计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(10):205−210 [doi: 10.15888/j.cnki.csa.007619]

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

基于 BLSTM-Attention 神经网络模型的化工事故分类①

葛　艳, 郑利杰, 杜军威, 陈　卓

(青岛科技大学 信息科学技术学院, 青岛 266061)
通讯作者: 郑利杰, E-mail: 794011210@qq.com

摘　要: 化工事故新闻数据包含新闻内容, 标题以及新闻来源等方面信息, 新闻内容的文本对上下文具有较强的依赖性. 为了更准确地提取文本特征并提高化工事故分类的准确性, 该文提出了一种基于 Attention 机制的双向 LSTM (BLSTM-Attention) 神经网络模型对化工新闻文本进行特征提取并实现文本分类. BLSTM-Attention 神经网络模型能够结合文本上下文语义信息, 通过正向和反向的角度来提取事故新闻的文本特征; 考虑到事故新闻中不同词对文本的贡献不大相同, 加入 Attention 机制对不同词和句子分配不同权重. 最后, 将该文提出的分类方法与 Naive-Bayes、CNN、RNN、BLSTM 分类方法在相同的化工事故新闻数据集上进行实验对比. 实验结果表明: 该文提出的神经网络模型 BLSTM-Attention 神在化工数据集上的效果更优于其他分类方法模型.

关键词: 化工事故新闻; 特征提取; BLSTM-Attention; 文本分类

引用格式: 葛艳,郑利杰,杜军威,陈卓.基于 BLSTM-Attention 神经网络模型的化工事故分类.计算机系统应用,2020,29(10):205–210. http://www.c-s-a.org.cn/1003-3254/7619.html

Chemical Accident Classification Based on BLSTM-Attention Neural Network Model

GE Yan, ZHENG Li-Jie, DU Jun-Wei, CHEN Zhuo

(College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: Chemical accident news data contains information such as news content, titles, and news sources. The text of news content is highly dependent on the context. In order to extract text features more accurately and improve the accuracy of chemical accident classification, this study proposes a Bidirectional LSTM (BLSTM-Attention) neural network model based on Attention mechanism to extract features of chemical news texts and realize text classification. The BLSTM-Attention neural network model can combine text context semantic information to extract text features of accident news through forward and reverse angles. Considering that different words have different contributions to the text in the accident news, the Attention mechanism is added to assign different weights to different words and sentences. Finally, the proposed classification method is compared with Naive-Bayes, CNN, RNN, BLSTM classification method on the same chemical accident news data set. Experimental results show that the BLSTM-Attention neural network model proposed in this study is better than other classification models in chemical data set.

Key words: news of chemical accidents; feature extraction; BLSTM-Attention; text classification

① 基金项目: 国家自然科学基金 (61973180, 61273180); 山东省重点研发计划 (2018GGX101052); 山东省自然科学基金 (ZR2019MF033)
Foundation item: National Natural Science Foundation of China (61973180, 61273180); Key Research and Development Program of Shandong Province (2018GGX101052); Natural Science Foundation of Shandong Province (ZR2019MF033)
收稿时间: 2020-02-28; 修改时间: 2020-03-17; 采用时间: 2020-04-03; csa 在线出版时间: 2020-09-30

随着化工领域的迅速发展, 由于化工生产品中易燃易爆、有腐蚀性、有毒的物质比较多, 化工事故也在频频发生[1]. 因此, 有必要对发生的化工事故进行分析, 知道是哪些行为、哪些物品或者哪些事件造成了化工事故的发生. 这对化工事故的监管, 预警以及处理等方面有着重要意义.

文本自动分类是信息处理的重要研究方向[2]. 目前分类方法有很多, 常用的有朴素贝叶斯 Naive-Bayes、卷积神经网络 CNN、循环神经网络 RNN 等分类方法.

罗慧钦针对朴素贝叶斯模型属性间条件独立假设不完全符合实际的问题, 提出只考虑属性间的依赖关系的基于隐朴素贝叶斯模型的商品情感分类方法[3]. 但朴素贝叶斯模型无法根据文本上下文的语义信息做出有效的特征提取.

