[文章编号] 1003－4684(2019)04-0068-05

基于深度学习的法院命名实体识别模型

龚启文，程 玉，陈建峡，李 超，张 帝，龙逸舒
（湖北工业大学计算机学院，湖北 武汉 430068）

[摘 要] 命名实体识别作为信息抽取、问答系统、句法分析、机器翻译等应用领域的重要基础工具，在法院判决书信息抽取系统中也得到了广泛应用。然而，已有的技术模型在文本中存在大量专有名词或术语时，实体识别的提取效果会变得很差。双向循环神经网络–条件随机场判别模型可对现有的法院判决书条件随机场模型进行优化，实现自动化特征的选取过程，准确率比已有的条件随机场模型更高。

[关键词] 命名实体识别；深度学习；条件随机场模型；双向循环神经网络
[中图分类号] TP391 [文献标识码] A

随着司法信息化建设的进展，法院案件判决领域已经逐步实现了信息化，但其数据库中仍然存在着大量的非结构化文本信息。在信息量日益庞大的客观环境下，案件裁决人员不得不耗费大量的精力在相关信息的查找和收集上。如果能对这些非结构化信息进行分析和提取，将各类案件中的信息点提取出来并进行智能化管理，相信对相关从业人员，以及日后的进一步建设都会有重大帮助。

命名实体识别（Named Entity Recognition，NER）这个概念最早是在 1995 年，于 MUC-6（Message Understanding Conference）会议中被提出[1]。命名实体识别的主要任务是识别和分类文本中的专有名词和有意义短语。常见的命名识别包括实体（组织名称、人名、地名）、时间表达式（日期、时间）、数字表达式（货币值、百分比）等[2]。近来也有使用命名实体识别针对某一领域的专有名词进行抽取的例子[3-4]。

命名实体识别作为信息抽取、问答系统、句法分析、机器翻译等应用领域的重要基础工具，得到了广泛的应用。其中也包括被应用于法院判决书信息抽取或罪犯信息报导与调查访谈系统等。然而，这些已经投入使用的系统却纷纷暴露出难以解决的问题：大部分已经投入使用的命名实体识别使用的是条件随机场算法（Conditional Random Fields，CRF）[5]，CRF 在训练时必须由测试员根据自身经验设置特征值，目前并没有可靠的标准用于表明选取的特征值是否合理，这导致 CRF 模型的测量精度时高时低。其次，由于 CRF 是完全依赖于已有训练词库进行实体判断，因此无法对词库中没有出现过的生词进行辨析，这一点在专业领域表现得更突出，在文本中存在大量专有名词或术语时，CRF 模型的提取效果会变得很差。

2012 年，加拿大多伦多大学的 Hinton 教授[6]提出深度学习的概念，被谷歌实际应用。时至今日，深度学习已经在图像识别，语音识别等方面获得了巨大成功[7-8]。由于深度学习可以从原始字符集上提取高级特征，使用深度学习来完成命名实体识别也成为了 CRF 识别之外的一种新思路。事实上，已经有相当数量的自然语言处理（Natural Language Processing，NLP）工作使用了深度学习进行实现，效果喜人[9-10]。本文利用基于深度学习的循环神经网络技术，对现有的法院判决书信息抽取系统做出改进，通过将大量的特征数据输入神经网络，让模型自动选取特征，让系统更稳定，准确率更高。

1 关键技术简介

1.1 条件随机场算法

条件随机场 CRF（Conditional Random Fields，CRF）[11]，是一种概率分布模型，在给定一组输入随机变量条件下，另外一组将会输出随机变量，它是一种判别式的概率无向图模型，是对条件概率分布建模。

[收稿日期] 2018－10－12
[基金项目] 湖北省教育厅科研计划研究项目基金（Q20141410）
[第一作者] 龚启文（1997－），男，湖北蔡甸人，湖北工业大学本科生，研究方向为机器学习

简而言之,CRF 是一种特征函数的集合,可以根据特征对一个标注序列进行打分,并据此给出最终结果,由于实际操作中特征有多种,因此最后的结果是一个综合式,而特征的规定方式则由外界决定。这也是为什么上面提到,CRF 的特征十分重要的原因。

CRF 的打分方式可以用下式表示:

$$P(l \mid s)=\frac{\exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i)]}{\sum_{l'}\exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i')]} \quad (1)$$

