·开发研究与工程应用·　　文章编号：1000-3428(2019)05-0308-07　　文献标志码：A　　中图分类号：TP391

基于 BiLSTM-CRF 的商情实体识别模型

张应成[1]，杨　洋[2]，蒋　瑞[3,4]，全　兵[5]，张利君[3]，任晓雷[6]

（1. 四川大学 计算机学院，成都 610065；2. 四川省计算机研究院，成都 610041；
3. 成都瑞贝英特信息技术有限公司，成都 610041；4. 四川智仟科技有限公司，成都 610041；
5. 中移（苏州）软件技术有限公司，江苏 苏州 215000；6. 四川黑马数码科技有限公司，四川 泸州 646000）

摘　要： 结合语言模型条件随机场（CRF）和双向长短时记忆（BiLSTM）网络，构建一种 BiLSTM-CRF 模型，以提取商情文本序列中的招标人、招标代理以及招标编号 3 类实体信息。将规范化后的招标文本序列按字进行向量化，利用 BiLSTM 神经网络获取序列化文本的前向、后向文本特征，并通过 CRF 提取出双向本文特征中相应的实体。实验结果表明，与传统机器学习算法 CRF 相比，该模型 3 类实体的精确率、召回率和 F1 值平均提升 15.21%、12.06% 和 13.70%。

关键词： 条件随机场；双向长短时记忆网络；语言模型；命名实体识别；深度学习

开发科学（资源服务）标志码（OSID）：

中文引用格式： 张应成，杨洋，蒋瑞，等. 基于 BiLSTM-CRF 的商情实体识别模型[J]. 计算机工程，2019，45(5)：308-314.

英文引用格式： ZHANG Yingcheng，YANG Yang，JIANG Rui，et al. Commercial intelligence entity recognition model based on BiLSTM-CRF[J]. Computer Engineering，2019，45(5)：308-314.

Commercial Intelligence Entity Recognition Model Based on BiLSTM-CRF

ZHANG Yingcheng[1]，YANG Yang[2]，JIANG Rui[3,4]，QUAN Bing[5]，ZHANG Lijun[3]，REN Xiaolei[6]

（1. College of Computer Science，Sichuan University，Chengdu 610065，China；
2. Sichuan Institute of Computer Sciences，Chengdu 610041，China；
3. Chengdu Ruibeiyingte Information Technology Co.，Ltd.，Chengdu 610041，China；
4. Sichuan Zhiqian Science and Technology Co.，Ltd.，Chengdu 610041，China；
5. China Mobile（Suzhou）Software Technolgy Co.，Ltd.，Suzhou，Jiangsu 215000，China；
6. Sichuan Heima Digital Technology Co.，Ltd.，Luzhou，Sichuan 646000，China）

【Abstract】 A BiLSTM-CRF model is constructed by combining the Conditional Random Field（CRF）model of Bidirectional Long Short-Term Memory（BiLSTM）network to extract three kinds of entity information，tenderer，bidding agent and bidding number，in a commercial text sequence. The normalized bidding text sequence is vectorized by word. The forward and backward text features of the serialized text are obtained by BiLSTM neural network，and the corresponding entities in the two-way text features are extracted by CRF. Experimental results show that compared with the traditional machine learning algorithm CRF，the precision，recall rate and F1 value of the three types of entities in the proposed model are improved by 15.21%，12.06% and 13.70% in average，respectively.

【Key words】 Conditional Random Field（CRF）；Bidirectional Long Short-Term Memory（BiLSTM）network；language model；Named Entity Recognition（NER）；deep learning

DOI： 10.19678/j.issn.1000-3428.0052810

基金项目： 四川省科技计划项目（18PTDJ0085，2019YFH0075，2018GZDZX0030）；泸州市科技计划项目（2017CDLZ-G25）。

作者简介： 张应成（1994—），男，硕士研究生，主研方向为自然语言处理、人工智能；杨　洋、蒋　瑞、全　兵、张利君，工程师、硕士；任晓雷，工程师。

收稿日期： 2018-10-08　　**修回日期：** 2018-11-11　　**E-mail：** yangyang@tccxfw.com

0　概述

标书是由发标单位编制,向投标者提供工程技术、质量和工期要求的文件。标书中隐藏了很多重要信息,具有较大的商业价值。例如,招标方和项目代理等信息,可以使企业迅速查找到感兴趣的项目并进行投标,进而带来经济效益。

