计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2018.20170506
55(7):1548-1556,2018

基于注意机制的化学药物命名实体识别

杨 培 杨志豪 罗 凌 林鸿飞 王 健
(大连理工大学计算机科学与技术学院 辽宁大连 116024)
(yangperasd@mail.dlut.edu.cn)

An Attention-Based Approach for Chemical Compound and Drug Named Entity Recognition

Yang Pei, Yang Zhihao, Luo Ling, Lin Hongfei, and Wang Jian
(*School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning* 116024)

Abstract Recognizing chemical compound and drug name from unstructured data in the field of biomedical text mining is of great significance. The current popular approaches are based on CRF model which needs large amounts of hand-crafted features, and these approaches inevitably have the tagging non-consistency problem (the same mentions in a document are tagged different labels). In this paper, we propose an attention-based BiLSTM-CRF architecture to mitigate these aforementioned drawbacks. First, word embedding is obtained from vast amounts of unlabeled biomedical text. Then the characters of current word are fed to a BiLSTM layer to learn the character representation of this word. After this, word and character representations are transformed to another BiLSTM layer and the current adjacency context representation of this word is generated. Then we use attention mechanism to obtain the current word's context at document level on the basis of the adjacency context of all words in this document and the current word. At last, a CRF layer is used to predict the label sequence of this document according to the integration of the current adjacency context and the document-level context. Experimental results show that our method improves the consistency of mention's label in the same document, and it can also achieve better performance (an F-score of 90.77%) than the state-of-the-art methods on the BioCreative IV CHEMDNER corpus.

Key words long short-term memory (LSTM); attention; conditional random fields (CRF); chemical compound and drug name recognition; deep learning

摘 要 在生物医学文本挖掘领域,化学药物命名实体识别具有重要意义.目前的主流方法是基于条件随机场(conditional random fields, CRF)的方法,但是该方法需要大量的人工特征,并且存在实体标签的全文非一致性问题.针对此问题,提出一种基于注意(Attention)机制的深度学习方法.该方法首先从海量生物文本中学习词向量,然后利用双向长短期记忆网络(BiLSTM)学习字符向量,随后将词向量和字符向量再经过另一个 BiLSTM 以获得词的上下文表示,然后再利用 Attention 机制获得词在全文范围下的上下文表示,最后利用 CRF 层得到整篇文章的标签序列.实验结果表明:相比之前的研究方法,提高了在同一篇文章中实体识别的一致性,并在 BioCreative IV 评测中的 CHEMDNER 数据集上取得了更好的结果(F 值为 90.77%).

收稿日期:2017-07-04;**修回日期**:2017-12-08

基金项目:国家自然科学基金项目(61272373,61572102,61572098);新世纪优秀人才支持计划(NCET-13-0084);国家重点研究计划项目(2016YFC0901902)

This work was supported by the National Natural Science Foundation of China (61272373, 61572102, 61572098), the Project for New Century Excellent Talents in University (NCET-13-0084), and the National Key Research and Development Program of China (2016YFC0901902).

关键词　长短期记忆网络;注意;条件随机场;化学药物命名实体识别;深度学习

中图法分类号　TP391

近年来伴随着生物医学领域的飞速发展,生物医学领域的相关文献也以指数级别快速增长.这也随之促进了生物医学文本挖掘技术的快速发展.而化学药物命名实体识别便是生物医学文本挖掘技术中非常重要的一步.

BioCreative评测是国际上关于生物医学文本挖掘的重要评测,而CHEMDNER(chemical compound and drug name recognition)是其中关于化学药物命名实体识别的一项子任务[1-2].在CHEMDNER这个任务上,很多研究者将该任务转化为序列标注问题,而条件随机场(conditional random fields, CRF)在处理这类问题上有着优秀的性能.因此,大多数研究者都采用CRF模型来处理该任务.Leaman等人[3]针对化学名实体识别提出了tmChem系统,该系统由2个使用了丰富特征的CRF模型组成,并运用了若干后处理,在CHEMDNER任务上该系统获得的最高F值为87.39%.Lu等人[4]同样利用2个CRF模型在CHEMDNER任务上获得了88.06%的F值.这2个CRF模型分别为单词级别的CRF和字符级别的CRF.在单词级别的CRF模型中,它们使用了词聚类的特征.基于传统的统计机器学习的方法在该任务上虽然取得了不错的成绩,但是大多数方法均需要领域专家来设计各种丰富的特征,而这些丰富特征的设计,需要大量的人力、物力,于此同时这些特征往往都与任务密切相关,无法扩展到其他任务.

目前,深度学习在各个领域受到了大量研究者的关注.深度学习主要是利用复杂的网络结构或由多重非线性变换构成的多个处理层对数据进行高层抽象的方法[5].最初,深度学习方法在图像和语音识别领域上获得较大成果,随后被运用到自然语言处理(natural language processing, NLP)领域.在命名实体识别任务上,Collobert等人[6]提出了一种有效的神经网络模型,该模型不再需要大量的人工特征,并且能够从海量的未标注文本中学习词向量,进而有助于模型的训练;Huang等人[7]提出BiLSTM-CRF模型,该模型是BiLSTM与CRF的结合,其中BiLSTM能有效利用当前词的上下文,而CRF层能利用句子级别的标签序列信息;Ma等人[8]提出BiLSTM-CNN-CRF模型,该模型首先利用CNN学习字符级别的表示,然后利用BiLSTM-CRF进行后续处理.与传统的机器学习方法相比,深度学习在表示学习上具有很大的优势,它不需要或只需要少量的特征便可以达到较好的性能.在当今的序列标注问题上,BiLSTM-CRF模型成为了主流的方法.该方法引入了句子级别的似然函数(sentence level log-likelihood),与原始的单词级别的似然函数(word level log-likelihood)——交叉熵相比,该函数考虑了句子中单词标签的相关性.虽然该方法利用了句子级别的信息,但是该方法并没有利用全文信息.目前的方法,无论是基于传统机器学习的方法,还是基于深度学习的方法,它们都没有很好地利用篇章信息,仅仅停留在句子信息.这就造成了实体标签的全文非一致性问题(一篇文章中的相同实体被赋予不同的类别标签)的产生.