式(1)中 P 表示概率值,λ 表示权重,s 为样本本身,l 为标注序列,f 为特征函数,i 表示特征系列中的第 i 个值,l_i 表示第 i 个值对应的特征值。

1.2　循环神经网络

在传统的神经网络模型中,层与层之间是完全连接的,而每层之间的节点是无连接的。但是这种普通的神经网络对于许多问题是无能为力的。例如,为了预测句子中的下一个单词是什么,通常需要使用前一个单词,因为句子中的前后单词并不独立。循环神经网络(Recurrent netrual network,RNN)[12]是一种能将自身状态在网络中进行传递,以接受更广泛的时间序列结构输入的网络。实现逻辑为将上一次的节点状态记录下来,并在接下来的训练和输出中将之前记忆下来的结果纳入计算,与正常前馈网络不同,RNN 网络中当前时刻的输出值不仅和输入相关联,还和记忆有关。理论上,循环神经网络能够对任何长度的序列数据进行处理,但是在实践中,迫于运算能力的问题,往往假设当前的状态只与前面的几个状态相关。图 1 展示了一个典型的循环神经网络(RNN)结构。

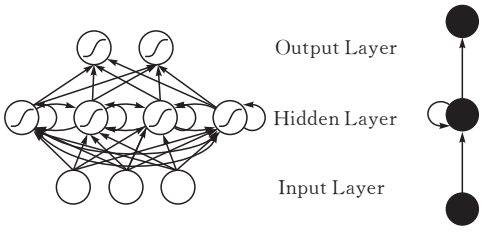

图 1　循环神经网络(RNN)结构

将循环神经网络(RNN)可视化的一种有效方法是将其在时间上进行展开,得到如图 2 所示结构。

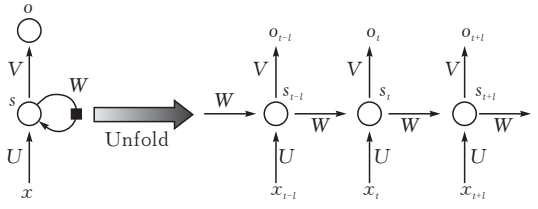

图 2　在时间上展开的 RNN

1.3　词向量

计算机无法直接使用字符状态的词进行计算,需要用其他方式进行表达,比较常用的两种方法为离散表示(one-hot representation)和分布式表示(distribution representation)。

离散表示把每个词表示为一个长向量。这个向量的维度是词表大小,向量中只有一个维度的值为 1,其余维度为 0,每种词对应一种长向量,这种方法相对传统,相当于给每个词分配一个 id,这导致离散表示所占用的空间极大,且非常容易引发维度灾难。但也带来一个好处,就是在高维空间中,很多应用任务线性可分。

分布式表示是将词转化成一种分布式表示,又称词向量。分布式表示将词表示成一个定长的连续的稠密向量,其有两个特性,首先词和词之间存在距离的概念,这对很多自然语言处理非常有帮助,其次词向量可以表达更多的信息,每一维都能用于表示隐性的含义。词向量本身需要通过训练获得,这一过程常用的架构包括 word2vec 或 glove 等。

2　BiRNN-CRF 算法模型设计

2.1　BiRNN-CRF 模型总体设计

本文设计了一个双向循环神经网络−条件随机场判别模型(BiRNN-CRF),见图 3。

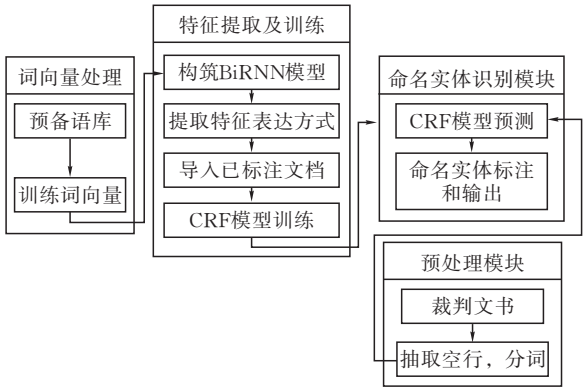

图 3　BiRNN-CRF 模型总体设计

如图 3 所示,与 CRF 模型相比,RNN-CRF 模型使用 BiRNN 构建了特征提取模块,替代了人工操作。同时增加了词向量训练模块用于处理数据的输入格式。

2.2　词向量语库训练

因为本文使用的分辨方式依然是 CRF,所以本文依然无法避免生词难以进行处理的问题,其一,CRF 本身对未知词的辨别效果比较差。其二,迫于人力有限,不太可能获得非常大的样本用于分辨指定实体词的已标记语料。所以本文使用词向量作为输入方式,词向量这种表示方法允许通过算法计算