在标书文档中,招标方的名称识别不仅具有领域特殊性、复杂性,还面临着名称用词随意的问题。与传统识别方法相比,深度神经网络以数据为驱动,可以自动地从数据中提取有用的特征,将其应用于非结构化、模式未知多变的数据分析具有显著优势[1]。在自然语言处理领域,深度神经网络方法已经在词性标注[2]、文本分类[3]、命名实体识别[4]以及情感分析[5]等任务中取得良好效果,并且在特征与模型层面无需依赖大量先验知识,可以自动地发现数据中的特征表达,从而对标书中的相关信息进行提取。本文提出一种基于深度神经网络的识别方法,以对标书中招标方的名称、招标编号以及招标代理等信息进行识别与提取。

1　相关工作

1.1　传统识别方法

传统信息提取方法利用先验知识,人工设计出识别模型,然后对模型进行定性、定量分析以及优化[1],在此基础上识别命名实体。这种识别方法主要包括 2 类:基于规则的方法和基于统计机器学习的方法。

基于规则的方法通过制定好的规则模板提取相应的信息,其需要大量的先验知识以及获知各实体出现的规律,这将大幅提升任务完成的难度。此外,该类方法还具有时间效率低、可移植性弱等缺点[5],其在处理结构化单一的数据集时有效,但随着大数据时代的到来,非结构化数据占有很大比例,基于规则的方法很难获取足够的先验知识以建立规则模板[1]。

基于统计机器学习的方法融合语言模型与统计机器学习算法,主要包括最大熵(Maximum Entropy, ME)模型[6]、隐马尔科夫模型(Hidden Markov Model, HMM)[7]、支持向量机(Support Vector Machine, SVM)[8]以及条件随机场(Conditional Random Field, CRF)[9]。但是,上述方法的特征提取仍然需要人工参与,并且容易丢失文本本身的情感信息,在模型训练时需要大量的人工标注样本,且效果也并不理想。

1.2　神经网络方法

近年来,基于深度神经网络的方法在自然语言处理领域取得了较好效果,该类方法主要包括机器翻译[10]、情感分析[5]、短文本分类[11]以及对话系统[12]等。对于一个序列模型,为实现可变长度的输入以及发掘序列前后的长期依赖关系,回复式神经网络(Recurrent Neural Network, RNN)[13]应运而生。RNN 的多种变体在处理时间序列数据时可以很好地获取并存储记忆[14-15],其结合词向量技术[16],可以避免人工提取特征而直接对原始数据进行处理[17-19]。与传统 n-gram 模型相比,RNN 模型可以保持非限定长度的上下文信息,并且具有位置相关性。

对于一个标准的 RNN,设长度为 T 的输入序列为 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T)$,从时间 $t=1$ 到 $t=T$ 迭代计算隐层神经元 $\boldsymbol{H}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_T)$ 和输出序列 $\boldsymbol{Y}=(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_T)$。元素 $\boldsymbol{h}_t$ 和 $\boldsymbol{y}_t$ 的计算公式如下:

$$\begin{cases}\boldsymbol{h}_t=F(\boldsymbol{W}_{xh}\boldsymbol{x}_t+\boldsymbol{W}_{hh}\boldsymbol{h}_{t-1}+\boldsymbol{b}_h)\\ \boldsymbol{y}_t=\boldsymbol{W}_{hy}\boldsymbol{h}_t+\boldsymbol{b}_y\end{cases}\tag{1}$$

其中,$\boldsymbol{W}$ 是权重矩阵,$\boldsymbol{W}_{xh}$ 是输入层 x 到隐藏层 h 的权重矩阵,$\boldsymbol{W}_{hh}$ 是隐藏层之间的权重矩阵,$\boldsymbol{W}_{hy}$ 是隐藏层 h 到输出层 y 的权重矩阵,$\boldsymbol{b}$ 是各层的偏置向量,$\boldsymbol{b}_h$ 是隐藏层 h 的偏置,$\boldsymbol{b}_y$ 是输出层 y 的偏置,F 是隐藏层的激活函数,激活函数可以是 Sigmoid 函数、tanh 函数等。

2　相关技术

2.1　长短时记忆模型

RNN 模型在语言模型[20]和语音识别[21]中取得了较好效果,但同时也显现出不足。传统的 RNN 模型在处理长序列数据时可能会发生梯度消失或梯度爆炸的现象,对此,研究者建立了长短时记忆(Long-Short Term Memory, LSTM)模型。与 RNN 模型不同,LSTM 模型在隐藏层加入了特别设计的记忆单元,这在一定程度上能够降低梯度消失或梯度爆炸发生的概率[22],从而发现数据在长时间内的相互依赖关系。图 1 所示为 LSTM 模型的一个存储单元。

