.

[14] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, KAWAKAMI K, DYER C. Neural architectures for named entity recognition[Z/OL]. arXiv:1603.01360v3 [cs.CL]. (2016-04-07)[2019-01-28]. *https://arxiv.org/pdf/1603.01360.pdf*.

[15] MA X, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: ACL, 2016: 1064-1074.

[16] 陶砾,杨朔,杨威. 深度学习的模型搭建及过拟合问题的研究[J]. 计算机时代,2018(2):14-17,21.

[17] 隋明爽,崔雷. 结合多种特征的 CRF 模型用于化学物质-疾病命名实体识别[J]. 现代图书情报技术,2016(10):91-97.

[本文编辑] 杨亚红