

基于命名实体识别任务反馈增强的中文突发事件抽取方法

武国亮, 徐继宁*

(北方工业大学 电气与控制工程学院, 北京 100144)

(* 通信作者电子邮箱 jxu0422@ncut.edu.cn)

摘要: 针对基于双向长短期记忆网络-条件随机场(BiLSTM-CRF)的事件抽取模型仅能获取字粒度语义信息, 可学习特征维度较低致使模型上限低的问题, 以开放领域的中文公共突发事件数据为研究对象, 提出了一种基于命名实体识别任务反馈增强的中文突发事件抽取方法FB-Lattice-BiLSTM-CRF。首先, 将Lattice(点阵)机制融合双向长短期记忆(BiLSTM)网络作为模型的共享层, 获取句子中的词语语义特征; 其次, 增加命名实体识别辅助任务, 以联合学习和挖掘实体语义信息, 同时将命名实体识别任务的输出反馈到输入端, 提取其中实体对应的分词结果作为Lattice机制的外输入, 以减少该机制自组词数量大带来的运算负荷并进一步强化对实体语义特征的提取; 最后, 通过最大化同方差不确定性的最大高斯似然估计方法计算模型总损失, 从而解决多任务联合学习产生的损失不平衡问题。实验结果表明, FB-Lattice-BiLSTM-CRF在测试集上的准确率达到81.25%, 召回率达到76.50%, F1值达到78.80%, 较基准模型分别提升7.63、4.41和5.95个百分点, 验证了该方法对基准模型改进的有效性。

关键词: 中文突发事件; 事件抽取; 命名实体识别; 多任务学习; 点阵双向长短期记忆网络; 损失平衡

中图分类号: TP389.1 **文献标志码:** A

Chinese emergency event extraction method based on named entity recognition task feedback enhancement

WU Guoliang, XU Jining*

(School of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China)

Abstract: Aiming at the problem that the Bidirectional Long Short-Term Memory network-Conditional Random Field (BiLSTM-CRF) based event extraction model can only obtain the semantic information of character granularity, and the upper limit of the model is low due to the low dimensionality of learnable features, a Chinese emergency event extraction method based on named entity recognition task feedback enhancement was proposed by taking the Chinese public emergency event data in open field as the research object, namely FeedBack-Lattice-Bidirectional Long Short-Term Memory network-Conditional Random Field (FB-Lattice-BiLSTM-CRF). Firstly, the Lattice mechanism was integrated with Bidirectional Long Short-Term Memory network (BiLSTM) as the sharing layer of the model to obtain the semantic features of words in sentences. Secondly, the named entity recognition auxiliary task was added to jointly learn and mine entity semantic information. At the same time, the output of the named entity recognition task was fed back to the input end, and the word segmentation results corresponding to the entities were extracted as the external input of the Lattice mechanism, so as to reduce the computing overhead brought by the large number of self-formed words of the mechanism and further enhance the extraction of entity semantic features. Finally, the total loss of the model was calculated by the maximum Gaussian likelihood estimation method to maximize the homoscedasticity uncertainty, so as to solve the problem of loss imbalance caused by multi-task joint learning. Experimental results show that FB-Lattice-BiLSTM-CRF has the accuracy of 81.25%, the recall of 76.50%, and the F1 value of 78.80% on the test set, which are 7.63, 4.41 and 5.95 percentage points higher than those of the benchmark model, respectively, verifying the effectiveness of the improvement performing to the benchmark model.

Key words: Chinese emergency event; event extraction; named entity recognition; multi-task learning; Lattice-Bidirectional Long Short-Term Memory network (Lattice-BiLSTM); loss balance

0 引言

近年来,公共安全突发事件频发。网络的普及使得公共事件相关的有价值的信息散布在各类网络文本中,事件抽取(Event Extraction, EE)在这类信息的挖掘和提取方面有着极其重要的作用。事件发生时,对社交媒体上的相关事件描述

文本进行事件抽取,能够及时地了解事态信息,使应急部门快速做出反应。在事件发生后,通过事件抽取对报道中事件信息的结构化提取,建立相应的案例库,从而为下一步的事件分析和制定防范预案做有力的支撑。

事件抽取主要分为模式匹配方法^[1-4]和机器学习方法^[5-8]。

收稿日期: 2020-09-25; **修回日期:** 2020-12-04; **录用日期:** 2020-12-07。 **基金项目:** 国家重点研发计划项目(2018YFC0807000)。

作者简介: 武国亮(1996—),男,北京人,硕士研究生,主要研究方向:自然语言处理、事件抽取; 徐继宁(1970—),女,陕西西安人,副教授,博士,主要研究方向:自动控制系统、现场总线。

