文章编号:1007-5321(2021)04-0129-06　　**DOI**:10.13190/j.jbupt.2020-241

基于 BERT-BiLSTM-CRF 的法律案件实体智能识别方法

郭知鑫,　邓小龙
(北京邮电大学 网络空间安全学院, 北京 100876)

摘要: 在智能法务系统应用中,人工智能自然语言处理相关技术常采用静态特征向量模型,算法效率低,精度偏差较大. 为了对法律文本中的案件实体进行智能识别,提高案件的处理效率,针对动态字向量模型提出以基于转换器的双向编码表征模型作为输入层的识别方法. 在其基础上通过融合双向长短期记忆网络和条件随机场模型,构建了高精度的法律案件实体智能识别方法,并通过实验验证了模型的性能.

关　键　词: 自然语言处理; 智能法务; 基于转换器的双向编码表征模型
中图分类号: TP391.1　　**文献标志码**: A

Intelligent Identification Method of Legal Case Entity Based on BERT-BiLSTM-CRF

GUO Zhi-xin,　DENG Xiao-long
(School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: In the past, artificial intelligence natural language processing related technologies often used static feature vector models in the application of intelligent legal systems, which had problems such as low algorithm efficiency and large accuracy deviations. To intelligently identify case entities in legal texts and improve case processing efficiency, the dynamic word vector model is studied, and a recognition method based on the bidirectional encoder representations from transformers model as the input layer is proposed. Based on the fusion of bi-directional long short-term memory and conditional random fields models, a high-precision method of intelligent identification of legal case entities is constructed. The performance of the model is verifiedthrough experiments.

Key words: natural language processing; intelligent legal affairs; bidirectional encoder representations from transformers model

随着我国法治建设步伐的不断加快及人工智能技术的迅速发展,人工智能技术在司法领域中的应用也日渐成熟. 司法文书不同于普通文本,通常包含大量法律专业术语并有一词多义等情况,智能标注和提取办案笔录等文字材料的相关实体能大幅提高办案效率,相关研究具有重要的意义. 但传统的中文命名实体识别模型存在提取能力不足的问题,为此,采用谷歌公司提出的基于转换器的双向编码

收稿日期: 2020-11-30
基金项目: 国家重点研发项目子课题(2017YFC0820603)
作者简介: 郭知鑫(1994—), 男, 硕士生.
通信作者: 邓小龙(1977—), 男, 副教授, E-mail:shannondeng@bupt.edu.cn.

表征(BERT,bidirectional encoder representations from transformers)模型[1]为特征表示层. 该模型是以Transformers模型为主要框架的双向编码表征模型,具有很强的文本特征表示能力. 笔者将BERT模型与双向长短期记忆网络(BiLSTM,bi-directional long short-term memory)命名实体识别模型相融合,最后结合状态转移矩阵(CRF,conditional random fields)输出全局最优的序列. 主要贡献包括2个方面.

1)将BERT模型作为输入层,获取字向量. 由于该模型能充分利用上下文之间的字词关系,可有效提取文本特征. 针对法律文本中存在很多指代性较强的内容,创新性地采用BERT模型进行上下文特征的综合提取,提高了文本预处理的准确率.

2)将BERT + BiLSTM + CRF模型与从实际法律大数据中抽取实体相结合,提出基于BERT + BiLSTM + CRF模型的法律智能实体识别模型,并使用《人民日报》和CAIL(challenge of artificial intelligence in law)法研杯相关法律文本经典数据集以及人工整理、标注的法律笔录文本数据集进行验证. 结果表明,所提方法能够有效提高法律案件中相关命名实体识别的效果,其性能优于Word2Vec经典模型.

1 相关研究及BERT模型简介

在自然语言处理中广泛使用的命名实体识别(NER, named entity recognition)一词最早由Grishman等[2]提出. 命名实体识别一直是自然语言处理领域的研究热点. 最早期的基于规则和字典方法的识别和归纳过程开销大,识别效率也普遍不高. 发展到传统机器学习阶段,解决命名实体识别问题的技术主要包括隐马尔可夫模型(HMM,hidden Markov models)[3]、决策树模型(DT,decision tree)[4]、最大熵模型(ME,maximum entropy model)[5]以及条件随机场(CRF,conditional random field)算法[6]等. 这些传统的机器学习方法存在对文本提取特征要求高等缺点. 近年来,基于深度学习的NER方法得到了快速发展,神经网络方法将预训练的词向量输入卷积神经网络、递归神经网络中,使用大型的未标记语料库进行词向量训练,实现了端到端的命名实体识别训练.