CNN 在很多自然语言处理任务中得到较好的结果. Kim Y 等人用 CNN 完成文本分类任务, 具有很少的超参数调整和静态向量, 在多个基准测试中获得较好的结果[4]. Zhang X 等人使用字符级卷积网络 (ConvNets) 进行文本分类, 并取得了较好的结果[5]. 但由于 CNN 模型需要固定卷积核窗口的大小, 导致其无法建立更长的序列信息[6].

Auli 等人提出了基于循环神经网络的联合语言和翻译模型[7]. 邵良杉等人基于 LSTM改进的 RNN 模型实现互联网中在线评论文本的情感倾向分类任务. RNN 能够解决人工神经网络无法解决的文本序列前后关联问题, 但无法解决长时依赖问题[8], 并且存在梯度消失问题, 对上下文的处理受到限制[9].

长短期记忆网络 (Long Short-Term Memory, LSTM)[10] 是一种时间递归神经网络[3], 也是一种特定形式的 RNN (Recurrent Neural Network, 循环神经网络). 可以解决 RNN 无法处理的长距离依赖问题, 但容易造成前面的信息缺失以及不同时刻信息的重要度得不到体现[11]. 万胜贤等人提出局部双向 LSTM 模型, 对高效的提取一些局部文本特征有重要意义[12]. LSTM 和双向 LSTM 是典型的序列模型, 能够学习词语之间的依赖信息但不能区分不同词语对文本分类任务的贡献程度[13], 有效的表达文本语义特征[14].

Attention 机制[15] 的一大优点就是方便分析每个输入对结果的影响, 可以自动关注对分类具有决定性影响的词, 捕获句子中最重要的语义信息, 而无需使用额外的知识和 NLP 系统[16]. 另外, Attention 机制[15] 在关系抽取任务上也有相关的应用[17].

针对以上分析, 本文提出一种基于 Attention 机制的 BLSTM (BLSTM-Attention) 神经网络模型对化工新闻文本进行特征提取, 实现事故文本分类. 模型的 BLSTM 层实现对文本上下文语义信息的高级特征提取; 在 BLSTM 层后引用 Attention 机制实现对不同词和句子分配不同权重, 合并成更高级的特征向量.

化工事故新闻数据包含新闻内容, 标题信息以及新闻来源等多个方面的信息, 其中, 新闻内容的文本对上下文具有较强的依赖性. 另外, 事故新闻数据集有其独有的一些特征, 有些分类算法在化工领域数据集上并不完全适用. 本文选取清洗完成的化工事故新闻标题、内容作为分类的依据, 新闻来源作为新闻的可靠性分析依据. 将本文的模型分类方法与 Naive-Bayes、CNN、RNN、BLSTM 方法在相同的化工事故新闻数据集上进行实验对比. 实验结果表明, 本文提出的 BLSTM-Attention 神经网络模型在化工事故新闻数据集上的效果要更优于其他方法模型.

1 BLSTM-Attention 神经网络模型

本文将 Attention 机制与双向 LSTM 神经网络相结合, 构成 BLSTM-Attention 模型, 该模型由 4 个部分组成: (1) 输入层: 输入训练好的词向量; (2) BLSTM 层: 根据输入层的信息获取高级特征; (3) Attention 层: 产生权重向量, 并通过乘以权重向量, 将每个时间步长的词级特征合并为句子级特征向量; (4) 输出层: 根据从 Attention 层获得的输出信息完成分类任务并输出结果. 模型结构如图 1 所示.

