计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(1):245−250 [doi: 10.15888/j.cnki.csa.006703]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 SVM-BiLSTM-CRF 模型的财产纠纷命名实体识别方法①

周晓磊[2], 赵薛蛟[1,2], 刘堂亮[3], 宗子潇[4], 王其乐[5], 里剑桥[6]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[3](辽宁省人民检察院沈阳铁路运输分院, 沈阳 110001)
[4](东北大学, 沈阳 110000)
[5](沈阳市第三十一中学, 沈阳 110021)
[6](大连理工大学, 大连 116621)
通讯作者: 赵薛蛟, E-mail: zhaoxuejiao16@mails.ucas.ac.cn

摘　要: 裁判文书中的命名实体识别是自动化审判的关键一步, 如何能够有效的分辨出案件的关键命名实体是本文的研究重点. 因此本文针对财产纠纷审判案件, 提出了一种基于 SVM-BiLSTM-CRF 的神经网络模型. 首先利用 SVM 筛选出包含关键命名实体的句子, 然后将正确包含此类实体的句子转化为字符级向量作为输入, 构建适合财产纠纷裁判文书命名实体识别的 BiLSTM -CRF 深层神经网络模型. 通过构建训练数据进行验证和对比, 该模型比其他相关模型表现出更高的召回率和准确率.

关键词: 命名实体识别; SVM; BiLSTM; CRF

引用格式: 周晓磊,赵薛蛟,刘堂亮,宗子潇,王其乐,里剑桥.基于 SVM-BiLSTM-CRF 模型的财产纠纷命名实体识别方法.计算机系统应用,2019,28(1):245–250. http://www.c-s-a.org.cn/1003-3254/6703.html

Named Entity Recognition Method of Judgment Documents with SVM-BiLSTM-CRF

ZHOU Xiao-Lei[2], ZHAO Xue-Jiao[1,2], LIU Tang-Liang[3], ZONG Zi-Xiao[4], WANG Qi-Le[5], LI Jian-Qiao[6]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[3](Shenyang Railway Transportation Branch of Liaoning People's Procuratorate, Shenyang 110001, China)
[4](Northeastern University, Shenyang 110000, China)
[5](Shenyang Thirty-first Middle School, Shenyang 110021, China)
[6](Dalian University of Technology, Dalian 116621, China)

Abstract: The recognition of the named entity in the judgment documents is the key step in the automatic trial. How to effectively distinguish the key named entity of the case is the key point in this study. Therefore, this study proposes a neural network model based on SVM-BiLSTM-CRF for property dispute of trial cases. First, the sentences containing the key named entities are selected by SVM, and then the sentences are converted into the character level vectors as input, and the BiLSTM-CRF deep neural network model suitable for the identification of the property dispute referee's named entity is constructed. By constructing training data for verification and comparison, the model shows higher recall and accuracy than other related models.

Key words: named entity; SVM; BiLSTM; CRF

① 收稿时间: 2018-05-24; 修改时间: 2018-06-15; 采用时间: 2018-07-10; csa 在线出版时间: 2018-12-26

随着国家法制建设不断进步, 人们的法律意识不断增强, 在遇到社会、经济生活中的纠纷时会自然的诉诸于法律审判. 这类案件虽然简单易断, 但由于数量急剧增多使得基层法院承受着十分沉重的工作压力. 因此对于简单的财产纠纷案件做到自动审判不但可以缓解基层法官的工作压力使得同类型案件审判一致, 更能增强民众用法律武器维护自身权利的动力. 而财产纠纷案件中命名实体的正确识别是完成自动化审判的非常重要的一步.