针对目前方法的不足,本文提出一种基于Attention机制的BiLSTM-CRF模型(Attended-BiLSTM-CRF).该方法以篇章为基础,首先通过大量无标注数据来学习低维、稠密的词向量;然后利用BiLSTM对文档的词向量和单词字符向量进行处理,抽取每个词的上下文表示;随后利用Attention机制来获得当前词在全文范围内的上下文表示;然后将全文范围内的上下文表示和该词的邻近上下文表示经过融合后送往CRF层;最后利用CRF层来获得这篇文章所对应的全文标签序列.实验结果表明,该模型在CHEMDNER评测数据集上获得了更好的表现.

本文方法的创新点在于利用Attention机制引入了篇章级别的信息,在篇章级别信息的帮助下,单词的类别标签的一致性获得了提升;同时将Attention机制和BiLSTM-CRF结合,并在CHEMDNER评测数据集上取得更高的F值.

1　基于Attention的BiLSTM-CRF的化学药物命名实体识别方法

本节我们首先描述模型使用的词特征和字符特征,然后描述BiLSTM-CRF模型,最后阐述本文提出的Attended-BiLSTM-CRF模型.模型的整体结构如图1所示:

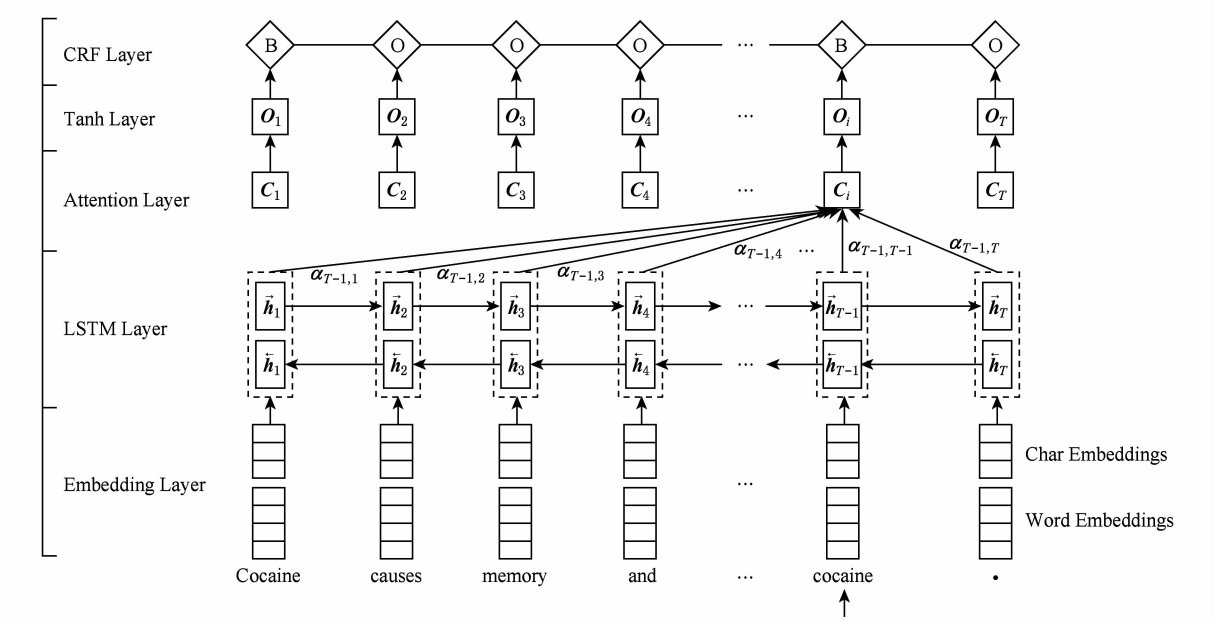

Fig. 1 The model architecture of Attended-BiLSTM-CRF

图 1 Attended-BiLSTM-CRF 模型结构

1.1 表示学习

1.1.1 词向量

目前在利用深度学习处理 NLP 任务上,词向量的表示方法被广泛运用. 词向量是一种分布式的词表示方法,它能够从大量的未标注数据中学习到词的语义和语法信息. 与传统的词袋模型(bag of words, BOW)表示相比,词向量表示具有维度低和稠密的特点. 目前已有很多工具可以用来学习词向量,如 word2vec[9] 和 GolVe[10] 等.

为了获得高质量的词向量,我们从 PubMed 利用"drug"关键字下载了 1 918 662 篇 MEDLINE 摘要. 随后利用 word2vec 工具使用这些无标注的数据进行训练,获得了 50 维的词向量表示. 我们使用这 50 维的词向量来初始化模型的词向量,模型的词向量是可以训练的,伴随训练不断更新.

1.1.2 字符向量

字符向量与词向量存在显著不同. 词向量主要关注词语本身的语义,而字符向量主要关注词语本身的拼写特点(如前缀、后缀等). 利用字符向量和词向量能更好地刻画单词的属性. 此外,化学药名与普通词在字符级别上存在较大的差异(药物命名存在一系列的标准规范,如 IUPAC),因此字符向量可以进一步帮助模型来识别化学药名.