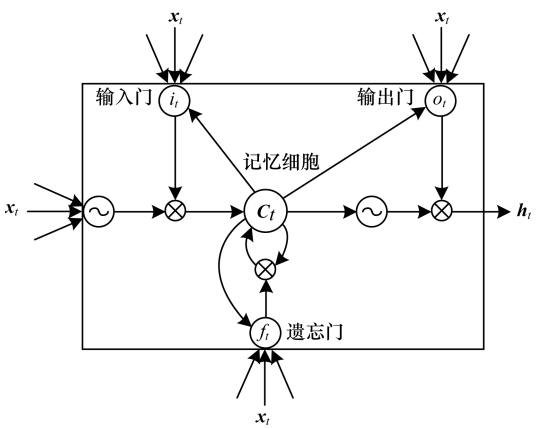

图 1　LSTM 模型存储单元

LSTM 模型的计算过程表示如下：

$$\begin{cases} \boldsymbol{i}_t = \sigma(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{ci}\boldsymbol{c}_{t-1} + \boldsymbol{b}_i) \\ \boldsymbol{f}_t = \sigma(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{cf}\boldsymbol{c}_{t-1} + \boldsymbol{b}_f) \\ \boldsymbol{o}_t = \sigma(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{co}\boldsymbol{c}_{t-1} + \boldsymbol{b}_o) \\ \boldsymbol{C}_t = \boldsymbol{f}_t\boldsymbol{C}_{t-1} + \boldsymbol{i}_t\tanh(\boldsymbol{W}_{xc}\boldsymbol{x}_t + \boldsymbol{W}_{hc}\boldsymbol{h}_{t-1} + \boldsymbol{b}_C) \\ \boldsymbol{h}_t = \boldsymbol{o}_t\tanh(\boldsymbol{C}_t) \end{cases} \quad (2)$$

其中，σ 是激励函数 Sigmoid 函数，$\boldsymbol{i}_t$、$\boldsymbol{f}_t$、$\boldsymbol{o}_t$、$\boldsymbol{C}_t$、$\boldsymbol{h}_t$ 分别是 t 时刻输入门、遗忘门、输出门、记忆细胞、隐藏门的激活向量，$\boldsymbol{W}$ 为权重矩阵，$\boldsymbol{b}$ 为偏置[23]。在一个文本序列中，对于第 t 个输入，LSTM 模型的输入为 $\boldsymbol{x}_t$、$\boldsymbol{C}_{t-1}$、$\boldsymbol{h}_{t-1}$，其中，$\boldsymbol{x}_t$ 为文本序列的第 t 个输入，$\boldsymbol{C}_{t-1}$和 $\boldsymbol{h}_{t-1}$分别为 LSTM 模型 $t-1$ 时刻的记忆单元和隐藏层的激活向量，相应的输出 $\boldsymbol{C}_t$ 和 $\boldsymbol{h}_t$ 可以由式(2)进行计算。

2.2 条件随机场模型

CRF 模型是在给定一组输入随机变量条件下另一组输出随机变量的条件概率分布模型，其特点是假设输出随机变量能构成马尔科夫随机场，其中，线性链上的特殊条件随机场通常用以解决标注问题，如图 2 所示。在条件概率模型 $P(Y|X)$ 中，Y 是输出变量，用来标记序列，X 是输入变量，用来表示需要标注的观察序列。在学习时，利用训练数据通过极大似然估计得到条件概率模型$\hat{P}(Y|X)$；在预测时，对于给定的输入序列 x，求出条件概率$\hat{P}(Y|X)$最大的输出序列$\hat{y}$[24]。

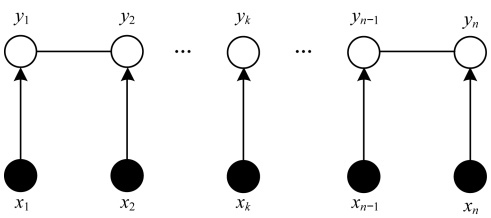

图 2 线性链条件随机场

目前，在标注过程中，利用相邻标注信息预测当前标签的方法分为 2 类：

1)以最大熵分类[25]和最大熵马尔科夫模型[26]为代表。该类方法先单独预测每个输入的标签，再对这些结果统一解码寻找概率最大的序列标注情况。

2)以条件随机场[27]为代表。与第 1 类方法不同，这类方法在处理标注问题时更侧重于对句子整体进行考虑。

在通常情况下，第 2 类方法在标注过程中可以取得更好的效果。

3 基于语言模型的双向 LSTM

为对标书中的招标人、招标编号、招标代理等信息进行有效识别，本文建立一种基于语言模型的双向 LSTM，即 BiLSTM-LM 模型。招标信息的识别是一个命名实体识别问题，给定一个文本序列作为输入，模型需要输出其中的实体。结合词向量技术，LSTM 在处理文本序列问题时有较大优势，其可以获取长时间、长距离的信息依赖关系。基于 LSTM 的 BiLSTM 能够获取更加全面的上下文信息，并且更容易学习到上下文之间的依赖关系。在 BiLSTM 模型中，本文引入语言模型，以期在促进模型获取特定领域语言结构关系的同时增强其泛化能力。