命名实体识别的目标是从语料中准确识别出专有名词或有意义的数量短语并加以归类[1]. 早期的命名实体识别主要是基于规则和字典的, 这种方法在处理复杂场景时会耗费人们的大量精力而且移植性差. 为了解决这些问题, 又出现了基于机器学习的方法, 但这些方法对特征选取的要求比较高. 而相比于上述的两类方法, 深度学习方法兼具泛化性和较少依赖人工特征的特点, 因此近年来, 深度学习在通用的命名实体识别领域运用广泛. CNN-CRF 例如: Collobert 等[2]提出了一种模型, 在 CNN 结构上运用 CRF 算法将标签转移得分加入到目标函数中. 在 CONLL2003 语料上取得了比较好的成绩. Huang 等[3]通过人工设计拼写特征提出训练了一种 BiLSTM-CRF 模型, 该模型在 CONLL2003 语料上的 *F*1 值达到了 88.83%.

财产纠纷裁判文书的关键实体主要包括案件涉及的财产形式、财产数额等. 经过分析实际的裁判文书后发现, 难点主要在于: (1) 纠纷涉及财产形式多样. (2) 裁判文书中包含法院认定的涉及纠纷的财产在整篇文书中出现比重较小. (3) 财产描述形式多样. 由于 BiLSTM-CRF 模型在通用领域的效果突出, 于是使用该模型对样本进行了模型训练, 但结果发现实际的输出并不理想. 在分析原因后发现是由于上述第二个难点导致了训练数据的不平衡. 为了解决这一问题, 本文提出一种基于 SVM-BiLSTM-CRF 的财产纠纷裁判文书命名实体识别模型. 以提高对财产纠纷裁判文书中涉案财产的识别精度.

1 数据集构建

本文通过从中国裁判文书网下载大量财产纠纷裁判文书, 在进行适当的数据预处理并手工标注后构建财产纠纷的语料库. 其中一半作为训练集进行模型训练, 另一半则作为测试集用于评价模型的性能.

1.1 数据预处理

裁判文书是一种半结构化的文本, 通常的结构如图 1 所示.

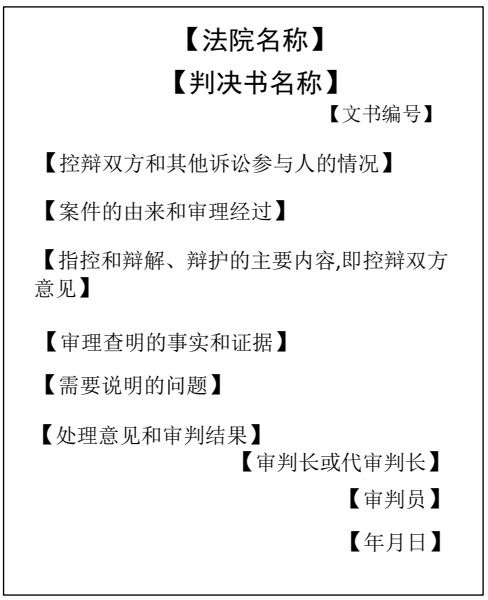

图 1 裁判文书的结构图

由于审判结果和审理查明的事实与证据存在直接关系, 所以从审理查明的事实和证据中提取的财产命名实体具有研究价值. 通过统计发现, 在审判文书中描述审理查明的事实的起始句包含以下说明词: “经审理查明”, “经审理认定”, “经开庭审理查明”, “经开庭审理认定”, “审理中查明”, “审理中认定”, “确定如下事实”, “认定如下事实”, “认定以下事实”, “查明如下事实”, “查明以下事实”, “本案事实如下”, “查明事实如下”, “确定事实如下”等. 同时, 在需要说明的问题部分起始句包含“本院认为”, 审判结果部分起始句包含“判决如下”. 通过这些触发词, 将审理查明的事实提取出来进行分句、分词、去停用词等处理.

1.2 数据标注

1.2.1 财产类别

我国的《民法通则》对财产有如下定义:财产是指拥有的金钱、物资、房屋、土地等物质财富:国家财产、私人财产, 具有金钱价值、并受到法律保护的权利的总称[4]. 根据上述定义, 将财产分为三种, 即动产, 不动产和知识财产. 据此, 本文将审判案件中涉及的财产分别标注为以下几个类别:

动产: 由于财产纠纷案件涉及金钱纠纷比例较大, 所以将动产的标注类别细分为 money 与 nonmoney.