两个词中间的距离，并以此作为词间的相似度。词向量本身的训练只需要词库即可，不需要手工标记的文本。

训练词向量常用的框架有 word2vec[13-14] 和 glove 两种，其中 word2vec 可直接用于训练中文语料，本模型中使用的即为 word2vec 框架，其原理如图 4 所示。

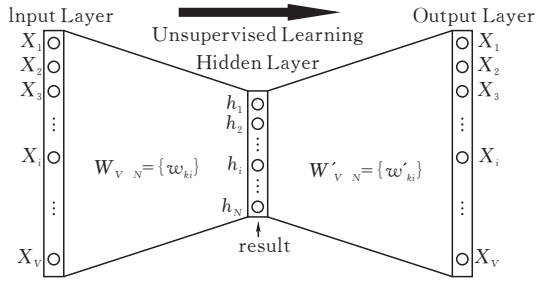

图 4 word2vec 模型原理图

2.3 双向循环神经网络

考虑到实际语言的组成逻辑，如果能同时获得前文和后文的信息，在语义判断时会有相当大的优势，因为语句的前后并非完全独立，后文中出现能作为判断条件的可能性也相当高，而且如果单纯使用阅读位置前的信息进行判断，由于没有足够的信息，句首的语义判断会变得非常困难。

然而，标准的循环神经网络（RNN）在时序上处理方式，导致它无法处理上下文信息。一种显而易见的解决办法是在输入和目标之间添加延迟，以便让网络有时步加入未来的上下文信息，也就是说在原有网络的基础上增加一定的时间帧用于预测输出。理论上，时间帧越大，能捕获的信息就会越多，但是从实际情况来看，过大的时间帧反而会导致准确率下降，这是由于捕获和记忆信息消耗了过多的运算能力，导致建模能力下降，因此在实际操作中时间帧的大小需要另行调节。

双向循环神经网络（BiRNN）[15] 即将两个相反的循环神经网络前后叠放，将两层循环神经网络的输出前后连接，以此达到传递前后信息的目的。图 5 展示的是一个沿着时间展开的双向循环神经网络。整个流程包括了输入到隐含层（w_1 和 w_3），隐含层自更新（w_2 和 w_5），从隐含层输出（w_4 和 w_6），同时由于整个过程中前后隐含层是独立的，所以不存在环状结构。

2.4 代价函数

本次使用的 BiRNN 模型逻辑回归使用的是自适应估计算法（AdamOptimizer）[16]，这是一种计算每个参数的自适应学习率的方法。

就思路而言，它的计算方式比较类似于 RMSprop 和 Momentum 算法，它会储存过去梯度的平

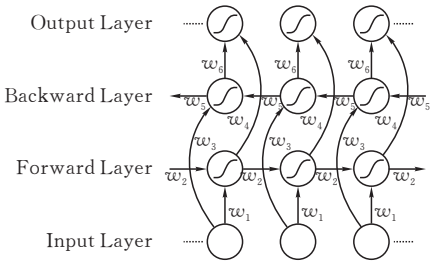

图 5 双向循环神经网络（BRNN）原理图

方 vt 的指数衰减平均值，同时也会保持过去梯度 mt 的指数衰减平均值，这是为了解决在自适应算法中学习率急剧下降的问题，它使得梯度方向不变的维度上变化速度更快，而在梯度有所改变的维度上更新减慢，以此减小抖动，并保持各维度的导数都在一个量级，在这个前提下则可以设置更大的学习率进行运算，以加快模型的训练速度。

运算过程如下：

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t. \tag{2}$$

$$\nu_t = \beta_2 \nu_{t-1} + (1-\beta_2) g_t^2. \tag{3}$$

如果式（2）、式（3）中的 mt 和 vt 没有经过随机初始化而是直接被初始化为 0 会导致向 0 偏置，为了避免这一情况，要进行偏差校正：

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}$$

$$\hat{\nu}_t = \frac{\nu_t}{1-\beta_2^t}$$

梯度更新算法如下：

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{\nu}_t} + \varepsilon} \hat{m}_t$$

