计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(4):242−247 [doi: 10.15888/j.cnki.csa.007364]

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

# 注意力机制的 BiLSTM 模型在招聘信息分类中的应用①

吕飞亚[1], 张英俊[1], 潘理虎[1,2]

[1](太原科技大学 计算机科学与技术学院, 太原 030024)
[2](中国科学院 地理科学与资源研究所, 北京 100101)
通讯作者: 张英俊, E-mail: 243418220@qq.com

**摘　要**: 目前 IT 招聘信息分类中传统算法存在长距离依赖, 且无法突出 IT 岗位关键词对文本分类特征影响等问题. 本文通过训练双向长短期记忆网络 BiLSTM 与注意力机制相结合的多层文本分类模型, 将其应用到招聘信息分类中. 该模型包括 One-hot 词向量输入层、BiLSTM 层、注意力机制层和输出层. 其中 One-hot 层构建招聘词典, 节省了大量训练词向量时间, BiLSTM 层可获取更多上下文不同距离的语义信息, 注意力机制层对经过 BiLSTM 层编码数据进行加权转变可提升序列化学习任务. 实验表明: 基于该模型的 IT 招聘信息分类准确率达到 93.36%, 与其他模型对比, 提高约 2%. 该模型更有针对性的分析不同岗位对就业者能力的要求, 实现了不同岗位招聘信息的分类, 对高校毕业生就业指导有重要意义.

**关键词**: 招聘信息; 文本分类; One-hot; BiLSTM 模型; 注意力机制

引用格式: 吕飞亚,张英俊,潘理虎.注意力机制的 BiLSTM 模型在招聘信息分类中的应用.计算机系统应用,2020,29(4):242–247. http://www.c-s-a.org.cn/1003-3254/7364.html

## BiLSTM Model of Attention Mechanism Application in Recruitment Information Classification

LYU Fei-Ya[1], ZHANG Ying-Jun[1], PAN Li-Hu[1,2]

[1](School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China)
[2](Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China)

**Abstract**: At present, traditional algorithms in IT recruitment information classification have long-distance dependence, and cannot highlight the impact of IT job keywords on text classification features. In this study, a multi-layer text classification model combining two-way long-term and short-term memory network BiLSTM and attention mechanism is applied to the classification of recruitment information. The model includes the one-hot word vector input layer, BiLSTM layer, attention mechanism layer, and output layer. One-hot layer builds a recruitment dictionary, which saves a lot of training word vector time; the BiLSTM layer can obtain more semantic information of different distances in the context; and the attention mechanism layer transforms the weights of the data encoded by BiLSTM enhancing the serialization learning task. The results show that the classification accuracy of IT recruitment information based on this model reaches 93.36%, which is about 2% higher than other models. The model analyzes the requirements of different positions on the ability of the employed in a more targeted manner, and realizes the classification of recruitment information in different positions, which is of great significance to the employment guidance of college graduates.

**Key words**: recruitment information; text classification; One-hot; BiLSTM; attention

① 基金项目: 山西省应用基础研究项目 (201801D221179); 山西省中科院科技合作项目 (20141101001); “十二五”山西省科技重大专项 (20121101001); 山西省社会发展科技攻关项目 (20140313020-1)

Foundation item: : Applied Basic Research Project of Shanxi Province (201801D221179); Science and Technology Collaborative Program Between Shanxi Province and Chinese Academy of Sciences (20141101001); Science and Technology Major Program during “Twelfth Five-Year Plan” of Shanxi Province (20121101001); Science and Technology Program for Social Development of Shanxi Province (20140313020-1)

收稿时间: 2019-07-26; 修改时间: 2019-09-03, 2019-09-30; 采用时间: 2019-10-24; csa 在线出版时间: 2020-04-05

## 引言

在大量招聘信息中获取到不同行业、不同岗位对应聘者的能力需求是现代工作中必不可少的, 其重要性不言而喻. 随着全球信息化的高速发展, 各大网站、公众号发布的招聘信息数据量越来越大, 在这些招聘信息中快速按行业和岗位进行准确分类, 才能更有针对性的为大学生就业提供指导.

中文的招聘信息具有非结构化特点[1], 计算机无法对其直接进行处理, 需对文本数据进行向量化表示和特征提取. 继图像处理之后深度学习在自然语言处理领域也取得了很好的效果, 深度学习可以更深层次的表达文本信息, 无需先验知识, 在训练过程中容纳海量数据还集特征提取和性能评价于一体, 有极大优越性. 其中循环神经网络 (Recursive Neural Network, RNN) 被广泛运用于基于时间序列的分类任务.