不动产: 标注为 realestate

知识财产: 标注为 intelpropert.

1.2.2 四词位法

在汉语语言文字中, 每个词都是由一个或多个字组成的. 例如: “现金”是两字词, “上轿礼”是三字词. 组成词语的每一个汉字在一个特定的词语中都占据一个特定的构词位置, 即词位. 词位的种类根据研究的需要可以自行定义. 在已有的工作中常用的有四词位标注集 (B、M、E、S) 和六词位标注集 (B、B_1、B_2、M、E、S)[5]. 在本文中, 采用的是四词位集, 用 B 表示词的开始, M 表示词的中部, E 表示词的结尾, O 表示其他非财产的字, 并结合财产类别进行标注. 表 1 是一个标注例子.

表 1 标注实例

字	词性	标注
经	p	O
其	rz	O
手	n	O
交	v	O
女	n	O
方	n	O
彩	n	B-money
礼	n	M-money
款	n	M-money
7	m	M-money
5	m	M-money
0	m	M-money
0	m	M-money
0	m	M-money
元	q	E-money

2 SVM-BiLSTM-CRF 模型

2.1 模型整体框架

SVM-BiLSTM-CRF 模型由三个模块组成: SVM 模块、BiLSTM 模块和 CRF 模块. 整体模型框架图如图 2 所示. 首先通过查询词向量表将输入的语句转换成相应的词向量序列, 然后输入 SVM 进行判断. 如果不含财产实体, 则将所有的字标记为 O, 否则则通过查询字符向量表获得相应的字符向量序列. 并将这些字符向量序列输入 BiLSTM 进行实体识别. 最后 CRF 模块将 BiLSTM 的输出进行处理得出一个最优的标记序列.

2.2 SVM 模块

支持向量机 (SVM) 是在 VC 维理论和结构风险最小化原理基础上建立起来的机器学习方法[6]. 它的基本模型是在特征空间中寻找间隔最大化的分离超平面的线性分类器, 在解决小样本、非线性和高维模式识别问题方面表现出特有优势[7]. 因此, 为了解决包含财产实体的句子占案件描述句子的比重不高的问题, 本文使用 SVM 将无用的句子直接筛除, 使得包含财产实体的句子中进行进一步的标注训练可以保持训练数据的平衡.