Word2Vec是目前自然语言处理领域使用较广泛的词向量训练工具,但是Word2Vec局限于自己的训练窗口大小,只是对词本身进行处理,并不符合实际语言环境中的不同含义或指代意义. 因此,在后续的研究过程中,长距离依赖的长短期记忆网络模型的应用,在一定程度上提高了训练效果. Zhang等[7]提出的用于中文处理的LSTM模型的格子模型很好地提升了中文处理的效果,但对于上述语言训练模型来说,很难符合人类在使用语言时上下文相结合的方式,准确率也没有太大的提升.

近年来,BERT模型被广泛应用于语言结构、语言相似度等实验[8-12],均取得了较好的效果,整体架构如图1所示. 其中,Trm为Transformer模型名称的缩写. BERT模型采用了双向Transformer模型结构,吸收了其他模型的优点. 实验的结果表明,双向训练的语言模型对语境的理解会比单向的语言模型更深刻、准确.

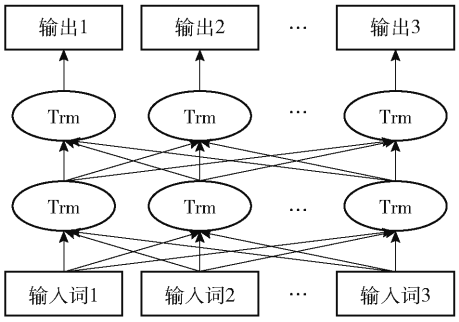

图1 BERT模型结构

2 BERT模型预训练任务简介

BERT模型是一种自然语言处理预训练的语言表征模型,其特点在于通过计算上下文的关系权重来确定文本中的主要特征,基于所有层融合上下文的语境来确定真正的上下文关系. BERT模型的预训过程包括掩码语言模型和下一句预测2个任务.

2.1 掩码语言模型

在BERT模型中采用了一种随机屏蔽部分输入“token”的方式来完成用训练深度双向表示的任务,在训练中只预测被屏蔽的“token”,这个过程称为掩码语言模型.

掩码语言模型由Word2Vec中的连续词汇学习演变而来. 根据中心词周围的词来预测中心词,并且对每一个词都做一遍预测. 而掩码语言模型在训练的过程中随机地掩盖每个序列,一般默认掩盖序列中15%的标签. 与从左到右的语言模型预训练方式不同,掩盖标签的目标是基于上下文来预测被掩

盖的词. 双向的 Transformer 模型并不清楚哪些词被掩盖,所以在训练过程中需要对每个词进行上下文的处理,随机掩盖 15% 的词,类似于英语考试中的完形填空,并不会影响模型对整体语言段落的理解.

2.2　下一句预测

在法律文本实体识别的任务中,不仅需要分析字词之间的上下文关系,还需要理解、推理句子之间的关系,而句子之间的关系无法由语言模型直接建模,所以在 BERT 模型中采用了一种二值任务下一句预测,如表 1 所示.

表 1　下一句预测

输入句子	标签
[CLS]要宣传我国禁毒工作取得的[MASK]成绩[SEP] 以[MASK]犯罪、教育群众[SEP]	IsNext
[CLS]要宣传我国禁毒工作取得的[MASK]成绩[SEP] 并竭力[MASK]开脱罪责[SEP]	NotNext
[CLS]被告人徐某[MASK]拳头击打被害人孙某的 面部[SEP] 造成[MASK]鼻骨骨折[SEP]	IsNext
[CLS]被告人徐某[MASK]拳头击打被害人孙某的 面部[SEP] 并如实供述[MASK]自己的犯罪事实	NotNext

输入序列中下一句是从文本中随机选择的,利用 Transformer 模型来判断文中的语句对是否存在连续关系,进而实现对语句关系之间的建模. 同时它还增加了一些特殊作用的标志:[CLS]标志放在第 1 个句子的首位,经过 BERT 模型得到的表征向量可以用于后续的分类任务;[SEP]标志用于分开 2 个输入句子,例如输入句子 A 和 B,要在句子 A 和 B 后面增加[SEP]标志;[MASK]标志用于遮盖句子中的一些单词,将单词用[MASK]遮盖之后,再利用 BERT 模型输出的[MASK]向量进行预测.