一般情况下文本分类分为基于统计学的传统文本分类、基于机器学习的文本分类方法以及基于深度学习的文本分类方法. 基于统计学的传统文本分类方法中, 首先是对特征词的预设置, 通过分析文中特征词出现的频率来确定文本的类别归属, 这种方法需要耗费大量的人力资源, 此外由于自然语言的灵活性较高, 其准确性也难以保证; 基于机器学习的文本分类方法[2], 常采用朴素贝叶斯 (NB)、最大熵 (ME)、支持向量积 (SVM) 等, 都是基于关键词设置和词频统计, 忽略了词语之间的关联和文本前后语义信息.

近年来, 深度学习算法被应用到了自然语言处理领域, 利用神经网络训练词向量来表示文本, 有效的避免了数据稀疏性问题, 同时还可以获取语义信息. 采用深度学习模型如递归神经网络[3]、卷积神经网络[4] (Convolutional Neural Network, CNN) 和循环神经网络 (RNN) 等进行文本分类, 获得比传统机器学习更好的效果. 2010 年 Mikolov等人对递归神经网络进行改进并在语言模型表征能力上取得巨大进步[5], 然而递归神经网络的输入是树/图结构, 这种结构需要花费很多人工去标注; 顾静航等人[6]提出基于卷积神经网络的实体关系提取, 证明卷积神经网络擅长于捕捉局部序列关系, 但不能很好的解决长距离依赖问题. 针对需要前后文语境的分类任务, 循环神经网络 RNN 具有先天优势, 如 Liu 等人[7]使用 RNN 模型进行多标签的文本分类. 为解决 RNN 在文本分类过程中出现的梯度消失和梯度爆炸问题, Hochreiter 和 Schmidhuber 等人[8]对 RNN 模型改进并提出长短期记忆模型 (Long Short Term Memory, LSTM), 其广泛应用于文本分类[9]和信息抽取[10]等任务, 取得了更好的效果.

本文主要针对 IT 行业的招聘信息按岗位进行分类, 招聘信息进行分类时对上下文有较强的依赖性, 并且某些关键词对分类又有较大的影响. 因此, 本文训练基于循环神经网络改进的双向长短期记忆网络 (Bi-directional Long Short Term Memory, BiLSTM) 进行文本分类模型, 构建招聘词典并用 One-hot 向量表示, 引入注意力机制, 加强岗位关键词的比重. 在各个招聘网站获取的 IT 行业的招聘数据进行实验, 最终招聘信息分类效果在准确率、召回率等其他指标得到明显的优化.

## 1 基于注意力机制的 BiLSTM 模型构建方法

基于注意力机制的 BiLSTM 模型主要由以下 5 部分组成, 结构如图 1 所示.

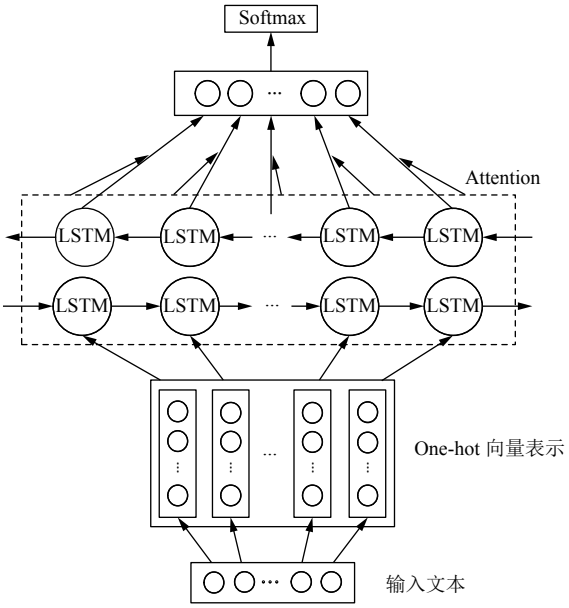


图 1 基于注意力机制的 BiLSTM 模型结构图

(1) 将文本做分词、去停用词, 并且设置自定义分词词典, 加入招聘信息不同类别中的特征词汇, 如“后端开发”, “界面设计”, 等具有强烈描述不同招聘类别的词汇;

(2) One-hot 向量表示, 把分类文本转换成向量;

(3) 训练 BiLSTM 模型获取招聘信息文本特征;

(4) 引入注意力机制突出招聘类别重点词的权重;

(5) 最后输入分类器进行分类, 得到分类结果.