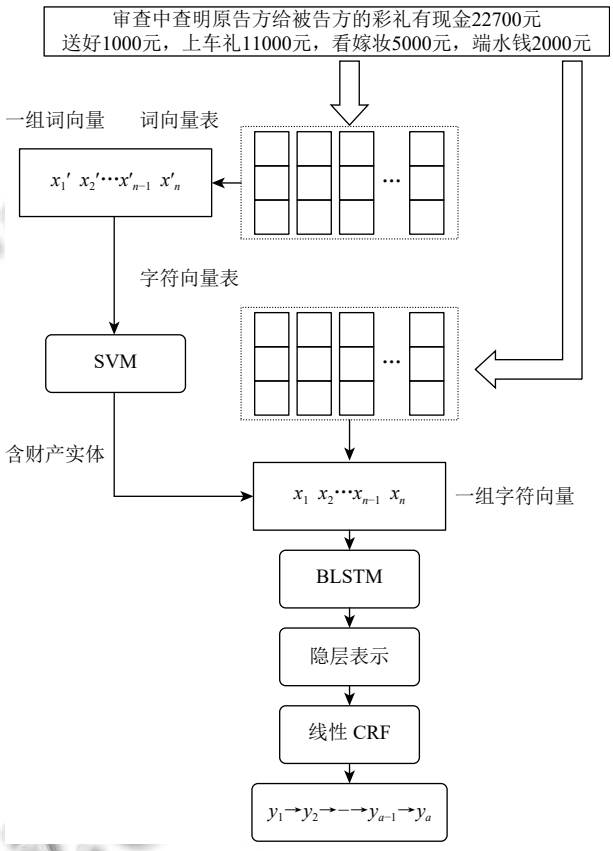

图 2 财产纠纷案件命名实体识别的 SVM-BiLSTM-CRF 模型

在训练开始, 首先将训练样本经过分词, 去停用词, 在不影响分类精度的情况下利用 tf-idf 进行特征降维形成词向量表$\{w_1, w_2, ..., w_n\}$, n=特征维度. 对于一个句子$S_i = \{w_1', w_2', ..., w_m'\}$, m=句子长度, 经过词向量表处理, 形成一个特征向量$\{x_1', x_2', ..., x_n'\}$, 利用核函数$\varphi$与标签$y_i$一起加入到式 (1) 中.

$$\begin{cases} \max\limits_{\alpha}\left(\sum\limits_{i=1}^{n}\alpha_i - \frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}\alpha_i\alpha_j y_i y_j(\Phi(x_i)\cdot\Phi(x_j))\right) \\ \text{s.t.}\ \sum\limits_{i=1}^{n}\alpha_i y_i = 0 \\ 0 \le \alpha_i \le C \end{cases} \tag{1}$$

其中, C 是惩罚系数, $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)^{\mathrm{T}}$为拉格朗日乘

子向量. 这是线性不可分的线性支持向量机的学习问题转化而成的对偶问题. 但是由于求解复杂度过高, 本文采用 SMO 算法来进行求解.

SMO 是 John C. Platt 于 1996 年提出一种启发式算法, 其思想是要将原问题分解成一系列小规模凸二次规划问题, 从而获得原问题最优解的方法. SMO 算法在每次迭代时选择两个拉格朗日乘子并同时固定其他乘子, 针对选择的乘子构建一个目标函数值更小的二次规划问题, 因为子问题可以通过解析方法求解, 所以可以大大提高整个算法的运算速度. SMO 算法的伪代码如算法 1.

算法 1. SMO 算法

1) 创建一个α并初始化为 0 向量.
2) 当迭代次数小于最大迭代次数时执行循环, 否则跳出循环返回结果.
3) 循环遍历数据集中的每一个数据向量, 如果该向量可以被优化, 则随机选择另外一个数据向量, 并同时优化这两个向量. 如果两个向量不能被优化, 则退出循环.
4) 如果所有向量都没有被优化, 则增加迭代次数, 进入下一次循环. 否则将迭代次数置 0, 重新进行迭代.

2.3 BiLSTM 模块

长短时记忆网络 (Long Short-Term Memory, LSTM) 是由 Schmidhuber 于 1997 年提出的. 它是一种具有特殊结构的 RNN 网络, 但是与传统 RNN 不同, 它解决了由于序列过长而产生的的长程依赖 (long-term dependencies) 问题. 网络模块示意图如图 3 所示. 其中包含四层神经网络, 最上面的一条线贯穿所有串联在一起的 LSTM 单元, 使得 LSTM 状态从第一个单元开始一直移动到最后一个单元, 在这个过程中只存在少量的线性干预和改变. LSTM 采用独特的门结构来控制 LSTM 单元对信息流中信息的添加和删减. 门结构一共有三类, 分别是输入门 (input gates), 忘记门 (forget gates) 和输出门 (output gates)[8]. 如果t时刻用i_t、f_t、o_t、C_t分别表示三种门和细胞状态, 则有:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{5}$$

$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{7}$$

其中, σ是激活函数 sigmod, tanh 是双曲正切激活函数. x_t 作为当前时刻的输入与上一层的输出 h_{t-1} 一起, 首先进入忘记门的 sigmod 层来决定从细胞状态中丢弃的信息. 接下来 x_t 和 h_{t-1} 通过输入门的 sigmod 层生成将更新的信息, 同时通过 tanh 层创建一个新的候选向量, 然后将之前的细胞状态信息加上新的候选信息确定前细胞状态中存储的信息[9]. 最后, X_t 和 h_{t-1} 进入输出门的 sigmod 层来确定细胞状态信息中的哪个部分输出然后与 tanh(C_t) 相乘得到输出值.