3　BERT-BiLSTM-CRF 模型简介

BERT-BiLSTM-CRF 模型中,采用 BERT 模型作为特征表示层进行词向量的获取;然后通过 BiLSTM 模型深度学习全文特征信息, 进行特定的法律案件实体识别;最后在 CRF 算法层对 BiLSTM 模型的输出序列进行处理, 结合 CRF 算法,根据相邻之间的标签得到一个全局最优序列[13],其结构如图 2 所示.

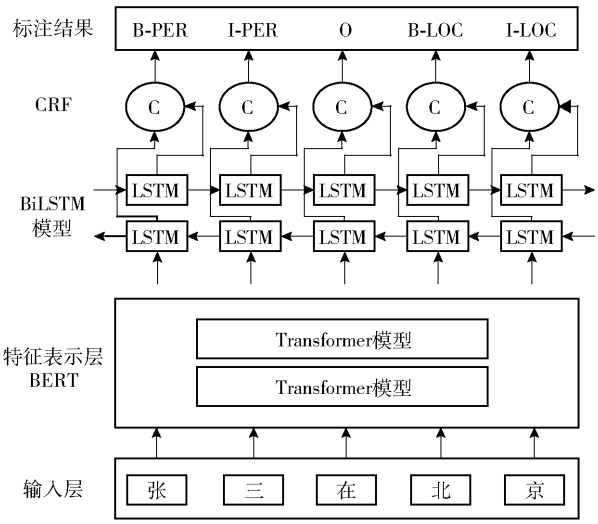

图 2　BERT-BiLSTM-CRF 模型结构

3.1　BiLSTM 模型

长短期记忆网络(LSTM,directional long short-term memory)模型是 Hochreiter 等[14]于 1997 年针对循环神经网络(RNN,recurrent neural networks)的梯度消失和梯度爆炸问题提出的改进模型. BiLSTM 模型是由前向 LSTM 模型与后向 LSTM 模型组合而成. 通常 LSTM 模型从前向后编码句子,只掌握了从前到后的上下文信息,没有掌握从后到前的上下文信息. 因此,可将前向 LSTM 模型和后向 LSTM 模型组合成 BiLSTM 模型来学习双向上下文信息[15]. LSTM 模型的计算过程如图 3 所示.

首先,LSTM 模型通过遗忘门决定上一个细胞中需要抛弃的信息,接收上一时刻的输出和本时刻的输入,通过式(1)计算出一个 0 ~ 1 的权值,表示从完全抛弃到完全保留. t 时刻遗忘门的结果为

$$f_t = \sigma(\boldsymbol{W}_{\mathrm{f}}[h_{t-1}, X_t] + \boldsymbol{b}_{\mathrm{f}}) \tag{1}$$

其中:$\boldsymbol{W}_{\mathrm{f}}$表示遗忘门的权重矩阵,$\boldsymbol{b}$ 为三门的偏置矩阵.

输入门控制本细胞需要加入的信息,计算公式如式(2)、式(3)所示.

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{\mathrm{i}}[h_{t-1}, X_t] + \boldsymbol{b}_{\mathrm{i}}) \tag{2}$$

$$C_t = f_t C_{t-1} + i_t \tanh(W_{\mathrm{f}}[h_{t-1}, X_t] + b_{\mathrm{c}}) \tag{3}$$

$$\boldsymbol{O}_t = \sigma(W_{\mathrm{o}}[h_{t-1}, X_t] + \boldsymbol{b}_{\mathrm{o}}) \tag{4}$$

$$H_t = O_t \tanh(C_t) \tag{5}$$

其中:X_t 为 t 时刻的输入内容;h_{t-1}为 $t-1$ 时刻的输出内容;C_t 为 t 时刻 LSTM 的细胞状态;$\boldsymbol{O}_t$ 为全连接层后逻辑回归的生成矩阵;$\boldsymbol{i}_t$ 为输入门中 t 时刻的 n 维向量,取 0 ~ 1 之间的数字;H_t 为 t 时刻当前状态