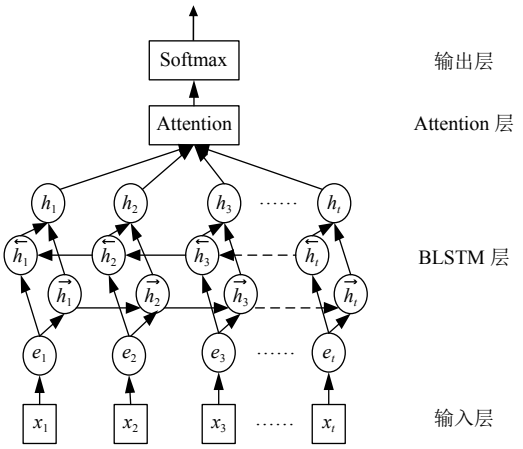

图 1 BLSTM-Attention 神经网络模型

1.1 输入层

给定一个由 t 个句子组成的训练文本 d, 即 $d=\{S_1,S_2,\cdots,S_t\}$, 每个句子 S 又由 n 个词语组成, 则 $S_i=\{w_1,w_2,\cdots,w_t\}$. 本文利用 Word2Vec 对数据进行向量化, 得到向量化表示的词语 x_t 作为 BLSTM 层的输入, 称为词向量 $w\in\Re^d$, 通过输入层实现文本向量化.

1.2 BLSTM 层

LSTM 网络信息的更新和保留是由输入门、遗忘门、输出门和一个 cell 单元来实现的. 长短时记忆网络的基本思想是在原始的 RNN 隐藏层只有一个对短期输入非常敏感的状态, 即隐藏层 h 基础上再增加一个状态单元 c (cell state) 来保存长期状态.

输入门 (input gate) 决定了当前时刻网络的输入 x_t 有多少保存到单元状态 c_t, 可以避免当前无关紧要的内容进入记忆. 表示为:

$$i_t=\sigma(W_{xi}x_t+W_{hi}h_{t-1}+W_{ci}c_{t-1}+b_i) \tag{1}$$

其中, W_{xi}, W_{hi}, W_{ci} 表示输入门 i_t 所对应的权重矩阵; x_t 表示输入的词向量; h_{t-1} 表示 LSTM 层上一时刻的输出结果; c_{t-1} 表示上一时刻的状态; b_i 表示一个常数向量.

遗忘门 (forget gate) 决定了上一时刻的单元状态 c_{t-1} 有多少保留到当前时刻 c_t, 可以保存很久很久之前的信息. 表示为:

$$f_t=\sigma\left(W_{xf}x_t+W_{hf}h_{t-1}+W_{cf}c_{t-1}+b_f\right) \tag{2}$$

其中, W_{xf}, W_{hf}, W_{cf} 表示遗忘门 f_t 所对应的权重矩阵; b_f 表示一个常数向量.

输出门 (output gate) 控制单元状态 c_t 有多少输出到 LSTM 的当前输出值 h_t, 可以控制长期记忆对当前输出的影响. 表示为:

$$o_t=\sigma(W_{xo}x_t+W_{ho}h_{t-1}+W_{co}c_{t-1}+b_o) \tag{3}$$

其中, W_{xo}, W_{ho}, W_{co} 表示遗忘门 o_t 所对应的权重矩阵; b_o 表示一个常数向量.

当前时刻状态单元的状态值由 c_t 来表示:

$$c_t=\sigma(i_tc_m+f_tc_{t-1}) \tag{4}$$

其中, c_m 表示候选状态单元值, 公式表示为:

$$c_m=\tanh(W_{xc}x_t+W_{hc}h_{t-1}+W_{cc}c_{t-1}+b_c) \tag{5}$$

其中, W_{xc}, W_{hc}, W_{cc} 表示选状态单元值 c_m 所对应的权重矩阵.

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

本文采用的 BLSTM 神经网络包含了两个隐藏层, 这两个隐藏层之间的链接是以相反的时间顺序流动的, 所以它是分别按前向与后向传递的. 自前向后循环神经网络层的公式表示为:

$$\overrightarrow{h_t}=\overrightarrow{LSTM}\left(w_i,\overrightarrow{h_{t-1}}\right) \tag{7}$$

自后向前循环神经网络层的公式表示为:

$$\overleftarrow{h_t}=\overleftarrow{LSTM}\left(w_i,\overleftarrow{h_{t-1}}\right) \tag{8}$$

BLSTM 层叠加后输入隐藏层公式表示为:

$$y_t=W_{\overrightarrow{hy}}\overrightarrow{h_t}+W_{\overleftarrow{hy}}\overleftarrow{h_t}+b_y \tag{9}$$