2.5 CRF 特征模型训练

训练 CRF 模型的部分与常规训练一致[17]（图 6）。需要特别说明的有两点，其一，此处导入 CRF 模型训练的语料会被转化为词向量，在完成训练之后，再将需要预测的语料转化为词向量并输入，即可得到命名实体提取结果。其二，此处特征学习和特征抽取时使用的特征由 Bi-RNN 给出，而非测试员手动选取特征。

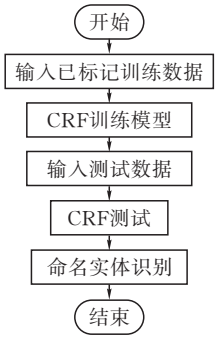

图 6 CRF 命名实体流程图

3　实验结果及分析

模型 BiRNN-CRF 的测试是在 Ubuntu16.06 环境下，使用 python2.7 和 tensorflow1.4.1 进行编写，使用显卡为 GeForce960M。

模型 BiRNN-CRF 使用中文维基百科的语料作为词向量训练素材，其中大约有 23 万篇中文语料，在经过去重，繁体简体统一，抽取空格，分词后，使用 word2vec 进行训练，可以得到一个 300 维度的词向量库，用于将标注文档和预测文档转换为词向量进行输入。

本模型中特征的提取方式由 Bi-RNN 训练结果得到，Bi-RNN 结构为常规结构，使用 tensorflow 自带的 BasicLSTMCell 和 bidirectional_dynamic_rnn 构建网络，隐藏层神经元数量和词向量维度一致，这里为 300 个神经元，使用的优化器为自适应估计算法，每次正向和反向传递之前要经过 dropout 层，dropout 系数设置为 0.6，输入的样本数据为已经标记完成的新闻语料库，分词由 jieba 完成，标注实体内容包括人物姓名、地名、职业、时间、罪名、机构名、法规名、案件名、法院名共计 9 种。

在 CRF 训练中导入 Bi-RNN 训练得到的特征，并以同样的语料库为准进行训练，将得到的模型用于法院判决书的预测即可得到测试结果，使用的判决书来源于 2012～2014 年间的滨州市已结案判决书文件，总计 21 668 份文档，预测用文档为从中抽取的十万行判决书内容文档。模型最后的抽取情况见表 1 与表 2。

表 1　BiRNN-CRF **实体抽取数量一览表**

人名	87.54	89.81	88.66
地名	87.88	99.57	93.36
职业	98.21	60.87	75.16
时间	91.54	99.74	95.46
罪名	90.91	84.61	87.65
机构名	98.11	96.36	97.23
法规名	88.24	89.47	88.85
案件名	94.28	99.24	96.70
法院名	99.24	97.19	98.20

表 2　CRF **实体抽取成功率一览表**

人名	89.78	90.58	88.64
地名	95.00	82.02	88.03
职业	94.92	86.35	90.43
时间	92.81	96.89	94.81
罪名	90.20	92.41	91.29
机构名	85.45	92.41	88.79
法规名	96.88	96.19	96.53
案件名	98.95	98.91	98.93
法院名	86.67	90.25	88.46

图 7 中人名召回率是以文本为单位进行统计的，即在同一个文本中重复出现的人名，只要能够成功提取一个即视为成功提取，但正确率依然是以词为单位进行计算的。此外，法院名是由提取内容进行正则判断二次提取得到的内容，因此合计中没有将其纳入计算。黑色柱状图表示 BiRNN-CRF 模型的实体抽取调和平均值，灰色柱状图表示 CRF 模型的实体抽取调和平均值，虽然职业相关的抽取召回率较低，但是其他实体抽取都相对稳定，尤其是机构名的抽取准确率极高。总体来看，BiRNN-CRF 模型相比于 CRF 模型比较突出的改进在于在地名上可以提高 6.05％的准确度，在机构名上可以提高 9.51％的准确度，但其他方向上提升并不明显，且对于职业实体词的精准度下降了 15.27％，各项实体数量统计后，模型的总体准确度提高约有 0.64％左右。