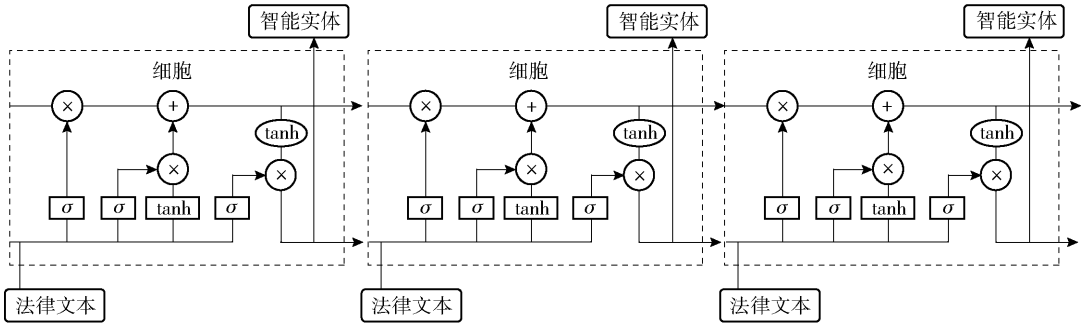

图 3 LSTM-Cell 内部结构

的输出内容.

输出门用于决定哪些信息可作为当前阶段的任务输出,计算过程如式(4)、式(5)所示. $\boldsymbol{W}_{\mathrm{i}}$表示输入门的权重矩阵,$\boldsymbol{W}_{\mathrm{o}}$表示输出门的权重矩阵;$\sigma$ 和 tanh 为激活函数.

3.2 CRF 算法

CRF 算法是以指定的随机变量作为输入,求解输出随机变量条件概率分布的一种算法,近年来被广泛应用于词性标注、句法分析和命名实体的识别领域[16]. CRF 算法能够利用相邻标注结果的关系,结合法律文本中存在的大量指代名词这个实际情况,得出全文中最优的标记序列结果,基本算法如下.

定义 p_{ij}为第 i 个符合第 j 个标签的概率. 输入的句子序列 $x=\{x_1,x_2,\cdots,x_n\}$ 与其预测序列 $y=\{y_1,y_2,\cdots,y_n\}$的得分为

$$S(x,y)=\sum_{i=0}^{n}B_{y_1,y_{i+1}}+\sum_{i=1}^{n}p_iy_i \tag{6}$$

其中 B 表示预测序列中的实体状态特征.

所有可能的序列路径归一化之后得到关于输出序列 y 的概率分布为

$$P(y|x)=\frac{\mathrm{e}^{S(x,\tilde{y})}}{\sum\limits_{\tilde{y}\in Y_x}\mathrm{e}^{S(x,\tilde{y})}} \tag{7}$$

在训练过程中标记序列的似然函数,即最大化关于正确标签序列 y^* 的对数概率,有

$$\lg\left(p(y^*|S)\right)=S(x,y^*)-\lg\left(\sum_{\tilde{y}\in Y_x}\mathrm{e}^{S(x,\tilde{y})}\right) \tag{8}$$

其中:$\tilde{y}$ 表示所有可能的标记集合,包括不符合 BIO(beginning-inside-outside)三元标记规则的序列. 采用句级似然函数的目的在于促进模型生成正确的标签序列. 预测结果是由式(9)得出的整体概率最大的一组序列,分类是通过 $K(x,\tilde{y})$ 函数完成的.

$$y^*=\arg\max K(x,\tilde{y}) \tag{9}$$

所有经 BiLSTM 模型层输出的分数将作为 CRF 算法层的输入. CRF 算法层最主要的特点是可以学习到句子的隐含约束条件. 例如,每个句子中的第 1 个字一定是“B-”或者“O”,绝对不可能是”I-”. 因为 LOC,ORG,PER 等命名实体开头的字都是用“B-”表示,没有“B-”绝对不可能有“I-”;另外,类似于“B-LOC I-ORG”这种组合一定是错的. 有了这些基本原则,再结合整体最大概率,错误的预测序列将会大量减少.