### 1.1 向量表示

招聘信息将按岗位需求分为软件开发、测试、运维、UI 和其他岗位 5 大类, 将其修改为统一的格式. 进行简单的预处理, 去掉多余的符号和无法识别的字符. 将数据分为训练集、测试集、验证集 3 部分. 本文采用 One-hot 表示文本, 简化步骤并缩短训练时间.

由于 IT 行业招聘信息内容在一个范围内, 不会像新闻类数据产生巨大的词典, 于是采用 One-hot 方法进行文本表示, 训练数据生成 IT 招聘能力要求的词汇表, 保留字符级信息. 具体步骤为:

(1) 选择招聘信息中出现频率最高的 5000 分类关键词生成词汇表, 来获取字符级信息;

(2) 每个文本含有词汇表中的字的部分, 对应词汇表生成 ID 表;

(3) 截取序列长度为 100, 将 ID 表中的 ID 生成 One-hot 向量. 将文本矩阵与标签矩阵对应, 生成输入矩阵. 图 2 为一个句子按照词汇表生成 One-hot 矩阵的过程.

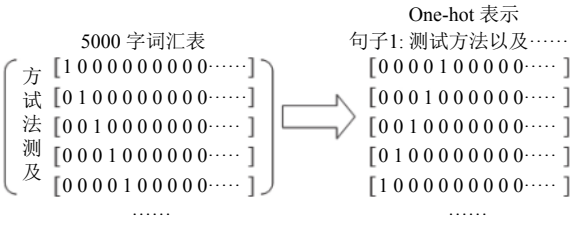


图 2 训练集 One-hot 表示示意图

### 1.2 BiLSTM 模型

循环神经网络 RNN 擅长于捕捉长依赖序列关系, 神经元的某些输出可作为输入再次传输到神经元中, 能够有效的利用之前的信息. 但是在训练过程中激活函数导数的不断累乘, 会导致“梯度消失”和“梯度爆炸”问题. 长短期记忆网络 LSTM 巧妙的引入“门”机制, 很好的解决了这一问题. 每一个“门”结构中都包含一个 Sigmoid 网络神经层和一个 pointwise 乘法操作, 来控制信息是否可以通过, 从而去除或者增强信息到细胞状态. LSTM 是由一系列重复时序模块组成, 每个模块包含三个“门”和一个记忆单元 (memory cell), 分别是遗忘门 (forget gate)、输入门 (input gate)、输出门 (output gate). 具体结构如图 3 所示.

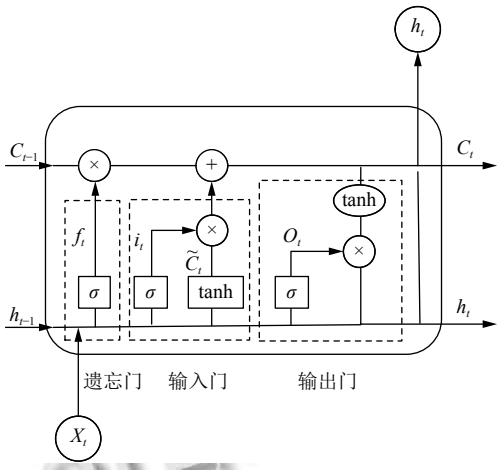


图 3 LSTM 单元结构图

遗忘门决定细胞将丢弃什么信息, 读取 $h_{t-1}$ 和 $x_t$, 输出一个 0 到 1 之间的数值给每一个在细胞状态 $C_{t-1}$ 中.

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

输入门确定什么样的新信息被存放在细胞状态中, 这里包含两个部分. 首先, 一个 Sigmoid 神经网络层决定什么值将要更新, 称“输入门层”. 然后, 一个 tanh 层创建一个新的候选值向量 $C_t$, $C_t$ 会被加入到状态中.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{3}$$

单元信息更新时将旧状态与 $f_t$ 相乘, 丢弃掉无关信息. 加上$i_t * \tilde{C}_t$, 形成新的候选值.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

输出门通过运行一个 Sigmoid 层来确定细胞状态的哪个部分将输出, 接着把细胞状态通过 tanh 函数进行处理, 得到一个在–1 到 1 之间的值, 并将它和 Sigmoid 门的输出相乘, 最终仅输出确定输出的那部分.