最近,循环神经网络(recurrent neural network, RNN)被广泛用于 NLP 领域,RNN 与传统前馈神经网络相比最大的特点在于:当前时刻的输出不仅依赖于当前时刻的输入,还依赖于前一时刻的输出,即会对前面时刻的信息进行记忆并应用于当前输出的计算中. RNN 这种链式特征对于处理序列化的数据具有很大的优势. 不过由于其存储记忆的结构过于简单,在后向传播算法中随着时间序列长度的增长,会产生梯度逐渐变小直至消失的现象,这样就导致 RNN 无法学习到离当前时刻较远的信息,即无法很好地解决长距离依赖的问题. 在一系列改进的 RNNs 中,目前使用最广泛最成功的模型便是长短期记忆模型(long short-term memory, LSTM[11]),该模型相对于传统的 RNN,主要引入细胞状态来存储记忆而不是依靠单一的隐藏层,这一改动有效的缓解了梯度消失的问题. 在 NLP 领域,例如机器翻译,语言模型等序列标注任务上该模型已经取得了不错的进展. 因此本文利用 LSTM 来学习字符向量. 由于 LSTM 本身序列化的特点,在当前时刻只能获取到上文信息,不能获取到下文信息. 为了能同时获得当前时刻的上文和下文,我们利用 2 个 LSTM 来组成 BiLSTM[12]:一个前向 LSTM,用来正向处理句子;一个反向 LSTM,用来逆向处理句子. 将这 2 个 LSTM 的输出 $\overrightarrow{\boldsymbol{h}}_t$ 和 $\overleftarrow{\boldsymbol{h}}_t$ 拼接到一起,便成为 BiLSTM 在时刻 t 的输出 $[\overrightarrow{\boldsymbol{h}}_t;\overleftarrow{\boldsymbol{h}}_t]$. 本文使用的 LSTM 的实现为

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{xi}\boldsymbol{x}_t+\boldsymbol{W}_{hi}\boldsymbol{h}_{t-1}+\boldsymbol{W}_{ci}\boldsymbol{c}_{t-1}+\boldsymbol{b}_i), \quad (1)$$

$$\boldsymbol{c}_t=(1-\boldsymbol{i}_t)\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\tanh(\boldsymbol{W}_{xc}\boldsymbol{x}_t+\boldsymbol{W}_{hc}\boldsymbol{h}_{t-1}+\boldsymbol{b}_c), \quad (2)$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{xo}\boldsymbol{x}_t+\boldsymbol{W}_{ho}\boldsymbol{h}_{t-1}+\boldsymbol{W}_{co}\boldsymbol{c}_t+\boldsymbol{b}_o),\quad(3)$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_t).\quad(4)$$

其中,σ 是非线性函数,本文使用 hard sigmoid 函数,$\{\boldsymbol{W}_{xi},\boldsymbol{W}_{hi},\boldsymbol{W}_{ci},\boldsymbol{W}_{xc},\boldsymbol{W}_{hc},\boldsymbol{W}_{xo},\boldsymbol{W}_{ho},\boldsymbol{W}_{co}\}$是 LSTM 的参数矩阵,$\{\boldsymbol{b}_i,\boldsymbol{b}_c,\boldsymbol{b}_o\}$是 LSTM 的偏置项.

本文首先为每个字符随机初始化一个 25 维的向量,然后将当前词对应的字符的向量分别顺序和逆序输入到 BiLSTM 中,最后的输出$[\overrightarrow{\boldsymbol{h}_t};\overleftarrow{\boldsymbol{h}_t}]$便是当前词的字符向量.

1.2　BiLSTM-CRF

对于一篇文章,经过上述处理后,便可以获得当前词的词向量和字符向量,随后将所有词的这 2 部分直接拼接后输入到另一个 BiLSTM 中,便可以得到每个词的上下文表示.

在序列标注任务上,BIO(begin,inside,outside)标签机制被广泛使用,因此本文也采用该标签机制.在实际文本中,化学药物命名实体通常不是由一个单词组成,而是由多个单词组成.这样,利用 BIO 标签机制后,一个实体的第 1 个词被赋予 B 标签,第 2 个词到最后一个词都被赋予 I 标签,同时实体之前和实体之后的词都被赋予 O 标签.由此可知,在实际的标注序列中,B,I,O 标签并不是任意出现,它们之间存在紧密的逻辑联系,即一个单词的实体标签不仅受到该单词的上下文环境和该词本身含义的影响,而且还受到该单词的上下文标签的影响.但是在普通的序列标注模型中并没有考虑到这种约束,它们在对当前词的标签判断上,仅仅使用当前词的上下文,并没有使用当前词标签的上下文.因此,在某些情况下会产生不可能的标签序列.如 I 标签出现在 O 标签之后等.为了进一步提高实体识别的准确性,这里我们借鉴 Collobert 等人[6]和 Huang 等人[7]的工作,结合 CRF 模型考虑标签转移概率的优点,我们在原始的 BiLSTM 基础上加入整个句子的标签转移信息.