在基于机器学习的事件抽取方法中,数据和特征决定模型学习的上限,因此提高数据质量和优化特征选择显得尤为重要。前者属于数据采集和数据预处理范畴,难以通过对模型的改进来提高,但后者可以通过语义特征挖掘来实现。命名实体识别(Named Entity Recognition, NER)和事件抽取是信息抽取的两个子任务,实体是事件抽取内容的组成部分,基于此学者们通过命名实体识别进行实体语义信息的挖掘,进而提高事件抽取任务的学习效果。Li等^[5]提出了一个结构化预测模型,能同时实现实体、关系和事件三个信息抽取任务。Yang等^[6]提出了篇章内的事件和实体联合抽取模型,通过采用联合因子图模型来联合学习每个事件内部的结构化信息、篇章内不同事件间的关系和实体信息。吴文涛等^[7]提出了一种混合神经网络模型HNN-EE(Hybrid Neural Network for Entity and Event Extraction),能同时对实体和事件进行抽取,深度挖掘了两者之间的依赖关系。仲伟峰等^[8]利用基于注意力机制的序列标注模型联合抽取句子级事件的触发词和实体,通过多层感知机判断实体在事件中扮演的角色。上述文献都证实了实体识别能够明显提高事件抽取效果。

与英文事件抽取不同的是,中文语料中表达语义的基本单位为词语,但词语间没有显示间隔,且中文语料中存在着大量的多义歧义词,如果分词错误则会使模型对句子语义理解造成明显的误差,从而对事件抽取结果带来负面影响。因此直接对中文语料进行分词预处理的效果大幅度依赖分词工具的准确性,适用性较弱。针对该问题,Zhang等^[9]基于长短期记忆(Long Short-Term Memory, LSTM)网络提出一种用于命名实体识别任务的Lattice(点阵)机制,通过该机制有效地利用丰富的词语边界信息,将词典融合到字符级的中文语料中使其附加词语语义信息。Liu等^[10]对前者的工作进行了优化和拓展,直接在字向量中融入词语向量,并采用4种不同的方法对多词同首和多词同尾的特殊情况进行处理。Xue等^[11]将自注意力机制与Lattice机制进行融合形成新型点阵变换编码器,能够以批处理模型捕获字符与匹配的词语之间的依存关系,使模型处理速度得到明显的提高。针对上述研究现状,本文以开放领域的中文突发事件数据作为研究对象,对仅能获取字粒度语义信息的双向长短期记忆网络-条件随机场(Bidirectional Long Short-Term Memory network-Conditional Random Field, BiLSTM-CRF)事件抽取模型进行词语维度和实体维度的语义特征增强,提出一种基于命名实体识别任务反馈增强的事件抽取改进模型FB-Lattice-BiLSTM-CRF(FeedBack-Lattice-BiLSTM-CRF)。该模型增加了命名实体识别辅助任务并将其输出反馈至输入端,反馈信息与输入融合得到动态实体分词结果;将Lattice机制与双向长短期记忆(Bidirectional Long Short-Term Memory, BiLSTM)网络进行结合,使反馈得到的动态实体分词信息通过该机制输入到模型后增强输入的实体语义特征,句子其他部分通过自组词查询得到相关词语结果;最后通过最大化同方差不确定性的最大高斯似然估计方法对各任务损失进行了平衡处理。实验结果表明,本文提出的改进模型相较于基准模型在事件抽取各项评价指标和模型训练收敛速度上都有明显的提升。

1 数据准备

1.1 数据来源及数据集划分

本文使用的实验数据来源之一是中文突发事件语料库

(Chinese Emergencies Corpus, CEC)^[12],该语料库是由上海大学语义智能实验室所构建。根据国务院颁布的《国家突发公共事件总体应急预案》的分类体系,从互联网上收集了合计332篇地震、火灾、交通事故、恐怖袭击和食物中毒共5类突发事件新闻报道作为生语料,然后再对生语料进行文本预处理、文本分析以及一致性检查等处理,最后将结果保存到语料库中。与ACE2005(Automatic Content Extraction 2005)中文语料库和TimeBank语料库相比,CEC语料库对事件和事件要素的标注更为全面。

由于训练本文提出的神经网络模型需要一定数量的数据,因此需要对原语料库进行扩充。以CEC语料库的数据为参考,本文爬取了国家突发事件预警信息发布网“近期突发事件”和微博平台“中国新闻网”发布的公共突发事件报道,通过数据筛选和数据清洗后,与CEC语料库合并最终得到1847篇实验语料,共计206383字符。

经过打乱数据排列顺序获得特征分布近似均匀的数据后,本文按照8:1:1的比例将数据划分为训练集、验证集和测试集。训练集用于训练模型参数;验证集用于每隔一定的迭代之后测试模型的效果,从而及时保存训练过程中达到更好效果的模型参数;测试集用于评估训练完成的模型效果。

1.2 标签定义

根据数据特征种类及稀疏程度,同时参考CEC语料库的数据标签及标注内容,本文将事件抽取标签重新定义并分为七类:事件触发词(Event Trigger Word)、事件类型(Event Type)、事发时间(Happened Time)、事发地点(Location)、事件参与者(Event Participant)、受影响对象(Affect Object)和受影响对象状态变化(Affect State)。其中,受影响对象状态变化是指该对象自身状态因受到事件发生的影响而产生的改变,如房屋“被摧毁”,车辆“被损坏”,人员“被困”“受伤”“死亡”等。