4 实验与分析

4.1 实验数据及指标

实验采用北京大学计算语言学研究所发布的 1998 年上半年语料、2018、2019 年 CAIL 法研杯相关案件文本、笔者标注的部分案件实体以及网上公开的案件相关文本作为数据集.《人民日报》语料中已经按照三元标记{B,I,O}进行了标注,B 表示分类实体的第 1 个字,I 表示分类实体第 2 字及其后面的字,O 表示不属于特定实体的名词,同时也将 CAIL 法研杯以及部分案件笔录相关案件文本进行标注. 实体名称包括 LOC,ORG 和 PER. LOC 表示案件地点、ORG 表示特定组织机构名,PER 表示案件人物. 将 1998 年 1~5 月的数据和 2018 年 CAIL 法研杯相关数据作为训练集,将 1998 年 6 月和 2019 年部分 CAIL 法研杯的相关数据作为测试集,训练和测试集的信息如表 2 所示.

表 2 相关语料数据统计

数据集	字数	LOC	PER	ORG
训练集	55 902 952	21 451	38 534	90 155
测试集	859 261	3 652	5 689	14 563

采用常用命名实体识别的评价指标，将每一类特定实体的准确率(p)、召回率(r)以及调和平均数(F_b)作为模型性能的评价标准，定义

$$p = \frac{A_{co}}{A_{co} + A_{in}} \times 100\% \tag{10}$$

$$r = \frac{A_{co}}{A_{co} + A_{d}} \times 100\% \tag{11}$$

$$F_b = \frac{2pr}{p + r} \times 100\% \tag{12}$$

其中：A_{co}为标注正确的实体数量，A_{in}为标准错了的实体数量，A_d 为未标注出的实体数量.

4.2　实验环境及参数

实验采用 macOS Catalina 操作系统，中央处理器的型号为 2.2 GHz 六核 Intel Core i7，实验语言为 Python3.6 版本，内存配置为 16 GB 2 400 MHz DDR4. 采用谷歌公司人工智能团队开发的深度学习框架 Tensorflow 1.12.0 搭建实验模型，该模型在各类深度学习任务中被广泛应用. 实验中的输入纬度 seq_length 为 128，训练集的 batch_size 为 32，测试集的 batch_size 为 8，learning_rate 为默认值.

4.3　实验结果及分析

表 3 所示为 BERT + BiLSTM + CRF 模型算法对{B,I,O}三元组的提取结果，其中 ORG 组织机构标注结果的准确率偏低，原因在于数据集中对特定组织机构的简称较多. 模型对于案件人物的标注准确率特别高，达到了 97% 以上.

表 3　BERT + BiLSTM + CRF 模型算法结果

实体名称	准确率	召回率	调和平均数
LOC(地点)	95.49	95.44	95.46
ORG(组织)	90.38	92.91	91.63
PER(人物)	97.95	97.89	97.92

表 4 所示的实验结果表明，Word2Vec 的各项实验结果均不理想，虽然它是目前自然语言处理领域使用比较广泛的词向量训练工具，但是其训练窗口小，并且只能针对词本身进行处理，不符合不同语境中具有相同词语差异的实际要求.

表 4　不同算法下实体识别的结果

算法名称	准确率	召回率	调和平均数
Word2Vec	23.12	24.54	23.81
BiLSTM	88.64	87.31	87.97
BiLSTM + CRF	90.45	89.72	90.08
BERT + BiLSTM + CRF	94.58	95.31	94.94

实验中采用了长距离依赖的双向 LSTM 语言模型，即 BiLSTM 模型，不再仅针对某个词本身进行处理，而是综合了长距离上下文得出词语本身的输出最大分值，从而使指标有了较大提升. 但是，由于 BiLSTM 模型的输出序列是根据当前词得分的最大值得出的，实验中容易将不必细分的词继续细分，如“供述自己的犯罪事实”中“犯罪事实”在这里应该当作宾语整体看待，但是 BiLSTM 模型容易将其拆分成“犯罪”和“事实”2 个实体，而在模型中加入 CRF 算法层可以兼顾实体的逻辑性和顺序性，尽量得到全局最优的序列，进一步提升准确性.

BERT 模型层的最大作用在于考虑了同一个词语在不同语境中有不同语义的现实情况，如“它”等代词在不同语境中的指代内容基本都是不同的. 所以从 BERT + BiLSTM + CRF 模型的实验结果可以看出，将 BERT 模型词向量的训练结果导入 BiLSTM + CRF 模型得出的结果更准确，说明 BERT 预训练语言模型在生成词向量时能更准确地反映信息.