首先,我们定义一个标签转移矩阵 $\boldsymbol{A}$,这里 $A_{i,j}$ 代表从标签 i 转移到标签 j 的得分,这是一个随模型一起训练的参数.定义 θ 为原始 BiLSTM 需要学习的参数,则 $\theta'=\theta\cup\{A_{i,j}\ \forall i,j\}$ 是整个模型要学习的所有参数.给定一个句子$[\boldsymbol{x}]_1^T$,T 为句子长度,定义$[f_\theta]_{i,t}$是第 t 个词、第 i 个标签的 BiLSTM 的输出分值,则一个句子在给定标签序列$[\boldsymbol{i}]_1^T$的总得分计算为

$$S([\boldsymbol{x}]_1^T,[\boldsymbol{i}]_1^T,\theta')=\sum_{t=1}^{T}(A_{[\boldsymbol{i}]_{t-1},[\boldsymbol{i}]_t}+[f_\theta]_{[\boldsymbol{i}]_t,t}).\quad(5)$$

最后我们使用函数 $softmax$ 计算在一个句子$[\boldsymbol{x}]_1^T$上真实标签序列$[\boldsymbol{y}]_1^T$的概率为

$$p([\boldsymbol{y}]_1^T\mid[\boldsymbol{x}]_1^T,\theta')=\frac{e^{S([\boldsymbol{x}]_1^T,[\boldsymbol{y}]_1^T,\theta')}}{\sum_{\boldsymbol{j}}e^{S([\boldsymbol{x}]_1^T,[\boldsymbol{j}]_1^T,\theta')}}.\quad(6)$$

这里$[\boldsymbol{j}]_1^T$表示所有可能的标签序列.最后,我们通过最大化对数似然概率来训练模型参数,计算为

$$\ln P([\boldsymbol{y}]_1^T\mid[\boldsymbol{x}]_1^T,\theta')=S([\boldsymbol{x}]_1^T,[\boldsymbol{y}]_1^T,\theta')-\ln\sum_{\forall[\boldsymbol{j}]_1^T}e^{S([\boldsymbol{x}]_1^T,[\boldsymbol{\varphi}]_1^T,\theta')}.\quad(7)$$

我们使用随机梯度下降法(stochastic gradient descent, SGD)来优化参数.

训练结束后,在预测标签时,我们需要寻找得分最高的标签序列作为预测标签序列,即:

$$\arg\max_{[\boldsymbol{j}]_1^T}S([\boldsymbol{x}]_1^T,[\boldsymbol{j}]_1^T,\theta').\quad(8)$$

这里本文采用维特比算法[13]来找到最佳标签序列.

1.3　Attention 机制

迄今基于深度学习和传统机器学习的方法中,无法避免的一个问题是单词标签的全文非一致性:在一篇文章中,相同的词、相同的实体却常常被模型赋予不同的实体标签.显然,这会降低模型的正确率,也不易于实际的工程使用.这种问题发生的主要原因在于现今的模型通常以句子作为单独的处理单元.在一个单独的处理单元中,模型根据该词的上下文来赋予标签,即这些模型仅仅利用了句子信息,是句子级别的方法.在同一篇文章中,如果同一实体在不同句子中的上下文不同,则句子级别模型所赋予的标签也会不同,这便是单词标签的全文不一致性产生的原因.同时,如果在同一篇文章中,对于同一个实体,在其众多的上下文中,如果只有唯一一处或几处的上下文对该实体的标签类别的判断起决定性的作用,则现今的句子级别的方法也不能很好地处理该问题.

句子级别方法在化学药名识别这一具体任务上,同样存在单词标签非一致性的问题.同时,单词标签非一致性在这一任务上的另一个体现便是缩写识别准确率的低下.在一篇文章中,作者通常只会在第一次提到缩写时给出该缩写的全称,通常普通模型能够根据此处的上下文对该缩写赋予正确的实体标签.在其后提到该实体时,作者通常不会给出全称,而仅仅给出该实体的缩写,而此处的上下文和该缩写的拼写与该缩写所代指的实体之间的关系较弱,普通模型无法仅根据这些信息作出正确的实体标签

类别判断. 模型要作出正确的判断,必须要从全文找到该缩写所对应的全称,并获得此处的上下文信息. 因此,只有引入篇章信息才能很好地解决这类问题.

在解决单词标签非一致性问题上,研究者通常采用基于规则的后处理. 但是规则的制定较为复杂,也无法根据任务学习、变化,而且很难制定出完善的规则. 对于该问题本文提出利用 Attention 机制来引入篇章信息,在篇章信息的帮助下,通过模型的不断学习来缓解该问题. Attention 机制最早被运用到图像领域,随后被运用到 NLP 领域,但目前并没有研究者将 Attention 机制运用到化学药物命名实体识别任务上. 在图像领域,Attention 机制主要是模拟人的注意机制[14]:人在观察一副图像时,并不会将自己的注意力平均地分散到整幅图像的每个部位,而大多是根据需求将注意力集中到图像的特定部分,例如看人物肖像时,通常会将注意力集中到脸部. 在本文中,对于每一个词,拟使用 Attention 机制来获取篇章级信息,进而改善相同词的标签的全文非一致性问题.