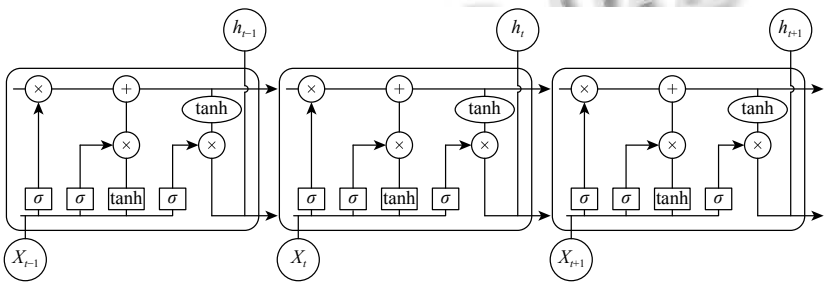

图 3 LSTM 网络模块示意图

而双向长短时记忆网络 (Bidirectional Long short-Term Memory, BiLSTM), 其原理是将两个时序方向相反的长短时记忆网络结构连接到同一输出, 以此来获取历史和未来信息. 因此相比于其他的 RNN 网络需要等到后面的时间节点才能获取未来信息, 该网络结构可以更充分的利用上下文信息. 我们利用该网络结构这一优势, 用 LSTM 对每个句子进行前向和后向的计算, 然后将得到的两个结果向量进行拼接得到最终的隐层表示.

2.4 CRF 模块

由于单独使用 BiLSTM 生成的结果可能在标注序列并不是全局最优, 为方便后续通过标注提取完整的

命名实体, 提高实体识别的正确率和召回率, 所以本文在 BiLSTM 层上加上一个线性 CRF 模块. 通过分析相邻标签的关系以获得一个全局最优的标记序列. 对于一个经过 BiLSTM 处理后的输出矩阵 P, P 的大小是 $n\times k$, 其中n是句子中包含的词数, k表示标签的种类. 其中$P_{i,j}$为该句第i个词映射到tag_j的非归一化概率, 然后引入状态转移矩阵A, 其中$A_{i,j}$表示时序上从第i个状态转移到第j个状态的概率, 则对于一个观测序列X的对应的标记序列$y=\{y_1,y_2,...,y_n\}$, 定义分数为:

$$s(X,y)=\sum_{i=1}^{n}(A_{y_i,y_{i+1}}+P_{i,y_i}) \tag{8}$$

对输入序列X的所对应的每个标记序列y计算 $\max s(X,y)$, 运用动态优化算法得到最终答案.

3 结果

3.1 评价指标

本文分别使用准确率、召回率和 $F1$ 值三个评价指标来对实验结果进行评价. 三种评价指标的表达式分别为:

$$Precision=\frac{correct}{correct+missing} \tag{9}$$

$$Recall=\frac{correct}{correct+spurious} \tag{10}$$

$$F=\frac{(beta^2+1.0)\times Precision\times Recall}{beta^2\times Precision+Recall} \tag{11}$$

其中, 准确率 (*Precision*) 为测试样本中识别正确的命名实体数量占总的命名实体数量的比例. 召回率 (*Recall*) 为正确识别为财产命名实体的数目占实际财产命名实体总数的比例. $F1$ 值则是当 *beta* 为 1 时对上述两个评价指标的加权平均.

3.2 实验比较

为了有效验证本文提出模型的合理性并证明模型中每个模块的必要性, 在仿真实验中得到 SVM-BiLSTM-CRF 模型的相关数据后, 又分别进行了 BiLSTM-CRF 模型、SVM-LSTM-CRF 模型以及 SVM-BiLSTM 模型在测试集上的性能评价实验. 并通过整合四次实验的结果, 进行了数据对比. 对比结果如表 2 所示.