3.1 BiLSTM-CRF 模型

标准 LSTM 按文本序列接收输入，其只能处理前文信息而忽略了下文信息。双向 LSTM 对每一个训练序列分别作用一个向前和向后的 LSTM 网络，且这 2 个 LSTM 网络连接着同一个输出层。这种网络结构可以给输出层提供每一个序列点完整的上下文信息。图 3 所示为 BiLSTM 模型在时间上的展开示意图。

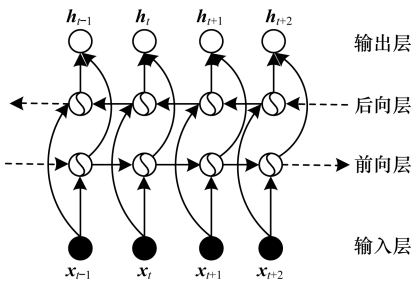

图 3 BiLSTM 网络结构

前向 LSTM 网络依次接受第 1 个时刻～第 t 个时刻的输入 $\boldsymbol{x}_1 \sim \boldsymbol{x}_t$，并依次计算前向隐藏状态$\overrightarrow{\boldsymbol{h}_1}$，$\overrightarrow{\boldsymbol{h}_2}, \cdots, \overrightarrow{\boldsymbol{h}_t}$。反向 LSTM 网络同样接受第 t 个时刻～第 1 个时刻的输入 $\boldsymbol{x}_t \sim \boldsymbol{x}_1$，并相应地计算出反向隐藏状态$\overleftarrow{\boldsymbol{h}_t}, \overleftarrow{\boldsymbol{h}_{t-1}}, \cdots, \overleftarrow{\boldsymbol{h}_1}$[24]。此时，即得到每个时刻前向和后向的双向特征，对 2 个方向上的特征进行拼接，得到一个双向表达：

$$\boldsymbol{h}_t = \begin{bmatrix} \overrightarrow{\boldsymbol{h}_t} \\ \overleftarrow{\boldsymbol{h}_t} \end{bmatrix} \quad (3)$$

向量$\boldsymbol{h}_t$就包含上下文信息，相比 LSTM 模型，双向表达的$\boldsymbol{h}_t$更关注当前词的周围信息[24]。

在标书文本中，关于招标人的描述通常有明显的提示，如“招标人:xxxx 有限公司”“招标方:xxxx 有限公司”。因此，在判断一个文本序列是否包含招标人

时,起始词能够起到关键作用,可以捕捉前后文本中的强依赖关系。BiLSTM 模型能够同时包含较长的前后文关系。例如,对于文本序列“组建了一家有限公司”,CNN 倾向于获取静态信息,当其接收到含有“有限公司”的特征信息时,会误以为该序列是一个公司名称。而 LSTM 更善于在时序序列中获取长时间、长范围的信息,该序列的开头属于干扰词,LSTM 获取的头尾依赖关系较弱,因此,LSTM 会判断该序列不是一个公司名。但是,对于如“xxxx 有限公司中标”或“合同授予 xxxx 有限公司”2 种文本序列,文本段中都含有公司名,但是序列前后存在干扰词,如果只采用 CNN 或单向 LSTM,可能无法排除这些干扰。CNN 倾向于捕捉关键词从而导致误判,单向 LSTM 对后面的干扰不够敏感也将导致误判。双向 LSTM 不仅能捕获时序的动态信息,而且能够利用当前词的前后文信息,最终获取较强的依赖关系。

将 BiLSTM 与 CRF 进行结合,这样既可以利用 BiLSTM 提取文本序列中的上下文信息,也可以通过 CRF 在整句层面上的标注信息来提高标注精确率。BiLSTM-CRF 模型结构如图 4 所示。