具体地,对于一篇文章$[\boldsymbol{s}]_1^N$,N 表示句子数,$[\boldsymbol{x}]_1^T$ 表示其中的一个句子,T 为句子长度. 我们定义 $\boldsymbol{attended}$ 为$[\boldsymbol{s}]_1^N$ 的词向量或字符向量以及它们的组合;定义 $\boldsymbol{state}_i$ 为第 i 个词在 $\boldsymbol{attended}$ 中相对应的一项;定义 $\boldsymbol{source}$ 为全文每个词对应的上下文,即$[\boldsymbol{s}]_1^N$ 经过 BiLSTM 的输出. 则可以用公式获得第 i 个词在全文范围内所应该分配的注意力 $\boldsymbol{\alpha}_i$:

$$\boldsymbol{energy}_i = f(\boldsymbol{attended}, \boldsymbol{state}_i, \boldsymbol{W}), \tag{9}$$

$$\boldsymbol{\alpha}_i = softmax(\boldsymbol{energy}_i), \tag{10}$$

其中,$f(\cdot)$ 是用于衡量 $\boldsymbol{state}_i$ 与 $\boldsymbol{attended}$ 之间相关性的函数,函数中的参数 $\boldsymbol{W}$ 随模型一同训练. 本文使用曼哈顿距离作为衡量相关性的函数. 由于 $\boldsymbol{a}$ 与自己的距离为 0,同时不同词的词义相关性越弱,其曼哈顿距离也就越大,但我们需要词义越相近的词的能量越大,故在实现中我们利用 $\max(\boldsymbol{energy}_i) - \boldsymbol{energy}_i$ 来修正所需的能量. 本文使用曼哈顿距离:

$$d(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{W}) = \sum_{i=1}^{N} w_i |a_i - b_i|. \tag{11}$$

在实现中,我们对 $\boldsymbol{W}$ 全部初始化为 $\boldsymbol{1}$,并且在训练过程中保持为正数.

随后,我们利用得到的注意力权重 $\boldsymbol{\alpha}_i$ 对 $\boldsymbol{source}$ 中的信息进行重新筛选和融合,获得当前词在全文范围下的上下文,这里我们定义为 $\boldsymbol{glimpse}_i$:

$$\boldsymbol{glimpse}_i = \boldsymbol{\alpha}_i^{\mathrm{T}}\, \boldsymbol{source}. \tag{12}$$

为了使 Attention 模型更容易训练,同时当前词的实体标签不仅取决于全文范围内的上下文信息,还应该取决于当前词邻近的上下文信息,因此,我们将 $\boldsymbol{glimpse}_i$ 与 $\boldsymbol{source}_i$ 相结合后,输入到后续的模型结构中:

$$\boldsymbol{context}_i = g(\boldsymbol{glimpse}_i, \boldsymbol{source}_i, \boldsymbol{U}). \tag{13}$$

其中,$g(\cdot)$为非线性函数,本文使用 tanh,$\boldsymbol{U}$ 为随模型训练的参数.

利用 Attention 机制和前文所述的 BiLSTM,对于每一篇文章$[\boldsymbol{s}]_1^N$ 便可以得到 $\sum_N \sum_T \boldsymbol{context}$(在图 1 中简写为 $\boldsymbol{C}$),随后再经过一个 tanh 层,便得到模型对于该文档的每个词在每个标签类别上的得分,记为 $\sum_N \sum_T \boldsymbol{output}$(在图 1 中简写为 $\boldsymbol{O}$),最后可计算文章$[\boldsymbol{s}]_1^N$ 在给定标签序列 $\sum_M [\boldsymbol{m}]_1^T$ 下的总得分:

$$S([\boldsymbol{s}]_1^N, \sum_M [\boldsymbol{m}]_1^T, \theta') = \sum_{\boldsymbol{M}} \sum_{t=1}^{T} (A_{[\boldsymbol{m}]_{t-1}, [\boldsymbol{m}]_t} + [\boldsymbol{output}]_{[\boldsymbol{m}]_t, t}). \tag{14}$$

式(14)与式(15)的区别是分别计算$[\boldsymbol{s}]_1^N$ 中每一个$[\boldsymbol{x}]_1^T$ 的得分,然后将所有$[\boldsymbol{x}]_1^T$ 的得分相加获得$[\boldsymbol{s}]_1^N$ 的总得分. 随后,与 1.2 节相同,使用函数 $softmax$ 获得概率后,通过最大化对数似然概率来训练模型参数.

在预测阶段,与 1.2 节不同的是,对每个句子分别采用维特比解码.

$$\sum_{\boldsymbol{M}} \arg\max_{[\boldsymbol{m}]_1^T} S([\boldsymbol{x}]_1^T, [\boldsymbol{m}]_1^T, \theta'). \tag{15}$$

2 实验分析

2.1 实验设置

本文在 BioCreative IV 命名实体识别子任务 CHEMDNER 的数据集上进行实验[1]. 表 1 展示了 CHEMDNER 原始数据集的构成:

Table 1 CHEMDNER Corpus Analysis

表 1 CHEMDNER 语料的数据统计

Item	Training Set	Development Set	Evaluation Set
Abstracts	3 500	3 500	3 000
Chemicals	8 520	8 677	7 563

实验中,我们将训练集(training set)和开发集(development set)合并,并从中随机抽取 10%作为新的开发集,并利用新的开发集来调参,测试集

(evaluation set)保持不变.在结果评估上,我们采用序列标注中常用的准确率(precision, P)、召回率(recall, R)、F 值(F-score, F)作为实验数据的评价指标.表 2 展示了本文模型使用的超参数:

Table 2　The Hyper-Parameters of Model

表 2　实验中模型的超参数列表

Parameters	Description	Value
word_embedding_dim	the dimension of word embedding layer	50
char_embedding_dim	the dimension of char embedding layer	25
char_for_lstm_dim	the dimension of forward char LSTM layer	25
char_rev_lstm_dim	the dimension of reverse char LSTM layer	25
for_lstm_dim	the dimension of forward LSTM layer	100
rev_lstm_dim	the dimension of reverse char LSTM layer	100
tanh_dim	the dimension of tanh layer	3
learning_rate	learning rate	0.001