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * \tanh(C_t) \tag{6}$$

其中, tanh() 代表激活函数, $\sigma$ 代表 Sigmoid 神经网络层, $x_t$ 为 $t$ 时刻输入的单元状态; $f_t$、$i_t$、$O_t$ 分别表示遗忘门、输入门、输出门的结算结果; $W_f$、$W_i$、$W_o$、$W_c$ 分别代表遗忘门、输入门、输出门和跟新后的权重; $b_f$、$b_i$、$b_o$、$b_c$ 为对应的偏置量.

在文本分类过程中, 为充分利用文本的上下文语境信息, 将使用双向长短期记忆网络 BiLSTM, 即将时序相反的两个 LSTM 模型相结合.

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(h_{t-1}, W_t, c_{t-1}), t \in [1, T] \quad (7)$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(h_{t+1}, W_t, c_{t+1}), t \in [T, 1] \quad (8)$$

$$H_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \quad (9)$$

$H_t$ 为 BiLSTM 模型输出的文本特征向量.

### 1.3 注意力机制

在招聘文本分类任务中, 将 BiLSTM 层每个时刻的输出向量求和取平均值, 得到每个特征词汇的权重, 无法突出对不同岗位分类起到重要作用词汇的重要性, 文本特征向量具有高维稀疏等特点, 特征向量直接求和取平均值对文本分类的准确率有一定程度的影响.

近年来注意力机制被应用于智能问答和文本检索等任务中[11], 都取得了良好的效果. 其应用了生物学的仿生学思想, 模拟人类大脑中的分配机制, 即对待处理的信息中比较关键的信息分配更多的注意力. 本文采用注意力机制对招聘信息进行处理, 学习其句子表示, 计算过程如式 (10) 和式 (11) 所示:

$$u_t = \tanh(W_w H_t + b_w) \quad (10)$$

$$a_t = \frac{\exp(u_t^T u_w)}{\sum_t \exp(u_t^T u_w)} \quad (11)$$

其中, $u_t$、$W_w$、$b_w$ 为注意力机制层参数, $a_t$ 为第 $t$ 个输入的特征词对区分文本类别贡献程度的权重. 从而得到新的输出特征值 $v$ 为:

$$v = \sum_{i=1} a_t H_t \quad (12)$$

## 2 实验结果及分析

### 2.1 数据来源与处理

本文自建数据集, 为了保证数据的多样性, 在各个招聘网站上采集 IT 行业招聘的数据信息, 经数据预处理获取招聘信息语料库. 该数据集共有招聘数据 60 000 条, 根据岗位需求分为软件开发、测试、运维、UI 和其他岗位 5 大类, 每一类 12 000 条文本, 取其中 10 000 条作为训练集, 另外 1000 条作为测试集和 1000 条验证集. 本文选取序列长度为 100.

### 2.2 参数设置

初始学习率为 0.001, 批处理文件数为 256, 正向和反向的 LSTM 隐藏单元数均为 512 层, 训练轮数为 1000. 模型的激活函数使用 ReLU, 采用 Softmax 函数作为分类器, 优化函数使用 AdamOptimizer. 针对模型训练过程中可能出现的过拟合现象, 利用 dropout 和 L2 正则化方法对网络参数进行约束.

### 2.3 对比实验

为了验证基于注意力机制的 BiLSTM 模型在招聘信息分类的有效性, 本文设置两个对比实验分析比较不同算法下特定类别招聘信息分类效果和不同算法、不同数据量对招聘信息准确率的影响.

本文实验结果分析采用正确率 (*precision*)、召回率 (*recall*) 和 *F* 值 (*F_score*)3 个标准作为模型性能的评价指标. 准确率即分类模型正确预测得样本数在总样本中所占的比例; 召回率又称为查全率, 体现系统分类结果的完备性; 实验结果希望准确率和召回率都是越高越好, 但是两者有一定矛盾性, 所以 *F* 值对准确率和召回率进行加权调和平均, 较为全面地评价一个分类器. 具体计算方法如下:

$$precision = \frac{out\_cor}{out\_all} \quad (13)$$

$$recall = \frac{out\_cor}{this\_all} \quad (14)$$

$$F\_score = \frac{2 \times precision \times recall}{precision + recall} \quad (15)$$

其中, $out\_cor$ 表示输出的判断正确的文本个数; *out_all* 表示输出的所有文本个数; *this_all* 表示测试集中的所有该文本的个数.