综上所述，由于 Word2Vec 训练窗口的限制且只针对词本身处理，导致各项结果均不理想；BiLSTM 模型增大了训练窗口，且尽可能多地结合了上下文，BiLSTM 模型的训练结果提升明显；针对 BiLSTM 模型对某些词组过度拆分的情况，对 CRF 算法做了进一步优化；引入 BERT 模型作为输入层处理，解决了不同语境下同一词语的不同语义和指代问题，使法律案件实体识别的效果得到进一步提升.

5　结束语

针对国内中文法律文本的特点提出了一种以 BERT 模型作为预训练的 BiLSTM + CRF 法律案件实体识别方法. 利用 BERT 模型能最大程度地解析出词在上下文中最准确的语义或者指代；使用 CRF 算法在特定的条件下，得出更加准确的标注序列. 针对国内中文法律文本的特点，使用 CRF 算法对法律智能实体识别中存在的现实逻辑进行约束，进一步提高了法律文本实体识别的成功率.

BERT + BiLSTM + CRF 模型算法的准确率、召回率、调和平均数这 3 个参数值均在 95% 左右，达到了较高水平. 笔者目前只对法律文本中的地点和人物进行了相应的标注，后续将进一步改进模型方法，对法律文本的其他要素，如时间、结果等进行标注，同时提高标注的准确率，为人工智能技术在司法领域中的应用提供方案.

参考文献：

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. MountainView:[s. n.], 2017: 5998-6008.

[2] Grishman R, Sundheim B. Message understanding conference-6: a brief history[C]// Proceedings of the 16th Conference on Computational Linguistics. Copenhagen: Association for Computational Linguistics, 1996: 466-471.

[3] Bikel D M, Schwartz R, Weischedel R M. An algorithm that learns what's in a name[J]. Machine Learning, 1999, 34(1/2/3): 211-231.

[4] Sekine S, Grishman R, Shinnou H. A decision tree method for finding and classifying names in Japanese texts[J]. Proceeding of the 6th Workshop on Very Large Corpora, 1998(5): 171-178.

[5] Borthwick A. A maximum entropy approach to named entity recognition[J]. Thesis New York University, 1999, 36(1): 4701-4708.

[6] McCallum A, Li Wei. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. Edmonton: Association for Computational Linguistics, 2003: 188-191.

[7] Zhang Yue, Yang Jie. Chinese NER using lattice LSTM[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne: Association for Computational Linguistics, 2018: 1554-1564.

[8] Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language?[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 3651-3657.

[9] Peinelt N, Nguyen D, Liakata M. tBERT: topic models and BERT joining forces for semantic similarity detection[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 7047-7055.

[10] Chai Duo, Wu Wei, Han Qinghong, et al. Description based text classification with reinforcement learning[C]// PMLR. Austria Vienna: [s. n.], 2020: 1371-1382.

[11] Qu Chen, Yang Liu, Qiu Minghui, et al. BERT with history answer embedding for conversational question answering[C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Paris: ACM, 2019: 1133-1136.

[12] Dai Zhuyun, Callan J. Deeper text understanding for IR with contextual neural language modeling[C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Paris: ACM, 2019: 985-988.

[13] 赵平, 孙连英, 万莹, 等. 基于 BERT + BiLSTM + CRF 的中文景点命名实体识别[J]. 计算机系统应用, 2020, 29(6): 169-174.
Zhao Ping, Sun Lianying, Wan Ying, et al. Chinese scenic spot named entity recognition based on BERT + BiLSTM + CRF[J]. Computer Systems & Applications, 2020, 29(6): 169-174.

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

[15] 张玉帅, 赵欢, 李博. 基于 BERT 和 BiLSTM 的语义槽填充[J]. 计算机科学, 2021, 48(1): 247-252.
Zhang Yushuai, Zhao Huan, Li Bo. Semantic slot filling based on BERT and BiLSTM[J]. Computer Science, 2021, 48(1): 247-252.

[16] 李丽双, 郭元凯. 基于 CNN-BLSTM-CRF 模型的生物医学命名实体识别[J]. 中文信息学报, 2018, 32(1): 116-122.
Li Lishuang, Guo Yuankai. Biomedical named entity recognition with CNN-BLSTM-CRF[J]. Journal of Chinese Information Processing, 2018, 32(1): 116-122.