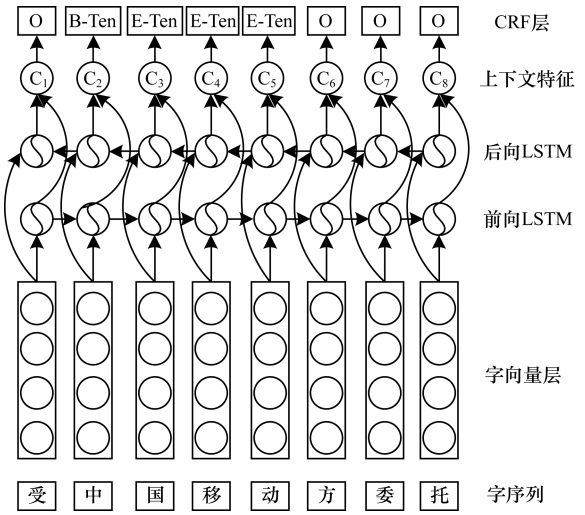

图 4　BiLSTM-CRF 模型结构

3.2　BiLSTM-CRF 网络总体结构

BiLSTM-CRF 模型总体结构主要由 3 个部分组成:词向量层,BiLSTM 网络层,神经网络语言模型 CRF 层。模型的输入为序列文本,按照每个字符进行输入,输出为每个字符的标签,代表是否为所需实体的一部分。输入序列中每个字符通过字向量表达依次输入到 BiLSTM 中,通过 BiLSTM 网络构建包含上下文信息的文本序列双向表达。在得到 BiLSTM 神经网络的双向表达后,对其进行合并,将合并后的表达进行一层隐射后输入到语言模型 CRF 中,通过 CRF 计算出序列文本中每个字符的标签,将其和标准标签对比得出输入序列的对数似然,然后将其定义为整体模型的损失。为提高模型的泛化能力并防止过拟合现象,本文在模型的神经网络部分加入了 dropout 层。

3.3　语言模型

在自然语言处理领域,语言模型广泛应用于词性标注、句法分析、信息检索以及机器翻译中,并发挥着重要作用。语言模型计算一个句子或者一个序列产生的概率,从而判断某一表达合法、通顺、有含义的概率。因此,本文在 BiLSTM-CRF 模型中引入语言模型。CRF 可以很好地将 BiLSTM 神经网络提取到的上下文信息转化为每个字符相应的标签,由这些标签组成实体。在模型训练中,使用时序反向传播算法 BPTT 计算参数梯度,通过 Adam 进行参数更新并最小化损失。

3.4　BiLSTM-CRF 模型在标书文档提取中的应用

招标书是招标单位或国家采购机构对一个招标事件的对外公示,一般可以通过互联网进行查阅。一份招标书背后蕴藏着巨大的商业价值,可以为企业带来经济效益。然而,随着经济水平的发展,近年来在互联网中充斥着大量的招标文件,这给企业的项目分析造成了很大困难。因此,如何自动地对招标文件中的相关信息进行提取,具有较大的现实意义。

招标文件中的重要信息包括招标人、招标编号、招标代理等,本文将 BiLSTM-CRF 模型应用于招标文件的信息提取,流程如图 5 所示。首先,从互联网中爬取招标文件,对每一份文本进行规范化处理并去除异常字符;然后,通过标点符号和段落控制符对整体文本进行切割,形成很多小的文本段,通过 BiLSTM-CRF 模型计算出每个小文本段对应的标签序列;最终,根据标签序列结合原文本找出最终的实体。利用 BiLSTM-CRF 模型进行招标文件信息提取的算法伪代码如算法 1 所示。

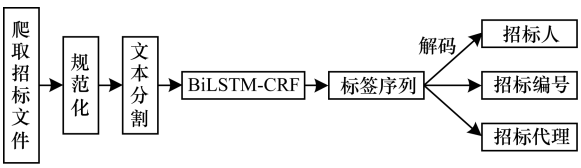

图 5　招标文件信息提取流程

算法 1　招标文件信息提取算法

输入　每一份招标文件的内容 *source*

输出　招标文件中的招标人集合 *tender*、招标编号集合 *number*、招标代理集合 *agent*

1. tender = { },number = { },agent = { };/ * 对集合进行初始化 * /

2. source = normalize(source)/ * 对招标文件内容进行规范化处理,去除异常字符 * /

3. clips = token(source)/ * 对处理后的招标文件内容按标点符号、段落进行切割,形成文本片段并存入列表 clips 中 * /

4. for clip in clips do / * 对 clips 中的每个小文本段 * /

5. temp_tender, temp_number, temp_agent = get_entity (clip)/ * 把每个 clips 输入到 BiLSTM-CRF 模型中,得到招标人、招标编号和招标代理 3 个实体 * /