表 2 对比实验结果 (单位: %)

模型	*Precision*	*Recall*	*F*1 值
SVM-BILSTM-CRF	93	92	92
BILSTM-CRF	80	43	56
SVM-LSTM-CRF	89	85	87
SVM-BILSTM	92	89	90

3.3 结果分析

(1) 移除 SVM 模块结果分析

由于提取出的财产纠纷案情包含财物命名实体的比例并不大, 所以会有大量标注为 O 的实体存在, 在未包含 SVM 模块的模型中, 训练得到的模型由于标注为 O 的实体占比过多, 造成了虽然准确率非常高但是召回率很低的情况. 而本文提出的模型比不包含 SVM 模块的模型的 F1 值高出 36%, 精确度高出 13%.

充分证明 SVM 模块在本模型中的重要作用.

(2) 替换 BiLSTM 模块结果分析

从 SVM-LSTM-CRF 模型与 SVM-BiLSTM-CRF 模型的结果数据对比中可以看到, 本文所提出的模型比使用 LSTM 的模型准确度高 4%, 召回率高 7%. 结果表明双向长短时记忆网络通过提取句子的上下文信息, 对结果产生了积极作用.

(3) 移除 CRF 模块结果分析

在本文提出的模型中, 线性 CRF 模块的主要作用就是根据相邻标签之间的关系优化神经网络输出的结果标签. 从实验数据中可以看到, 有 CRF 模块会比无 CRF 模块 F1 高 2%, 召回率高 4%. 在结果分析中发现, CRF 对长度较大或带有形容词的实体识别性能较高, 诸如“彩礼人民币九万九千五百元”、“‘海尔’牌电冰箱一台”等都能被 SVM-BiLSTM-CRF 正确识别, 但是 SVM-BiLSTM 则无法正确识别. 由此可见线性 CRF 模块的加入有助于提高模型的识别精度.

4 结论

本文针对财产纠纷审判文书中的财产实体识别问题进行了研究, 提出了通过 SVM 首先进行筛选判断是否包含财产实体, 然后通过神经网络和 CRF 进行进一步识别的方法. 为了训练模型和验证模型的有效性, 构建了裁判文书标注数据集. 实验最后的结果表明, 本文提出的 SVM-BiLSTM-CRF 模型在对财产纠纷审判文书中的关键实体识别有非常高的准确率和召回率, 从而能够为后续的财产纠纷审判案例自动判决工作奠定基础.

参考文献

1 刘海娟, 刘文展. 基于双向量模型的话题跟踪. 无线电工程, 2016, 46(2): 27–30. [doi: 10.3969/j.issn.1003-3106.2016.02.07]

2 Pinheiro PHO, Collobert R. Recurrent convolutional neural networks for scene parsing. arXiv: 1306.2795, 2013.

3 Huang ZH, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv: 1508.01991, 2015.

4 戴靖. 篡夺公司机会禁止制度研究[硕士学位论文]. 长沙: 中南大学, 2013.

5 汤步洲. 序列标注问题的监督学习方法及应用[博士学位论文]. 哈尔滨: 哈尔滨工业大学, 2011.

6 郑红军, 周旭, 毕笃彦. 统计学习理论及支持向量机概述. 现代电子技术, 2003, (4): 59–61. [doi: 10.3969/j.issn.1004-373X.2003.04.022]

7 丁晟春, 吴靓婵媛, 李红梅. 基于 SVM 的中文微博观点倾向性识别. 情报学报, 2016, 35(12): 1235–1243. [doi: 10.3772/j.issn.1000-0135.2016.012.001]

8 刁琦, 古丽米拉·克孜尔别克, 钟丽峰, 等. 基于循环神经网络序列标注的中文分词研究. 计算机技术与发展, 2017, 27(10): 65–68. [doi: 10.3969/j.issn.1673-629X.2017.10.014]

9 任智慧, 徐浩煜, 封松林, 等. 基于 LSTM 网络的序列标注中文分词法. 计算机应用研究, 2017, 34(5): 1321–1324, 1341. [doi: 10.3969/j.issn.1001-3695.2017.05.009]