1.3 Attention 层

考虑到不同词对文本的贡献不大相同, 本文采用能够通过分配不同的注意力来获得较高的权重的 Attention 机制来对重要词语和句子进行特征提取. Attention 机制结合 BLSTM 模型将利用每个时刻下状态和最终状态, 计算得到每个时刻状态的注意力概率分配, 以此来对最终状态进行更进一步的优化, 得到最终的文本特征, 并将 Attention 机制接入全连接进行分类. 首先对 BLSTM 层的输出信息 y_t 通过非线性变换得到隐含表示 u_{it}; 然后, 经过随机初始化注意力机制矩阵 u_w 与 u_{it} 进行点乘运算得到 Attention 层的词级别权重系数 α_{it}, 并最终得到句子特征向量 s_{it}, 表示为:

$$u_{it}=\tanh(W_wh_{it}+b_w) \tag{10}$$

$$\alpha_{it}=\frac{\exp\left(u_{it}^{\mathrm{T}}u_w\right)}{\sum_t\exp\left(u_{it}^{\mathrm{T}}u_w\right)} \tag{11}$$

$$s_{it}=\sum_t\alpha_{it}h_{it} \tag{12}$$

其中, W_w 为权重矩阵, b_w 为偏置量.

同样, 采用与词级别相同的方式对文章贡献不同的句子赋予不同的权重参数, 通过句子级的 Attention 机制得到文章的特征向量. 具体表示为:

$$u_t=\tanh(W_wh_t+b_w) \tag{13}$$

$$\alpha_t=\frac{\exp\left(u_t^{\mathrm{T}}u_w\right)}{\sum_t\exp\left(u_t^{\mathrm{T}}u_w\right)} \tag{14}$$

$$v_t=\sum_t\alpha_th_t \tag{15}$$

1.4 输出层

本文采用计算简单, 效果显著的 Softmax 分类器

对经过 Attention 机制得到的文章特征向量v_t进行归一化得到预测分类. 表示为:

$$y = Softmax(W_v v_t + b_v) \tag{16}$$

另外, 本文采用正则化的方法来提高模型的泛化能力, 防止出现过拟合的情况. 目标函数用带有 L2 正则化的类别标签 y 的负对数似然函数, 表示为:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}(t_i \log(y_i)) + \lambda \|\theta\|_F^2 \tag{17}$$

其中, t_i是独热编码 (one-hot represented) 表示的基础值; $y_i \in \Re^m$是每个类的估计概率, m 是类别的个数; λ 是 L2 正则化超参数.

2 实验

2.1 实验数据

本文采用 Python 程序爬取各大网站得到的 4 万多条事故新闻作为语料库. 语料库中包括事故标题、事故内容以及新闻来源等信息, 本文选取事故标题、事故内容作为事故类型分类的依据. 由于网站信息杂乱, 新闻的重复报道以及与化工事故相关性不大等原因使得评论中存在较多的噪声数据, 为保证实验质量对数据进行了清洗. 通过相似度计算, 打标签等方法对数据进行预处理, 清洗去掉重复以及非化工事故新闻 3 万 4 千多条噪声数据, 选取 10 314 条数据进行实验. 数据统计如图 2 所示.

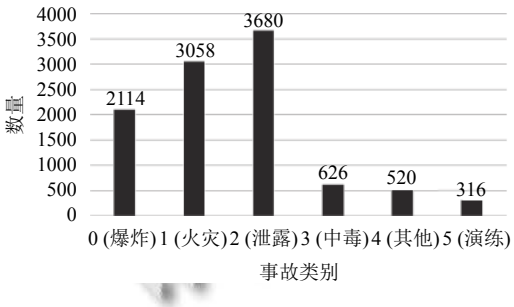

图 2 数据统计

图 2 横坐标表示事故类别, 纵坐标表示每一类别的个数. 化工事故的类别基本分为 5 大类: 爆炸, 火灾, 泄露, 中毒以及其他. 根据化工领域常举行化工事故演练来加强官兵快速反应能力和实战能力, 新增了演练这一类别, 能够训练模型将事故与演练区分开, 防止因为事故演练而做出错误事故的分析. 这样大大提升了我们对事故分析的准确性和有效性, 同时也提高了分类的准确性. 表 1 对数据做了进一步说明.