本文设计的算法中使用命名实体识别作为辅助学习任务,因此需要定义命名实体识别的标签信息。根据数据的特征定义了五类标签,分别是:时间(Time)、地点(Location)、组织(Orgnization)、人(Person)和设备(Equipment)。

1.3 数据标注

在本文实验中,事件抽取任务和命名实体识别任务全部转化为序列标注问题,并选择BIO标注法对数据进行标注。尾字母B表示标签对应的第一个字,I表示标签对应的中间或最后一个字,O表示无关字符。对训练集进行事件抽取和命名实体识别联合标注,对验证集和测试集仅依据事件抽取标签进行了标注。

虽然事件抽取和命名实体识别的标签定义中有重合部分,但在标注方法上存在以下不同:

1)EE标注仅针对描述事件的信息,NER标注针对所有符合实体标签的信息。以时间信息标注为例:“中新网4月12日电据巩义市政府新闻办公室官方微博消息,4月10日22时左右,巩义市发生一起自备火车冲出避难线倾覆事故。”EE标注只关注后半句中的事发时间“4月10日22时左右”,而NER标注将句中的两个时间实体信息全部进行标注。

2)EE标注以事件要素整体为单位,NER标注以实体为单位。以地点信息标注为例,EE标注和NER标注对比如表1所示。

表1 在地点信息上标注事件抽取和命名实体识别

Tab. 1 Labeling event extraction and named entity identification on location information

原文	EE 标注	NER 标注	原文	EE 标注	NER 标注
滨	LOC-B	LOC-B	珠	LOC-I	LOC-I
海	LOC-I	LOC-I	路	LOC-I	LOC-I
四	LOC-I	LOC-I	交	LOC-I	0
路	LOC-I	LOC-I	叉	LOC-I	0
与	LOC-I	0	口	LOC-I	0
明	LOC-I	LOC-B			

2 模型算法

2.1 BiLSTM-CRF

BiLSTM-CRF 网络结构如图1所示:首先将输入文本通过 Embedding 层进行字符编码后输入到 BiLSTM 网络中,再通过正反向传播计算使得各字符融合相应的上下文语义信息;线性层对 BiLSTM 层的输出进行特征提取并将其维度调整到符合条件随机场(Conditional Random Field, CRF)层的输入维度要求;最后通过 CRF 层学习标签序列的潜在规则信息并计算出最优的预测序列进行输出。

BiLSTM-CRF 综合了 BiLSTM 对上下文的语义获取能力和 CRF 对序列预测结果的优化能力,使得其在处理事件抽取、命名实体识别等序列标注任务上展现出较好的效果。

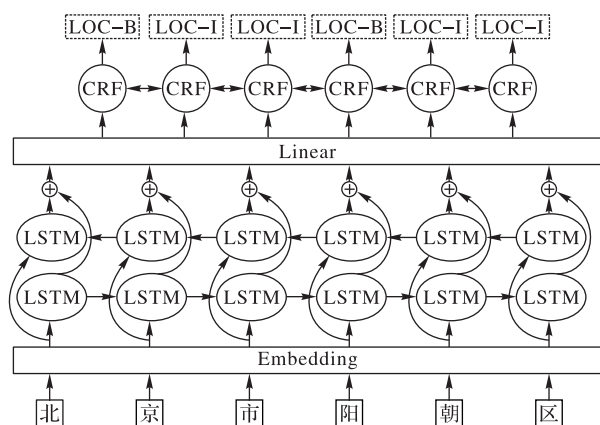


图1 BiLSTM-CRF 网络结构

Fig. 1 BiLSTM-CRF network structure

2.1.1 LSTM

BiLSTM-CRF 的 BiLSTM 网络层主要由 LSTM^[13] 神经元构成。LSTM 是一种特殊的循环神经网络,其单元内部结构如图2所示。

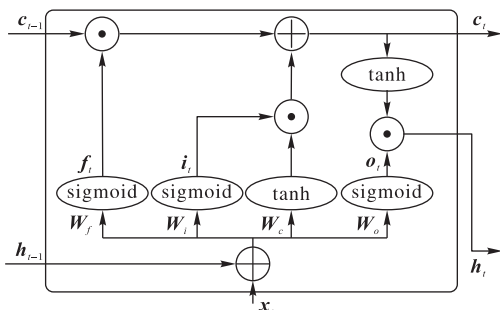


图2 LSTM 单元内部结构

Fig. 2 Internal structure of LSTM unit

LSTM 单元的内部结构为其增加了信息存储和更新机制,而这些机制通过一系列门结构来实现,其中遗忘门 f_t 选择性

丢弃一些传递信息,输入门 i_t 整合该时刻的输入信息,输出门 o_t 为输出提供输入分量信息,从而解决了循环神经网络(Recurrent Neural Network, RNN)的长远依赖问题。LSTM 单元内部信息计算公式如下:

$$f_t = \text{sigmoid}(\mathbf{W}_f \times [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \quad (1)$$

$$i_t = \text{sigmoid}(\mathbf{W}_i \times [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \quad (2)$$

$$\tilde{c}_t = \tanh(\mathbf{W}_c \times [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c) \quad (3)$$

$$o_t = \text{sigmoid}(\mathbf{W}_o \times [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o) \quad (4)$$

$$\mathbf{c}_t = \mathbf{c}_{t-1} \times f_t + i_t \times \tilde{c}_t \quad (5)$$

$$\mathbf{h}_t = \tanh(\mathbf{c}_t) \times o_t \quad (6)$$

其中: $[\mathbf{a}, \mathbf{b}]$ 表示向量 $\mathbf{a}$ 与向量 $\mathbf{b}$ 进行拼接操作; $\mathbf{h}_{t-1}$ 和 $\mathbf{h}_t$ 分别表示神经元上一时刻和该时刻状态的输出信息; $\mathbf{c}_{t-1}$ 和 $\mathbf{c}_t$ 分别表示神经元上一时刻和该时刻状态的细胞状态信息; $\tilde{c}_t$ 表示候选细胞状态, $\mathbf{W}_f, \mathbf{W}_i, \mathbf{W}_c$ 和 $\mathbf{W}_o$ 表示权重,对应的偏置项为 $\mathbf{b}_f, \mathbf{b}_i, \mathbf{b}_c$ 和 $\mathbf{b}_o$ 。

2.1.2 CRF

CRF 是一种特殊的隐马尔可夫模型。在序列标注任务中,模型虽然能够直接预测输出,但每个输出的标签前后之间没有关联性,难以符合标签序列中隐含的规则。CRF 能够学习标签序列中的规则信息,例如输出序列不能以 LOC-I 开头、LOC-I 不能接在 TM-B 后面等,从而大幅度提升了序列标注任务的准确率。

若 CRF 层的给定序列为 $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$, 标签序列为 $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$, 其最大条件似然估计公式如下:

$$L = \sum_{i=1}^n \lg(P(y_i | x_i)) + \frac{\lambda}{2} \|\theta\|^2 \quad (7)$$

其中: P 表示原序列到预测序列对应的概率。定义 $s(\mathbf{x}, \mathbf{y})$ 为输入序列到标签序列的评分函数,则概率 P 计算公式如下:

$$P(y_i | x_i) = \frac{e^{s(x_i, y_i)}}{\sum_{j=1}^n e^{s(x_i, y_j)}} \quad (8)$$

2.2 FB-Lattice-BiLSTM-CRF

基于单任务事件抽取基准模型 BiLSTM-CRF 的缺陷,本文提出了一种基于命名实体识别任务反馈增强的事件抽取改进模型 FB-Lattice-BiLSTM-CRF, 主要结构如图3所示。针对基准模型仅能学习到字粒度语义特征的问题,本文将 Lattice 机制融入 BiLSTM 网络,借用词向量对输入语料的语义增强效果提高了模型对数据特征的学习能力;此外通过增加命名实体识别任务预测输出反馈机制,使模型获得动态实体信息,为 Lattice-LSTM 机制提供了定向的实体分词结果,减少实体部分自组词数量大带来的运算负荷;针对多任务损失不平衡问题,本文使用最大高斯似然估计的方法,使各任务的损失同方差不确定性最大化,从而实现各任务损失达到平衡,同时增加主辅任务控制系数引导命名实体识别任务辅助事件抽取任务的学习。

2.2.1 Lattice-LSTM 机制

中文语料的事件抽取和命名实体识别关注的目标一般是词语或是词语构成的短语,所以与分词具有很强的关联性, He 等^[14] 通过研究证实了将分词结果的词语信息作为输入能够增加语义特征从而有效的提升模型的性能。在序列标注任务中,将分词结果融入输入序列通常有两类方法:一类是使用第三方工具直接对输入语料进行分词预处理,然后通过

Embedding 层进行词粒度嵌入,但这种方法的效果很大程度上取决于分词结果的准确性。例如“南京市长江大桥”应分词为“南京市,长江大桥”,也可能被分词为“南京,市长,江大桥”,后者出现了歧义词,将会对模型训练产生负面影响。另

一类是将遍历的词信息融入字粒度的输入中,代表方法是 Zhang 等^[9] 基于 LSTM 网络提出的点阵长短期记忆网络 (Lattice-Long Short-Term Memory network, Lattice-LSTM) 机制,其主要结构如图 4 所示。