2.2　实验结果

我们将目前的主流方法 BiLSTM-CRF 与加入 Attention 机制的 BiLSTM-CRF 进行对比.此外,为了探索词特征和字符特征对模型性能的影响,对于词特征和字符特征我们分别采用了 2 种形式的处理方式:1)词特征或字符特征是否经过 Attention 层处理,即 Attention 利用何种特征来决定在全文范围内的对齐;2)词特征和字符特征是否经过 LSTM 层,即是否将词特征或字符特征用于最后的分类.实验结果如表 3 所示:

Table 3　Experment Results for the Feature of Attention

表 3　Attention 特征的实验结果

Model	Word LSTM	Char LSTM	Word Attention	Char Attention	P/%	R/%	F/%
BiLSTM-CRF	○	×	×	×	88.28	87.06	87.66
BiLSTM-CRF	×	○	×	×	84.96	81.77	83.33
BiLSTM-CRF	○	○	×	×	91.31	87.73	89.48
Attended-BiLSTM-CRF	○	×	○	×	88.45	87.15	87.8
Attended-BiLSTM-CRF	×	○	×	○	86.15	83.34	84.72
Attended-BiLSTM-CRF	○	○	○	×	91.09	90.25	90.67
Attended-BiLSTM-CRF	○	○	×	○	91.4	90.15	90.77
Attended-BiLSTM-CRF	○	○	○	○	91.23	89.92	90.57

Note: "○" indicates that our model uses this feature. "×" indicates that our model does not use this feature. Word LSTM or char LSTM indicates that our model whether use LSTM layer to process word feature or char feature. Word attention or char attention indicates that our model whether use attention layer to process word feature or char feature.

由表 3 我们可以获得:

1) 无论是词特征、字符特征或它们的组合,运用 Attention 机制均能提高性能.

2) 在 LSTM 层,单独的词特征要比单独的字符特征好,而同时运用词特征和字符特征能进一步提高性能.

3) 在 Attention 层,单独的词特征要比单独的字符特征差,而同时运用词特征和字符特征反而会降低性能.

对于以上结果,我们分析如下:

1) Attention 机制能够学习到篇章级的信息,能够帮助模型提高全文一致性和缩写识别的准确率.在进行结果分析时,我们得到如表 4 的结果.表 4 中有背景颜色的单词是模型识别出的实体.由于篇幅限制,我们省略了文中的部分句子和除 ANIT 以外的标注.从表 4 中可以看到,本文提出的模型在缩写识别上优于 BiLSTM-CRF 模型.

2) 在做标签类别判断时,主要依靠词义,而不

是字符含义,同时词义与字符含义并不是相互对斥,而是可以相互互补.所以在 LSTM 层,单独词特征好于单独的字符特征,而同时使用也能提高性能.

3) 由于在词级别存在未登录词,而化学名中也存在大量的未登录词,所以在使用词来做 Attention 时,未登录词会影响 Attention 权重 $\boldsymbol{\alpha}_i$ 的生成,造成权重分配不恰当;而字符却不存在这种问题.所以在 Attention 层单独的字符特征要好于词特征.而这两者同时使用时,模型无法完全消除词特征的缺点,所以造成联合使用时性能有所下降.

为了展示 Attention 权重的学习效果,我们可视化了 Attention 权重,如图 2 所示.该表是当前词为 Retinoic 时,全文每个单词所获得的权重.由于全文单词过多,无法一一显示,故做相应的省略.

Table 4 An Sample of Tagging Consistency

表 4 标签一致性展示

Attended-BiLSTM-CRF	BiLSTM-CRF
… Here, we found that mice lacking A(1) AR were resistant to alpha-naphthyl isothiocyanate (ANIT)-induced liver injury, as evidenced by lower serum liver enzyme levels and reduced extent of histological necrosis. … In the kidney, A(1) AR deficiency prevented the decrease of glomerular filtration rate caused by ANIT. Treatment of WT mice with A(1) AR antagonist DPCPX also protected against ANIT hepatotoxicity. Our results indicated that lack of A(1) AR gene protects mice from ANIT-induced cholestasis by enhancing toxic biliary constituents efflux through biliary excretory route and renal elimination system and suggested a potential role of A(1) AR as therapeutic target for the treatment of intrahepatic cholestasis.	… Here, we found that mice lacking A(1) AR were resistant to alpha-naphthyl isothiocyanate (ANIT)-induced liver injury, as evidenced by lower serum liver enzyme levels and reduced extent of histological necrosis. … In the kidney, A(1) AR deficiency prevented the decrease of glomerular filtration rate caused by ANIT. Treatment of WT mice with A(1) AR antagonist DPCPX also protected against ANIT hepatotoxicity. Our results indicated that lack of A(1) AR gene protects mice from ANIT-induced cholestasis by enhancing toxic biliary constituents efflux through biliary excretory route and renal elimination system and suggested a potential role of A(1) AR as therapeutic target for the treatment of intrahepatic cholestasis.