6. tender←temp_tender; number←temp_number; agent←temp_agent;/ * 将步骤 5 中得到的招标人、招标编号、招标代理放入集合 tender,number,agent 中 * /

7. end for

8. return tender,number,agent

4 实验结果与分析

为验证本文算法对招标文件信息的识别提取效果,本节通过互联网爬取部分招标文件数据进行实验与对比。

4.1 数据集

本次实验从中国政府采购网、招投标采购网等招投标平台上爬取了总计 50 000 条招标文件,对这些招标文件人工查找出其中的招标人、招标编号、招标代理信息,去除无实际内容和 3 类信息为空的数据,最终得到约 45 000 条数据。其中,包含实体招标代理、招标人、招标编号的比例分别约为 43% 、71% 、63% 。对于这些数据,本文按照 7∶3 的比例随机抽取形成训练集和测试集。

4.2 数据预处理

数据预处理是整个模型训练的前提和关键步骤。在实际中,公司名所含词语为专用词或不常用词,例如,词语“苹果”代表一种水果,而在公司名“美国苹果公司”中,“苹果”则完全不具备水果的含义。再如,公司名“诺基亚”是一个音译词,在此之前的汉语中“诺基亚”不具备任何含义。考虑到以上问题,本文以字为边界对句子进行切割。在实验的数据预处理中,首先对训练集、测试集中的字进行统计,形成一个字的集合,对集合中的每个字在[-0.25,0.25]区间内随机初始化,最终得到一个 300 维的字向量。

4.3 模型搭建与参数设置

Tensorflow 是由谷歌人工智能团队开发的深度学习框架,被广泛应用于实现各类机器学习算法。本文采用 Tensorflow 进行模型搭建。实验参数设置:dropout 值为 0.5,训练集的 batch_size 为 32,测试集的 batch_size 为 64,BiLSTM 中前后方向隐藏状态的维度为 100,训练学习率为 0.001。使用梯度裁剪技术来防止 BiLSTM 梯度爆炸。在每一轮训练前,将所有训练数据打乱,采用 Adam 优化器最小化模型损失,当模型训练次数超过 50 或测试集上的 F1 值连续下降 10 次以后结束训练,并保存测试集上 F1 值最高的模型参数。

4.4 性能分析

本次实验对 CRF、CNN、BiLSTM、CNN-CRF、LSTM-CRF 以及 BiLSTM-CRF 6 种方法进行对比,采用召回率、精确率以及 F1 值作为模型评价指标,实验结果如表 1 ~ 表 3 所示。

表 1 各方法的招标人识别结果对比 %

方法	精确率	召回率	F1 值
CRF	64.52	71.83	67.98
CNN	70.13	71.68	70.90
BiLSTM	67.84	79.45	73.19
CNN-CRF	77.83	78.71	78.27
LSTM-CRF	76.73	83.10	79.79
BiLSTM-CRF	82.16	86.32	84.19

表 2 各方法的招标编号识别结果对比 %

方法	精确率	召回率	F1 值
CRF	69.52	70.87	70.19
CNN	74.87	71.34	73.06
BiLSTM	73.46	76.36	74.88
CNN-CRF	75.46	79.04	77.21
LSTM-CRF	81.06	82.47	81.76
BiLSTM-CRF	85.74	86.58	86.16

表 3 各方法的招标代理识别结果对比 %

方法	精确率	召回率	F1 值
CRF	77.13	80.84	78.94
CNN	80.51	81.65	81.08
BiLSTM	73.68	85.12	78.99
CNN-CRF	85.73	83.07	84.38
LSTM-CRF	87.85	82.05	84.85
BiLSTM-CRF	88.91	86.83	87.86

由表 1 ~ 表 3 可以看出:

1)基于 RNN 的方法(如 BiLSTM)整体优于基于 CNN 的方法,原因是 CNN 能有效提取静态特征,但对于动态的序列问题,基于 RNN 的模型则表现得更好。BiLSTM 是 LSTM 模型的改进,从结果上看,BiLSTM 方法优于 LSTM 方法,原因是 BiLSTM 可以从序列的前向和后向获取特征,这样能够更全面地得到序列的知识信息。