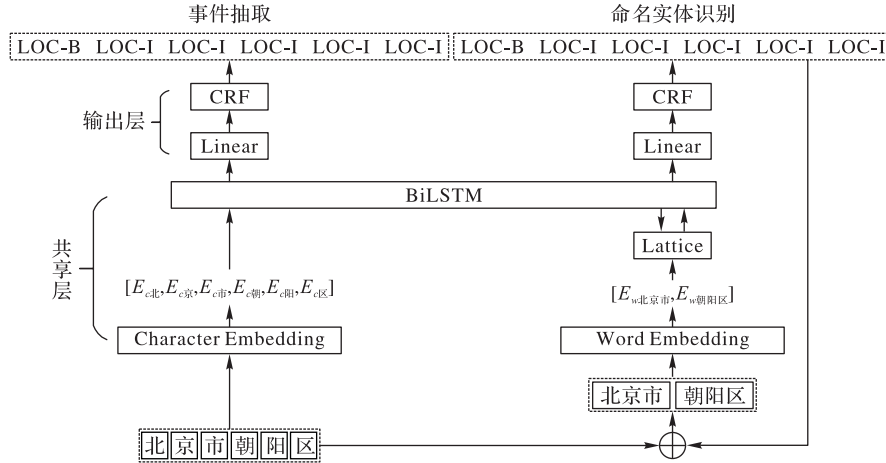


图 3 FB-Lattice-BiLSTM-CRF 结构

Fig. 3 FB-Lattice-BiLSTM-CRF structure

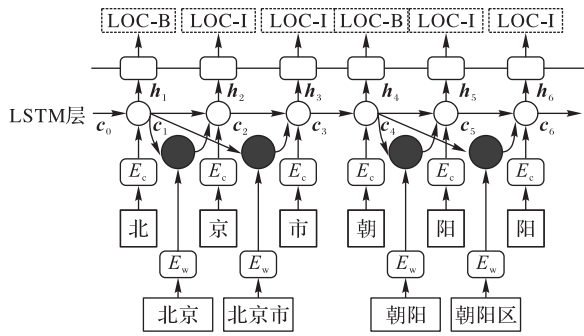


图 4 Lattice-LSTM 结构

Fig. 4 Lattice-LSTM structure

Lattice-LSTM 机制定义了一个新神经元,如图 4 中阴影单元所示,其内部结构与 LSTM 单元相似,唯一不同的是该单元没有输出门,即传递信息中只有细胞状态 c 而没有直接输出信息 h 。此外,字嵌入和词嵌入过程分别使用不同的 Embedding 网络 E_c 和 E_w 进行。

若词的长度为 n ,词首字对应时刻的 LSTM 细胞状态信息为 c_1 ,则新单元将 c_1 作为输入,计算得到新单元输出的细胞状态 c_w ;Lattice-LSTM 机制会把词信息融入词尾字作为输入对应时刻 LSTM 网络的输入细胞状态 c_{n-1} 中,即新单元输出的细胞状态 c_w 和 c_{n-1} 进行拼接形成新的细胞状态 $[c_w, c_{n-1}]$,由于新的细胞状态维度与原维度不同,因此需要定义一个新的门 $\tilde{c}_w^c$ 来进行转换,公式如下:

$$\tilde{c}_w^c = \text{sigmoid}(\mathbf{W}_w^c \cdot [c_w, c_{n-1}] + b_w^c) \quad (9)$$

如果多个自组词的词尾指向相同位置字符,则进行归一化处理求和得到融合后的细胞状态 c_{n-1}' 如下:

$$c_{n-1}' = \sum_j \frac{\exp(\tilde{c}_{w_j}^c)}{\exp(c_{n-1}) + \sum_i \exp(\tilde{c}_{w_i}^c)} \quad (10)$$

针对 Lattice-LSTM 自组词的词量多,语义泛化增强可能引起特征过度增加的问题,本文将通常作为辅助任务的命名实体识别结果反馈到输入端,和原始语料数据进行融合,从而形成实体分词信息;再将分词信息输入到 Lattice-LSTM 机制

中进行定向词信息融入。一方面,分词信息的加入对事件抽取任务具有增强效果;另一方面,定向的实体分词使得命名实体识别任务形成正反馈,能够有效加速模型的收敛过程。与对训练数据进行分词预处理效果不同的是,该模型融合反馈信息的分词结果是动态的,会随着命名实体识别任务效果的提高而不断优化,避免了因分词错误而产生歧义的问题。

对于除实体以外包含动词、介词等的剩余句子部分,进行逐个字符遍历,同时分别查询从当前字符起的前 2、3、4 个连续字符组合是否存在于中文词典中,如果存在则返回自组词结果。由于歧义性大且自组词数量较多的实体部分已经得到确定结果,句子整体的自组词数量会明显减少,从而使 Lattice-LSTM 部分的运算量得到有效的优化。

2.2.2 多任务损失平衡

为了解决多任务损失失衡问题,Cipolla 等^[15] 提出一种通过最大化同方差不确定性的最大高斯似然估计的方法。该方法以回归任务和分类任务为例进行了任务分析、方法定义和相应的公式推导。参考该方法解决多任务损失平衡问题的思路,本文对 CRF 层输出的序列标注任务定义了如下概率分布函数:

$$p(y|f^w(x), \sigma) = \text{CRF}\left(\frac{1}{\sigma^2} f^w(x)\right) \quad (11)$$