表 1 数据说明

数据	相关说明	
新闻类别	爆炸, 火灾, 泄露, 中毒、其他事故、演练	
分类依据	新闻内容、新闻标题	
新闻来源类别及得分	省级以上	5
	省级	3
	地市级	2
	其他级别	0

本文将新闻标题以及新闻内容作为模型的输入, 考虑到标题所包含的信息量比较少, 将获得的分类结果按照 2:8 的比重进行计算作为分类的最终结果. 另外, 由于新闻来源杂乱繁多, 通过打标签的方式将其分为 4 类: 省级以上、省级、地市级以及其他类别, 并对其进行打分. 新闻来源的可靠性通过其得分来判断, 化工事故新闻数据文本的得分情况如图 3 所示.

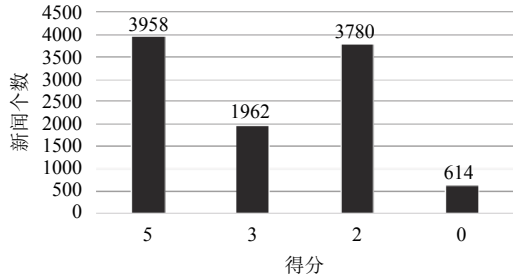

图 3 新闻数据得分

图 3 的横坐标表示新闻的得分, 纵坐标表示获得相应分数的新闻个数. 从图 3 看出, 新闻得分为 0 的新闻数并不是很多, 大多数的新闻都有可靠的来源, 具有较强的可靠性.

2.2 参数设置

本文采用 Adam 优化方法, 其学习率设为 0.001; 经过多次实验后选取结果最优参数设置: 迭代次数设为 20; 词嵌入维度设为 200; 神经元个数为 [256, 128]. 为了防止过拟合现象, 目标函数加入 L2 正则化项, 正则化的因子取值设为 10^{-5}; 另外, 还加入 dropout 策略, 并把它应用在输入层和 BLSTM 层, 经过多轮试验, 当 dropout rate 为 0.5 的时候, 模型能够达到比较好的性能. 根据实验过程中的最佳实验效果选取各个模型的参数, 具体参数设置如表 2 所示.

本文采用 jieba 分词工具对数据做分词处理, Word2Vec 训练数据产生所需要的词向量. 网络模型设置好之后, 不需要借助 GPU 在自己的电脑上就能够实

现, 实现成本低, 运算复杂度并不高. 另外, 本文采用准确率和 $F1$ 值来评估模型效果.

表 2 模型参数设置

参数	Naive-Bayes	CNN	RNN	BLSTM	BLSTM-Attention
迭代次数		10	10	20	20
学习率		le–3	le–3	0.001	0.001
词嵌入维度		64	64	200	200
神经元个数		128	128	[256, 128]	[256, 128]
dropout保留比例		0.5	0.5	0.5	0.5
去除高频词个数	20				
测试集占比(%)	20	20	20	20	20
序列长度		600	600	150	150

2.3 实验结果

图 4、图 5 分别表示 BLSTM 和 BLSTM-Attention 神经网络模型在模型训练时损失值随迭代次数的变化图.

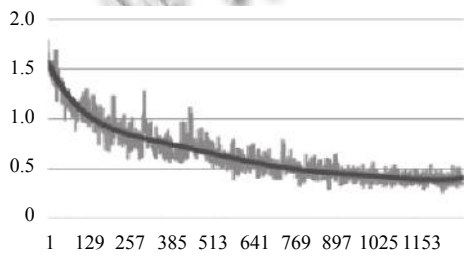

图 4 BLSTM 模型训练损失变化图 (10314)

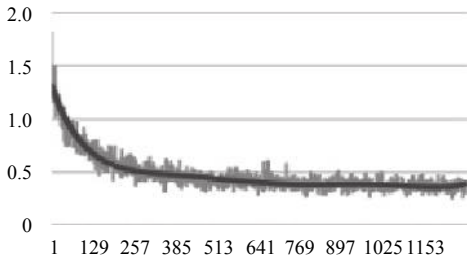

图 5 BLSTM-Attention 模型训练损失变化图 (10314)

从图 4 和图 5 中可以看出函数损失值是逐渐下降的, 并且最终趋于稳定收敛状态. BLSTM-Attention 模型与 BLSTM 模型相比, 起始的损失值相差不大, 但收敛速度明显增加, 也更加稳定, 并且最终收敛的损失值较小.