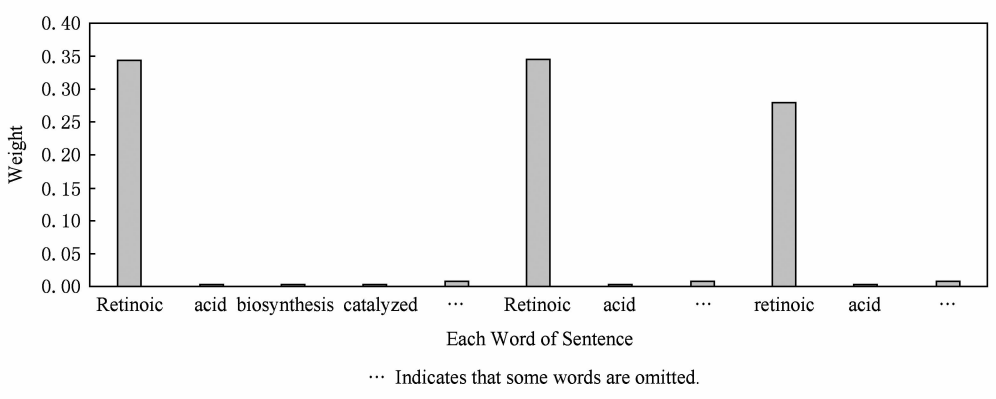

Fig. 2 A sample's attention weight from CHEMDNER dataset

图 2 CHEMDNER 数据集中某一样本的 Attention 权重展示

2.3 方法性能对比实验

为了验证 Attended-BiLSTM-CRF 的性能,我们和其他现存的一些方法进行了对比,实验结果如表 5 所示:

Table 5 The Results of Different Method

表 5 各种方法对比实验结果

Method	P/%	R/%	F/%	Δ/%
CRF model of Ref [3]	89.09	85.75	87.39	3.87
CRF model of Ref [4]	88.72	87.41	88.06	3.10
BiLSTM-CRF	91.31	87.73	89.48	1.48
Attended-BiLSTM-CRF	91.4	90.15	90.77	

Δ indicates the percentage of promotion in F-value.

对比方法简介:

1) CRF 模型[3].利用多个 CRF 组成的 tmChem 系统,并作相应的后处理,如缩写识别等,该模型在 BioCreative CHEMDNER 测评任务上排名第 1.

2) CRF 模型[4].同样利用多个 CRF 组成的系统,并运用了词聚类特征和使用了相应后处理,如括号匹配等,该模型在 BioCreative CHEMDNER 测评任务上排名第 2,但测评结束后利用多粒度的词聚类进一步提高模型性能达到88.06%.

3) BiLSTM-CRF.利用 BiLSTM-CRF 构成的模型,仅仅使用词和字符向量特征,未做后处理.

4) Attended-BiLSTM-CRF. 本文提出的基于 Attention 的 BiLSTM-CRF 的模型,未做后处理.

以上模型可以分为:传统机器学习方法(前 2 种方法)和深度学习方法(后 2 种方法).从实验结果来看,深度学习方法好于传统机器学习方法,而且在特征构造上也无需大量的人工特征,同时也无需复杂的后处理. 本文提出的 Attended-BiLSTM-CRF 模型相比其他方法在 F 值上均有提高,相比传统的机器学习方法 F 值提高超过 3%,同样相比目前主流的 BiLSTM-CRF 模型 F 值提高 1.48%. 在表 5 中,可以得到:本文提出的方法与 BiLSTM-CRF 方法最大的不同在于保持 P 值基本不变的情况下,大幅度提高 R 值,这从侧面显示本文提出的方法提高了标签的一致性. 同时,在对 Attended-BiLSTM-CRF 模型的结果分析中,我们发现单词标签的一致性的确得到提高,缩写识别的正确率也得到提高.

3　总结与展望

本文提出了一种用于化学药物命名实体识别的 Attended-BiLSTM-CRF 方法. 实验结果表明 Attended-BiLSTM-CRF 相比现存的方法均能获得更好的结果,这主要有 3 个原因:

1) 低维、稠密的词向量和字符向量比传统的机器学习有更好的表现能力,同时深度模型如 LSTM 能更好地学习到高层的抽象信息;

2) CRF 层的引入,降低了不可能出现的类别序列,利用了句子级别的标签之间的依赖信息;

3) Attention 机制利用了篇章级别的信息,有效降低了单词标签的全文的非一致性,同时,也带来了缩写识别的准确率的提高.

但是深度学习和 Attention 机制在生物医学文本挖掘中仍然有很大的提升空间. 从本文的实验结果可以看出 Attention 机制对利用篇章级别的信息很有帮助. 将来我们会进一步探索 Attention 机制在生物医学文本挖掘中的应用.

参　考　文　献

[1] Krallinger M, Leitner F, Rabal O, et al. CHEMDNER: The drugs and chemical names extraction challenge [J]. Journal of Cheminformatics, 2015, 7(S): 1-11

[2] Krallinger M, Rabal O, Lourenço A, et al. Overview of the CHEMDNER patents task [C/OL] //Proc of the 15th BioCreative Challenge Evaluation Workshop. 2015 [2017-07-04]. https://jcheminf. springeropen. com/articles/10. 1186/1758-2946-7-S1-S1

[3] Leaman R, Wei Chih-Hsuan, Lu Zhiyong. NCBI at the BioCreative IV CHEMDNER Task: Recognizing chemical names in PubMed articles with tmChem [C/OL] //Proc of the 15th BioCreative Challenge Evaluation Workshop. 2013 [2017-07-04]. http://www. biocreative. org/media/store/files/2013/bc4_v2_2. pdf

[4] Lu Yanan, Ji Donghong, Yao Xiaoyuan, et al. CHEMDNER system with mixed conditional random fields and multi-scale word clustering [J]. Journal of Cheminformatics, 2015, 7(S): 4-9

[5] Lecun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436-444

[6] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12(8): 2493-2537

[7] Huang Zhiheng, Xu Wei, Yu Kai. Bidirectional LSTM-CRF models for sequence tagging [OL]. [2017-07-04]. https://arxiv. org/abs/1508. 01991

[8] Ma Xuezhe, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF [C] //Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2016