其中 $\text{CRF}(x) = \lg\left(\frac{\exp(x_i)}{\sum_i \exp(x_i)}\right)$, 得到似然函数如下:

$$p(y = C|f^w(x), \sigma) = \frac{1}{\sigma^2} f_c^w(x) - \lg \sum_i \exp\left(\frac{1}{\sigma^2} f_{c_i}^w(x)\right) \quad (12)$$

定义新的交叉熵损失 $L(W, \sigma) = p(y = C|f^w(x), \sigma)$, 则两个序列标注任务的总交叉熵损失计算过程如下:

$$\begin{aligned} L(W, \sigma_1, \sigma_2) &= p(y_1 = C_1, y_2 = C_2|f^w(x), \sigma_1, \sigma_2) = \\ &= p(y_1 = C_1|f^w(x), \sigma_1) + p(y_2 = C_2|f^w(x), \sigma_2) = \\ &= \frac{1}{\sigma_1^2} L_1(W) - \lg \sum_i \exp\left(\frac{1}{\sigma_1^2} f_{c_i}^w(x)\right) + \end{aligned}$$

$$\begin{aligned}
& \frac{1}{\sigma_2^2} L_2(\mathbf{W}) - \lg \sum_j \exp \left(\frac{1}{\sigma_2^2} f_{c_j^w}^w(\mathbf{x}) \right) = \\
& \frac{1}{\sigma_1^2} L_1(\mathbf{W}) + \lg \frac{\sum_i \exp \left(\frac{1}{\sigma_1^2} f_{c_i^w}^w(\mathbf{x}) \right)}{\left[\sum_i \exp \left(f_{c_i^w}^w(\mathbf{x}) \right) \right]^{\frac{1}{\sigma_1^2}}} + \\
& \frac{1}{\sigma_2^2} L_2(\mathbf{W}) + \lg \frac{\sum_j \exp \left(\frac{1}{\sigma_2^2} f_{c_j^w}^w(\mathbf{x}) \right)}{\left[\sum_j \exp \left(f_{c_j^w}^w(\mathbf{x}) \right) \right]^{\frac{1}{\sigma_2^2}}} \approx \\
& \frac{1}{\sigma_1^2} L_1(\mathbf{W}) + \lg \sigma_1 + \frac{1}{\sigma_2^2} L_2(\mathbf{W}) + \lg \sigma_2 \quad (13)
\end{aligned}$$

其中: $L_1(\mathbf{W})$ 、 $L_2(\mathbf{W})$ 分别为事件抽取任务和命名实体识别任务的交叉熵损失; σ_1 和 σ_2 是可学习参数, 通过网络训练进行更新。随着 σ_1 和 σ_2 的优化, 各任务的损失将逐渐趋于平衡。

在 FB-Lattice-BiLSTM-CRF 模型中, 命名实体识别只是事件抽取的辅助任务, 为避免两任务在模型训练后期出现过度对抗竞争的问题, 增加超参数主辅任务控制系数 K 作为事件抽取任务关联项的系数, 总损失如下:

$$\begin{aligned}
L(\mathbf{W}, K, \sigma_1, \sigma_2) = & \frac{K}{\sigma_1^2} L_1(\mathbf{W}) + \lg \sigma_1 + \\
& \frac{1}{\sigma_2^2} L_2(\mathbf{W}) + \lg \sigma_2 \quad (14)
\end{aligned}$$

在模型训练过程中, σ_1 和 σ_2 的更新受到学习率的限制, 使得两任务的损失不会迅速达到平衡, 因而即使超参数 K 取相同的值仍可以在模型的迭代过程中持续影响 σ_1 和 σ_2 的更新。若希望 K 能够全程调节主辅任务在网络中的作用, 则要求 K 为一个变量。

3 实验结果与分析

3.1 实验环境与参数设置

本文的实验环境为: 操作系统 Ubuntu 18.04, CUDA 10.01; GPU 1080ti, 显存 11 GB; 编程语言 Python 3.7, 机器学习框架 Pytorch 1.6; 编译器 Pycharm Profession

实验参数如表2所示。

表2 实验参数设置

Tab. 2 Experimental parameter setting

超参数	设置	超参数	设置
optimizer	Adam	Embedding dim	128
Learning Rate	0.01	Hidden dim	200
Batch Size	32		

可学习参数 σ_1 和 σ_2 初始值为 0.1。此外, 经实验测定 K 满足 $\{K|K^2 = d + 4, K > 0\}$ 时, 模型有较好效果, 其中 d 为迭代次数。