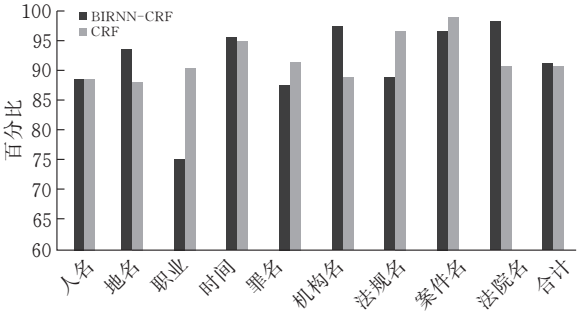

图 7　模型调和平均值对比图

4　结论及展望

本文利用基于深度学习的 BiRNN 模型，对 CRF 模型进行改进，使用神经网络模型学习得到特征选择模型，再通过词向量使 CRF 模型可以对未进行训练的词进行识别，取得了较好的结果。就目前来看，系统依然存在一定的问题和局限性。

从实际测试中得到的结果来看，虽然调和平均值确实有提高，但是提升并不大，推测可能是由于过于严苛的提取条件导致召回率偏低所致，通过修改 RNN 的超参数有可能进行优化，但是目前而言尚未发现神经网络的超参数与得到的 CRF 模型的准确率间的直接联系，如果能使用一定的标准对其超参数进行规范，得到的模型的效果有可能会更好。

无论是 CRF 本身还是词向量训练，使用的都是进行过分词的语料，这导致模型的训练效果和分词结果关系密切，本模型中使用的是 jieba 分词，就结果而论 jieba 分词对人名的识别效果并不理想，如果有精准度更高的分词手段，模型的训练效果能大大提升。

[参 考 文 献]

[1] 李保利,陈玉忠,俞士汶.信息抽取研究概述[J].计算机工程与应用,2003,39(10):1-5,

[2] 郭喜跃,何婷婷,信息抽取研究综述[J].计算机科学,2015,42(2):14-17

[3] Chen H.Chung W,Xu J J,et al. Crime data mining:a general framework and some examples[J].Conputer,2004,37(4):50-56.

[4] Chau M,Xu J J,Chen H.Extracting meaningful entities from police narrative reports[C] // Proceedings of the National Conference for Digital Government Research,2002:271-275

[5] 张帆,王敏.基于深度学习的医疗命名实体识别[J].计算机技术与自动化,2017,36(1):22-25

[6] Hinton G E,Salakhutdinov R R.Reducing the dimensionality of data with neural networks [J]. Science,2006,313(5786):504-507.

[7] Mikkolov T,Kombrink S,Burget L,et al.Extensions of recurrent neural network language mode[C] // 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE, 2011: 5528-5331.

[8] Jiang Z , Li L , Huang D . An unsupervised graph based continuous word representation method for biomedical text mining [M]. IEEE Computer Society Press, 2016.

[9] 庞亮,兰艳艳,徐君,等. 深度文本匹配综述[J].计算机学报,2017,40:55-70

[10] 张亮. 基于深度学习的中文微博文本实体识别研究[D].长沙:湖南大学,2017:83-97

[11] Lafferty J D, Mccallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]// Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 2001:282-289.

[12] Graves A. Generating sequences with recurrent neural networks[J]. Computer Science, 2013:2-6.

[13] 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.

[14] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013.

[15] Schuster M, Kuldip K. Bidirectional recurrent neural networks[C]// IEEE Transactions on Signal Processing, IEEE, 1997.

[16] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. Computer Science, 2014:14-21.

[17] 刘稳,王锦,李锐,等. 法院判决书关键信息抽取系统设计与实现[J].湖北工业大学学报,2018,33(1):63-67

Research on the Recognition Model of Court Judgment Named Entity Based on Deep Learning

GONG Qiwen,CHENG Yu,CHEN Jianxia,LI Chao,ZHANG Di,LONG Yishu

(*School of Computer Science,Hubei Univ. of Tech.,Wuhan* 430068, *China*)

Abstract: Named entity recognition, as an important basic tool in such application fields as information extraction, question and answer system, syntactic analysis, machine translation and others, has been widely used in court judgment information extraction system. However, the extraction effect of entity recognition becomes poor when existing technical models have a large number of proper nouns or terminologies in the text. In this paper, a two-way cyclic neural network-conditional random-airport discriminant model is thus developed to optimize the existing court judgment condition with the airport model. The experiment proves that the model can realize the selection process of automatic feature, and the accuracy rate is higher than that in the airport model.

Keywords: named entity recognition; deep learning; conditional random fields; bidirectional recurrent neural networks

[责任编校: 张岩芳]