[9] Mikolov T, Sutskever I, Chen Kai, et al. Distributed representations of words and phrases and their compositionality [C/OL] //Proc the 26th Advances in Neural Information Processing Systems. 2013 [2017-07-04]. https://arxiv. org/pdf/1310. 4546. pdf

[10] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation [C/OL] //Proc of the 2014 Conf on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532-1543 [2017-07-04]. http://www. aclweb. org/anthology/D14-1162

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

[12] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Networks, 2005, 18(5): 602-610

[13] Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm [J]. IEEE Trans on Information Theory, 1967, 13(2): 260-269

[14] Mnih V, Heess N, Graves A. Recurrent models of visual attention [C/OL] //Proc of the 27th Advances in Neural Information Processing Systems. 2014 [2017-07-04]. http://papers. nips. cc/paper/5542-recurrent-models-of-visual-attention. pdf

Yang Pei, born in 1991. Master candidate. His main research interests include information retrieval and deep learning.

Yang Zhihao, born in 1973. PhD, professor and PhD supervisor. His main research interests include text mining, machine learning and natural language processing.

Lin Hongfei, born in 1962. PhD, professor and PhD supervisor. His main research interests include information retrieval and data mining natural language understanding.

Luo Ling, born in 1988. PhD candidate. His main research interests include information extraction and deep learning.

Wang Jian, born in 1967. PhD, professor and PhD supervisor. Her main research interests include text mining, machine learning and natural language processing.

2016 年《计算机研究与发展》高被引论文 TOP10

排名	论文信息
1	刘峤，李杨，段宏，刘瑶，秦志光. **知识图谱构建技术综述**[J]. 计算机研究与发展，2016，53(3)：582-600 Liu Qiao, Li Yang, Duan Hong, Liu Yao, Qin Zhiguang. Knowledge Graph Construction Techniques [J]. Journal of Computer Research and Development, 2016, 53(3): 582-600
2	刘知远，孙茂松，林衍凯，谢若冰. **知识表示学习研究进展**[J]. 计算机研究与发展，2016，53(2)：247-261 Liu Zhiyuan, Sun Maosong, Lin Yankai, Xie Ruobing. Knowledge Representation Learning: A Review [J]. Journal of Computer Research and Development, 2016, 53(2): 247-261
3	张蕾，章毅. **大数据分析的无限深度神经网络方法**[J]. 计算机研究与发展，2016，53(1)：68-79 Zhang Lei, Zhang Yi. Big Data Analysis by Infinite Deep Neural Networks [J]. Journal of Computer Research and Development, 2016, 53(1): 68-79
4	王兴伟，李婕，谭振华，马连博，李福亮，黄敏. **面向“互联网＋”的网络技术发展现状与未来趋势**[J]. 计算机研究与发展，2016，53(4)：729-741 Wang Xingwei, Li Jie, Tan Zhenhua, Ma Lianbo, Li Fuliang, Huang Min. The State of the Art and Future Tendency of “Internet＋” Oriented Network Technology [J]. Journal of Computer Research and Development, 2016, 53(4): 729-741
5	孟小峰，杜治娟. **大数据融合研究：问题与挑战**[J]. 计算机研究与发展，2016，53(2)：231-246 Meng Xiaofeng and Du Zhijuan. Research on the Big Data Fusion: Issues and Challenges [J]. Journal of Computer Research and Development, 2016, 53(2): 231-246
6	甘丽新，万常选，刘德喜，钟青，江腾蛟. **基于句法语义特征的中文实体关系抽取**[J]. 计算机研究与发展，2016，53(2)：284-302 Gan Lixin, Wan Changxuan, Liu Dexi, Zhong Qing, Jiang Tengjiao. Chinese Named Entity Relation Extraction Based on Syntactic and Semantic Features [J]. Journal of Computer Research and Development, 2016, 53(2): 284-302
7	单言虎，张彰，黄凯奇. **人的视觉行为识别研究回顾、现状及展望**[J]. 计算机研究与发展，2016，53(1)：93-112 Shan Yanhu, Zhang Zhang, Huang Kaiqi. Visual Human Action Recognition: History, Status and Prospects [J]. Journal of Computer Research and Development, 2016, 53(1): 93-112
8	庄严，李国良，冯建华. **知识库实体对齐技术综述**[J]. 计算机研究与发展，2016，53(1)：165-192 Zhuang Yan, Li Guoliang, Feng Jianhua. A Survey on Entity Alignment of Knowledge Base [J]. Journal of Computer Research and Development, 2016, 53(1): 165-192
9	付志耀，高岭，孙骞，李洋，高妮. **基于粗糙集的漏洞属性约简及严重性评估**[J]. 计算机研究与发展，2016，53(5)：1009-1017 Fu Zhiyao, Gao Ling, Sun Qian, Li Yang, Gao Ni. Evaluation of Vulnerability Severity Based on Rough Sets and Attributes Reduction [J]. Journal of Computer Research and Development, 2016, 53(5): 1009-1017
10	曹珍富，董晓蕾，周俊，沈佳辰，宁建廷，巩俊卿. **大数据安全与隐私保护研究进展**[J]. 计算机研究与发展，2016，53(10)：2137-2151 Cao Zhenfu, Dong Xiaolei, Zhou Jun, Shen Jiachen, Ning Jianting, Gong Junqing. Research Advances on Big Data Security and Privacy Preserving [J]. Journal of Computer Research and Development, 2016, 53(10): 2137-2151

数据来源：CSCD，中国知网；统计日期：2017 年 11 月 27 日