3.2 评价指标

本文实验中使用序列标注任务常用的准确率 P (Precision)、召回率 R (Recall) 和 F_1 (F1 值) 作为模型性能的评价指标, 计算公式如下:

$$P = TP / (TP + FP) \times 100\% \quad (15)$$

$$R = TP / (TP + FN) \times 100\% \quad (16)$$

$$F_1 = 2PR / (P + R) \times 100\% \quad (17)$$

其中: TP (True Positive) 表示测试集中被正确识别的标签数

量, FP (False Positive) 表示被错误识别的标签数量, FN (False Negative) 表示测试集中没有被识别的标签数量。

3.3 结果分析

为了验证模型的有效性, 本文设置了三组对比实验, 各对比实验模型分别是:

1) BiLSTM-CRF (单任务): 基于 BiLSTM-CRF 的单任务事件抽取模型, 在本实验中该模型为基准模型;

2) BiLSTM-CRF (多任务): 基于 BiLSTM-CRF 的事件抽取和命名实体识别多任务联合学习模型, 较基准模型增加了对实体语义特征的提取;

3) BiLSTM-CRF (多任务+Lattice 机制): 基于 Lattice-BiLSTM-CRF 的事件抽取和命名实体识别多任务联合学习模型, 该模型中的 Lattice 机制全部使用自组词查询结果, 较基准模型增加对实体语义特征提取的同时, 将词语语义信息融入相应的字向量中。

此外, 由于 FB-Lattice-BiLSTM-CRF 模型存在输出反馈机制, 在训练方式上与对比实验模型不同。该模型单次迭代的训练过程如图5所示。

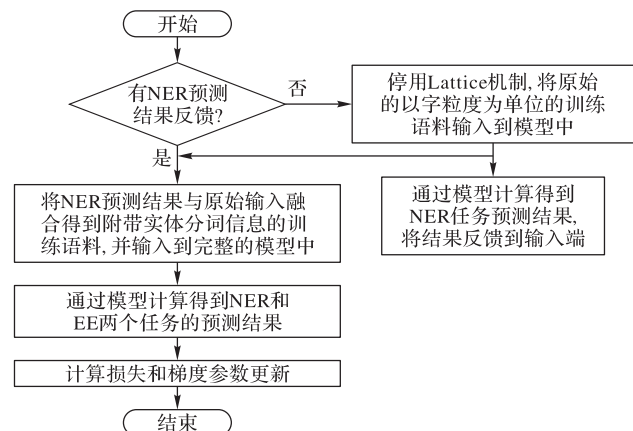


图5 FB-Lattice-BiLSTM-CRF 单次迭代训练过程

Fig. 5 Single iteration training process of FB-Lattice-BiLSTM-CRF

四种模型在相同的超参数设置下进行训练, 训练过程中分别取验证集上有最好结果的模型进行保存, 最后在测试集上进行事件抽取的预测结果如表3所示。可以看出, 实体维度和词语维度的语义增强都对基准模型有明显提升; 同时, 本文提出的 FB-Lattice-BiLSTM-CRF 相较于基准模型, 在准确率、召回率和 F1 值分别提升了 7.63、4.41 和 5.95 个百分点, 验证了针对基准模型缺陷进行的模型结构改进和损失平衡优化是有效的。

表3 四种事件抽取模型在测试集的结果评价 单位: %

Tab. 3 Result evaluation of four event extraction models on test set unit: %

模型	P	R	F_1
BiLSTM-CRF (单任务)	73.62	72.09	72.85
BiLSTM-CRF (多任务)	76.93	73.06	74.95
BiLSTM-CRF (多任务+Lattice 机制)	78.89	73.94	76.33
FB-Lattice-BiLSTM-CRF	81.25	76.50	78.80

FB-Lattice-BiLSTM-CRF 模型由于将输出结果反馈到输入, 实体分词位置信息与结果预测的实体位置信息相同, 在命名实体识别任务上相当于正反馈作用, 这会加速该任务的训练过程, 进而提高整个模型训练的收敛速度。本模型和基准模型在验证集上 F1 值评价变化如图6所示。