#### 2.3.1 在不同模型下特定类别分类效果的比较

本实验部分采用招聘信息语料库, 比较本文提出的基于注意力机制下的双向长短期记忆网络 (Attention-BiLSTM) 与注意力机制下的长短期记忆网络 (Attention-LSTM)、长短期记忆网络 (LSTM)、CNN 和 FastText 模型的准确率、召回率和 *F* 值, 验证基于注意力机制的 BiLSTM 模型的稳定性和有效性.

由于 Attention-BiLSTM 利用双向的 LSTM 通过对词向量的计算得到更高级别的句子向量, 使得模型对文本信息的理解更准确、更完整. 其中运维岗位和测试岗位的在不同算法下的对比实验如表 1 和表 2 所示. 实验结果表明本文提出的模型在 3 个实验指标上均要高于其他模型 1%~2%.

#### 2.3.2 不同数据量对模型准确率的影响

该实验分别在本文提出的基于注意力机制下的 BiLSTM 模型与 Attention-LSTM、LSTM 三组模型进

行对比实验, 并且设置对比实验的超参数和文本模型参数相同.

表 1 运维岗位对比实验结果

| 模型 | 准确率 (%) | 召回率 (%) | *F* 值 |
|---|---|---|---|
| Attention-BiLSTM | 94.73 | 94.96 | 94.84 |
| Attention-LSTM | 93.58 | 93.46 | 93.52 |
| LSTM | 91.86 | 92.53 | 92.19 |
| CNN | 92.51 | 92.96 | 92.73 |
| FastText | 92.43 | 92.74 | 92.58 |

表 2 测试岗位对比实验结果

| 模型 | 准确率 (%) | 召回率 (%) | *F* 值 |
|---|---|---|---|
| Attention-BiLSTM | 93.36 | 93.85 | 93.60 |
| Attention-LSTM | 91.78 | 92.84 | 92.31 |
| LSTM | 90.45 | 91.23 | 90.84 |
| CNN | 91.23 | 91.68 | 91.45 |
| FastText | 91.21 | 91.47 | 91.34 |

使用招聘信息语料库训练数据 50 000 条, 分别取数据集的 10%、20% 至 100% 共 10 组不同百分比的数据量进行训练. 数据量为 10%(5000 条)、50% (25 000) 和全部 (50 000) 时各分类模型准确率变化如图 4、图 5、图 6 所示.