2)引入语言模型 CRF 可以给各模型带来不同程度的效果提升。以 CNN 模型为例,加入 CRF 后模型效果得到了显著提升,原因是语言模型的引入

弥补了 CNN 模型只关注局部信息而忽视上下文关系的缺陷。

3)将单一 CRF 和深度神经网络结合 CRF 进行对比可以看出,深度神经网络的加入为模型提供了更好的序列特征,进而提升了模型的效果。此外,分析语言模型 CRF 对各种深度神经网络模型在训练上的作用时发现,CRF 可以使模型更快地收敛。对比分析前 20 轮的参数更新情况,在训练初期,BiLSTM-CRF 模型能够更快地达到一个较高的水平,并且有一个持续的提升,而 BiLSTM 模型在初期提升较为缓慢,经过多次更新后才慢慢达到一个较高的水平。

本文 BiLSTM-CRF 模型在招标人实体、招标代理实体、招标编号实体上的精确率分别为 82.16% 、88.91% 、85.74% ,召回率分别为 86.32% 、86.83% 、86.58% ,说明该方法可以取得较好的实体抽取效果。如果采用多层 BiLSTM-CRF 方法,虽然可以取得更好的结果,但是由于其参数规模较大,在模型训练和实体提取时都会产生很大的时间开销。

4.5　模型实际应用评价

本文利用 BiLSTM-CRF 模型对真实的 500 份招标文件进行信息提取并与人工标注的结果进行对比。其中,3 类实体完全正确的有 469 例,对不正确的部分进行排查,发现有多个招标人导致提取结果漏项的有 7 例,招标人前文用全称、后文用简称导致的结果多项有 5 例,招标人前文用全称、后文再具体到下一级行政单位的情况导致的结果多项有 3 例,提取错误的有 4 例,其余为招标文件无招标人。招标代理上的错误主要体现在未明确指明代理公司,在文件最后以落款签名的形式给出,这种情况有 14 例,其余为标书中不存在招标代理的情况。招标编号上的错误主要体现在对于编号的表述过于宽泛,例如“文件编号”“审批编号”“项目编码”等,有些表述由于缺少足够样本导致模型无法正确识别,这种情况有 23 例,其余为标书中无招标编号的情况,对此,下一步考虑加入规则的方法对模型进行优化。

5　结束语

本文将深度神经网络方法应用到商情实体识别中,构建一种 BiLSTM-CRF 模型,以对招标文件中的招标人、招标代理以及招标编号信息进行识别。将中文文本按字进行向量化,利用 BiLSTM 网络获取前文和后文的双向语义特征。相较于 CNN 网络,BiLSTM 可以更全面地获取序列化文本的整体特征,相较于 LSTM 网络,BiLSTM 可以同时获取前文和后文的双向特征。实验结果表明,与 CNN、RNN 相比,BiLSTM-CRF 模型能够取得较好的识别效果。多层 BiLSTM 模型效果优于单层结构,但其训练和预测的时间消耗较大,在保证高准确率的前提下提升多层 BiLSTM 模型的效率将是下一步的研究方向。

参考文献

[1] ZHANG Lei, ZHANG Yi. Big data analysis by infinite deep neural networks[J]. Journal of Computer Research and Development, 2016, 53(1): 68-79.

[2] BOSCO A, LAGANÀ D, MUSMANNO R, et al. Modeling and solving the mixed capacitated general routing problem[J]. Optimization Letters, 2013, 7(7): 1451-1469.

[3] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[4] CHIU J P C, NICHOLS E. Namedentity recognition with bidirectional LSTM-CNNs [EB/OL]. [2018-09-30]. https://arxiv.org/pdf/1511.08308.pdf.

[5] ZHANG Suxiang, WANG Xiaojie. Automatic recognition of Chinese organization name based on conditional random fields [C]//Proceedings of International Conference on Natural Language Processing and Knowledge Engineering. Washington D. C., USA: IEEE Press, 2007: 229-233.

[6] BORTHWICK A E. A maximum entropy approach to named entity recognition[D]. New York, USA: New York University, 1999.

[7] BIKEL D M, MILLER S, SCHWARTZ R, et al. Nymble: a high-performance learning name-finder[C]//Proceedings of the 15th Conference on Applied Natural Language Processing. Washington D. C., USA: IEEE Press, 1997: 194-201.

[8] ASAHARA M, MATSUMOTO Y. Japanese named entity extraction with redundant morphological analysis[C]// Proceedings of NAACL'03. Stroudsburg, USA: Association for Computational Linguistics, 2003: 8-15.

[9] MCCALLUM A, LI Wei. Early results for named entity recognition with conditional random fields, feature induction and Web-enhanced lexicons[C]//Proceedings of CONLL'03. Stroudsburg, USA: Association for Computational Linguistics, 2003: 188-191.

[10] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[EB/OL]. [2018-09-30]. http://anthology.aclweb.org/D/D14/D14-1179.pdf.