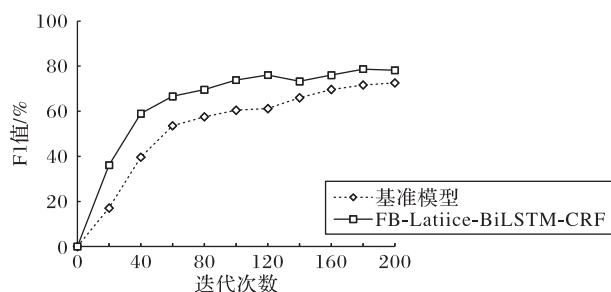


图6 模型训练过程中在验证集上的F1值曲线

Fig. 6 F1 value during model training on verification set

4 结语

针对 BiLSTM-CRF 模型在中文突发事件抽取任务上可学习语义特征维度较低的问题,本文提出了端到端的改进模型 FB-Lattice-BiLSTM-CRF。该模型通过增加命名实体识别任务输出反馈机制和 Lattice 机制,使用最大化同方差不确定性的最大高斯似然估计方法平衡各任务损失,提高了模型对词语和实体语义特征的学习能力。对比实验表明,改进后的模型在准确率、召回率和 F1 值上均有明显提升,同时提高了模型训练中的收敛速度。该改进方法未对模型中间层进行限制,因此也可应用于其他使用 LSTM 或门控循环单元 (Gate Recurrent Unit, GRU) 网络为底层网络的语言模型中,例如增加注意力机制的 BiLSTM-Att-CRF (Bidirectional Long Short-Term Memory network-Attention-Conditional Random Field) 模型,以 BiLSTM 作为 Encoder 的 Seq2Seq 模型,预训练语言模型 ELMo (Embeddings from Language Models) 等。

本文提出的模型由于存在输出反馈机制,使用通用方法进行模型训练时,复杂度有所提高;此外,模型中主辅任务控制系数 K 的测定还缺少针对性的研究,目前只是在小范围值域内进行了局部优化。今后将针对这两个方向做进一步研究探索。

参考文献 (References)

- [1] 杨佳乐,王俊豪,钱卫宁,等. 裁定文书中企业破产事件的自动化抽取[J]. 华东师范大学学报(自然科学版), 2020(4): 88-97. (YANG J L, WANG J H, QIAN W N, et al. Automatic extraction of corporate bankruptcy-related events from ruling documents [J]. Journal of East China Normal University (Natural Science), 2020 (4): 88-97.)
- [2] 邱奇志,周三三,刘长发,等. 基于文体和词表的突发事件信息抽取研究[J]. 中文信息学报, 2018, 32(9): 56-65, 74. (QIU Q Z, ZHOU S S, LIU C F, et al. Emergency information extraction based on style and terminology [J]. Journal of Chinese Information Processing, 2018, 32(9): 56-65, 74.)
- [3] 余晨,毛喆,高嵩. 基于规则的海事自由文本信息抽取方法研究[J]. 交通信息与安全, 2017, 35(2): 40-47. (YU C, MAO Z, GAO S. An approach of extracting information for maritime unstructured text based on rules [J]. Journal of Transport Information and Safety, 2017, 35(2): 40-47.)
- [4] 陈亚东,洪宇,王潇斌,等. 利用框架语义知识优化事件抽取[J]. 中文信息学报, 2017, 31(2): 117-125. (CHEN Y D, HONG Y, WANG X B, et al. Event extraction optimization via frame semantic knowledge [J]. Journal of Chinese Information Processing, 2017, 31(2): 117-125.)

- [5] LI Q, JI H, HONG Y, et al. Constructing information networks using one single model [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2014: 1846-1851.
- [6] YANG B, MITCHELL T M. Joint extraction of events and entities within a document context [C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2016: 289-299.
- [7] 吴文涛,李培峰,朱巧明. 基于混合神经网络的实体和事件联合抽取方法[J]. 中文信息学报, 2019, 33(8): 77-83. (WU W T, LI P F, ZHU Q M. Joint extraction of entities and events by a hybrid neural network [J]. Journal of Chinese Information Processing, 2019, 33(8): 77-83.)
- [8] 仲伟峰,杨航,陈玉博,等. 基于联合标注和全局推理的篇章级事件抽取[J]. 中文信息学报, 2019, 33(9): 88-95, 106. (ZHONG W F, YANG H, CHEN Y B, et al. Document-level event extraction based on joint labeling and global reasoning [J]. Journal of Chinese Information Processing, 2019, 33(9): 88-95, 106.)
- [9] ZHANG Y, YANG J. Chinese NER using lattice LSTM [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2018: 1554-1564.
- [10] LIU W, XU T, XU Q, et al. An encoding strategy based word-character LSTM for Chinese NER [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2019: 2379-2389.
- [11] XUE M, YU B, LIU T, et al. Porous lattice transformer encoder for Chinese NER [C]// Proceedings of the 28th International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020: 3831-3841.
- [12] 刘炜,刘菲京,王东,等. 一种基于事件本体的文本事件要素提取方法[J]. 中文信息学报, 2016, 30(4): 167-175. (LIU W, LIU F J, WANG D, et al. A text event elements extraction method based on event ontology [J]. Journal of Chinese Information Processing, 2016, 30(4): 167-175.)
- [13] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.
- [14] HE J, WANG H. Chinese named entity recognition and word segmentation based on character [C]// Proceedings of the 6th SIGHAN Workshop on Chinese Language Processing. [S. l.]: Asian Federation of Natural Language Processing, 2008: 128-131.
- [15] CIPOLLA R, GAL Y, KENDALL A. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7482-7491.

This work is partially supported by the National Key Research and Development Program of China (2018YFC0807000).

WU Guoliang, born in 1996, M. S. candidate. His research interests include natural language processing, event extraction.

XU Jining, born in 1970, Ph. D., associate professor. Her research interests include automatic control system, field bus.