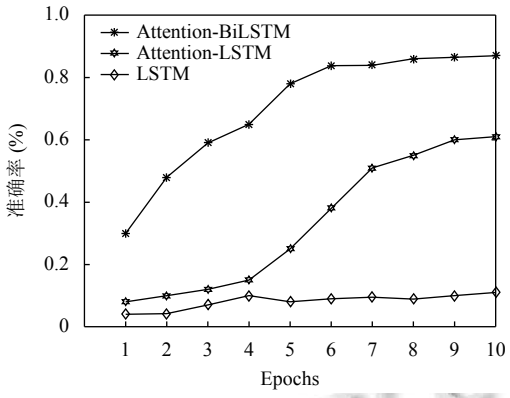


图 4 数据量为 5000 训练结果

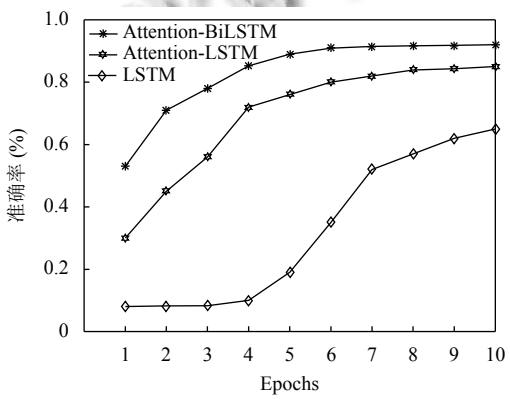


图 5 数据量为 25 000 训练结果

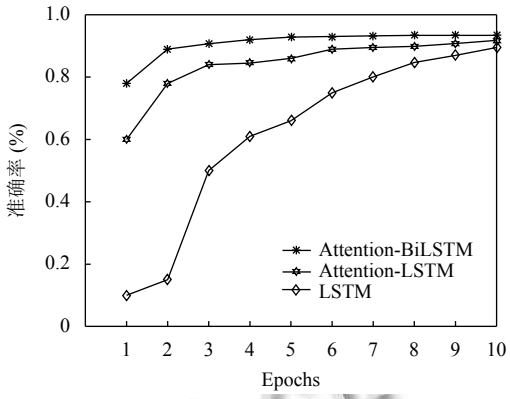


图 6 数据量为 50 000 训练结果

由图 4 可以看出, 招聘信息数据量较少时, 本文提出的模型整体准确率明显高于另外两组, Epoch 到 6 时准确率可以达到 84% 以上. 并趋于稳定, 模型 (Attention-LSTM) 变化幅度较大, Epoch 到 10 时准确率达到 60%, 由于模型 (LSTM) 虽然实现了文本序列化, 但是单向的 LSTM 缺乏上文语义信息, 当训练数据量较少时, 文本向量特征高维稀疏, 模型学习能力差, 导致模型 (LSTM) 分类效果较差.

从图 5、图 6 看出, 招聘信息数据量增多时, 模型 (Attention-LSTM) 融入注意力机制, 有效提升了模型学习的能力并且学习到不同距离的上下文语义依赖关系, 导致模型 (Attention-BiLSTM) 可以最快达到稳定并且有 93.36% 准确率, 模型 (Attention-LSTM) 达到稳定性的速度和准确率略次于模型 (Attention-BiLSTM), 模型 (LSTM) 在数据量达到 50 000 时, 随着 Epoch 改变分类性能才有明显提升.

## 3 结论与展望

本文采用基于注意力机制的 Bi-LSTM 多层文本分类模型, 并构建了招聘词典, 使用 One-hot 方法进行向量表示, 有效的解决了招聘信息分类中准确率低、无法突出分类关键词等问题. 通过与现有几种文本分类算法比较, 本文的模型在准确率、召回率和 *F* 值上均有明显提高. 实验证明, 将基于注意力机制的 Bi-LSTM 文本分类模型应用于招聘信息分类具有可行性和有效性.

针对 Ont-hot 对于招聘类的专业名词的深层语义辨析上存在一定的缺陷, 本文下一步将继续查阅相关领域文献, 改进 One-hot 语义方面的表示[12], 提高文本分类准确率.

## 参考文献


1 郑阶财. 非结构化数据的相关问题研究[博士学位论文]. 济南: 山东大学, 2017.

2 杜昌顺, 黄磊. 分段卷积神经网络在文本情感分析中的应用. 计算机工程与科学, 2017, 39(1): 173–179. [doi: 10.3969/j.issn.1007-130X.2017.01.024]

3 Socher R, Huval B, Manning CD, *et al*. Semantic compositionality through recursive matrix-vector spaces. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island. 2012. 1201–1211.

4 王吉俐, 彭敦陆, 陈章, 等. AM-CNN: 一种基于注意力的卷积神经网络文本分类模型. 小型微型计算机系统, 2019, 40(4): 710–714. [doi: 10.3969/j.issn.1000-1220.2019.04.004]

5 Mikolov T, Karafiát M, Burget L, *et al*. Recurrent neural network based language model. Proceedings of the 11th Annual Conference of the International Speech Communication Association. Makuhari, Japan. 2010. 10451048.

6 顾静航. 面向生物医学领域的实体关系抽取研究[博士学位论文]. 苏州: 苏州大学, 2017.

7 Liu PF, Qiu XP, Huang XJ. Recurrent neural network for text classification with multi-task learning. arXiv: 1605.05101, 2016.

8 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

9 赵淑芳, 董小雨. 基于改进的 LSTM 深度神经网络语音识别研究. 郑州大学学报 (工学版), 2018, 39(5): 63–67.

10 王竣平, 白宇, 蔡东风. 采用 BI-LSTM-CRF 模型的数值信息抽取. 计算机应用与软件, 2019, 36(5): 138–144. [doi: 10.3969/j.issn.1000-386x.2019.05.025]

11 谢金宝, 侯永进, 康守强, 等. 基于语义理解注意力神经网络的多元特征融合中文文本分类. 电子与信息学报, 2018, 40(5): 1258–1265. [doi: 10.11999/JEIT170815]

12 梁杰, 陈嘉豪, 张雪芹, 等. 基于独热编码和卷积神经网络的异常检测. 清华大学学报 (自然科学版), 2019, 59(7): 523–529.