[11] SANTOS C N D, GATTIT M. Deep convolutional neural networks for sentiment analysis of short texts[EB/OL]. [2018-09-30]. http://www.aclweb.org/anthology/C14-1008.

[12] LI Jiwei, GALLEY M, BROCKETT C, et al. A diversity-promoting objective function for neural conversation models[EB/OL]. [2018-09-25]. https://arxiv.org/pdf/1510.03055.pdf.

[13] KARJALA T W,HIMMELBLAU D M,MIIKKULAINEN R. Data rectification using recurrent (Elman) neural networks [C]//Proceedings of International Joint Confe-rence on Neural Networks. Washington D.C.,USA:IEEE Press,1992: 901-906.

[14] GRAVES A. Long short-term memory[M]//GRAVES A. Supervised sequence labelling with recurrent neural networks. Berlin,Germany:Springer,2012:1735-1780.

[15] ZHOU Guobing, WU Jianxin, ZHANG Chenlin, et al. Minimal gated unit for recurrent neural networks[J]. International Journal of Automation and Computing, 2016,13(3):226-234.

[16] MIKOLOV T, CHEN Kai, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. [2018-09-10]. http://export.arxiv.org/pdf/1301.3781.

[17] BENGIO Y,SCHWENK H,SENÉCAL J S,et al. Neural probabilistic language models[J]. Journal of Machine Learning Research,2001,3(6):1137-1155.

[18] MIKOLOV T,KARAFIÁT M,BURGET L,et al. Recurrent neural network based language model[EB/OL]. [2018-09-25]. http://www.fit.vutbr.cz/research/groups/speech/publi/2010/mikolov_interspeech2010_IS100722.pdf.

[19] MIKOLOV T, ZWEIG G. Context dependent recurrent neural network language model[C]//Proceedings of 2012 IEEE Spoken Language Technology Workshop. Washington D.C.,USA:IEEE Press,2012:234-239.

[20] MIKOLOV T,DEORAS A,POVEY D,et al. Strategies for training large scale neural network language models[C]//Proceedings of 2011 IEEE Workshop on Automatic Speech Recognition and Understanding. Washington D.C.,USA:IEEE Press,2011:196-201.

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

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

[23] GRAVES A,JAITLY N,MOHAMED A R. Hybrid speech recognition with deep bidirectional LSTM [C]// Proceedings of 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. Washington D.C.,USA: IEEE Press,2013:273-278.

[24] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. [2018-09-15]. https://arxiv.org/pdf/1409.0473.pdf.

[25] RATNAPARKHI A. A maximum entropy model for part-of-speech tagging[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Washington D.C.,USA:IEEE Press,1996:133-142.

[26] MCCALLUM A, FREITAG D, PEREIRA F C N. Maximum entropy Markov models for information extraction and segmentation[C]//Proceedings of the 17th International Conference on Machine Learning. Washington D.C.,USA:IEEE Press,2000:591-598.

[27] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning. [S.l.]: Morgan Kaufmann Publishers Inc., 2001: 282-289.

编辑 吴云芳

(上接第 307 页)

[11] HOANG T S,DGHAYM D,SNOOK C,et al. A composition mechanism for refinement-based methods[C]//Proceedings of International Conference on Engineering of Complex Computer Systems. Washington D.C., USA: IEEE Press, 2017:100-109.

[12] ABRIAL J R, SU Wen. Transforming guarded events into pre-conditioned operations [EB/OL]. [2018-02-28]. http://wiki.event-b.org/images/JR2_A_sld_Proc_in_EVB_V5.pdf.

[13] 吴劲,陈志慧. 基于 Event-B 的形式化建模关键技术研究[J]. 电子科技大学学报,2014,43(3):405-408.

[14] LEUSCHEL M, BUTLER M. ProB: an automated analysis tool-set for the B method[J]. International Journal on Software Tools for Technology Transfer, 2008,10(2):185-203.

[15] HOFNER P, KHEDRI R, MOLLER B. An algebra of product families[J]. Software and Systems Modeling, 2011,10(2):161-182.

[16] ABRIAL J R. Event model decomposition[EB/OL]. [2018-02-28]. https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/69255/1/eth-4958-01.pdf.

[17] MEDTRONIC Inc. The minimed paradigm real-time insulin pump and continuous glucose monitoring system insulin pump user guide[M]. Minneapolis, USA: Medtronic MiniMed Inc.,2008.

[18] BUTLER M. Decomposition structures for Event-B[J]. Lecture Notes in Computer Science,2009,5423:20-38.

编辑 刘盛龄