为了验证 BLSTM-Attention 神经网络模型的有效性, 与 Naive-Bayes、CNN、RNN 以及 BLSTM 方法模型在相同的数据集上做对比实验, 实验结果如表 3 所示.

表 3 不同分类模型的平均分类结果比较

模型(Model)	准确率(Precision)	$F1$ (F_score)
Naive-Bayes	0.8403	0.8392
CNN	0.8810	0.8792
RNN	0.7783	0.7752
BLSTM	0.9023	0.9032
BLSTM-Attention	0.9303	0.9212

2.4 实验分析

从表 3 的各个模型的对比实验结果能够看出, 在相同的数据集上, 除了 RNN 的实验效果较差, 以 Word2Vec 训练的词向量作为文本特征的其他分类实验结果的效果都不错. 能够说明 Word2Vec 训练的词向量可以很好地描述文本特征.

BLSTM 神经网络模型相较于其他分类模型的分类效果更好, 这也说明 BLSTM 模型在学习词语之间的依赖信息和反映文本上下文语义信息上有着重要作用.

BLSTM-Attention 模型与 BLSTM 模型相比, 实验结果表明 Attention 机制对不同词语和句子所分配的权重对文本的特征提取有一定的意义, 提升了文本分类的准确度.

3 结束语

为了解决化工新闻文本的语义特征提取及对上下文的依赖问题, 本文提出一种应用于化工事故领域的基于双向 LSTM-Attention 机制的神经网络模型. BLSTM-Attention 模型能够实现对词语之间以及句子之间的特征学习和提取, 并且通过 Attention 机制对不同的词语和句子分配不同的权重.

本文采用 Word2Vec 对清洗好的数据训练得到词向量; 将 BLSTM-Attention 神经网络模型与 Naive-Bayes、CNN、RNN 以及不加 Attention 机制的 BLSTM 方法模型在相同的化工数据集上做对比实验. 实验结果表明, Word2Vec 训练的词向量可以很好地描述文本特征. 另外, 相较于 Naive-Bayes、CNN、RNN 以及不加 Attention 机制的 BLSTM 方法模型, BLSTM-Attention 模型能够实现对词语之间以及句子之间的特征学习和提取, 并且通过 Attention 机制对不同的词语和句子给予不同的关注度, 对提高分类性能有一定作用. 本文提出的 BLSTM-Attention 模型能够更有效地提取出文本特征, 对于文本分类效果有一定的提升.

参考文献

1 曹赵娅. 我国化工行业发展的现状分析. 中国化工贸易, 2015, 7(25): 47. [doi: 10.3969/j.issn.1674-5167.2015.25.030]

2 张志强. 基于自学习向量空间模型文本分类算法的研究与应用. 软件, 2016, 37(9): 118–121. [doi: 10.3969/j.issn.1003-6970.2016.09.028]

3 罗慧钦, 陆向艳, 张雄宝, 等. 基于隐朴素贝叶斯的商品评论情感分类方法. 计算机工程与设计, 2017, 38(1): 203–208.

4 Kim Y. Convolutional neural networks for sentence classification. arXiv: 1408.5882, 2014.

5 Zhang X, Zhao JB, LeCun Y. Character-level convolutional networks for text classification. arXiv: 1509.01626, 2015.

6 刘婷婷, 朱文东, 刘广一. 基于深度学习的文本分类研究进展. 电力信息与通信技术, 2018, 16(3): 1–7.

7 Auli M, Galley M, Quirk C, *et al*. Joint language and translation modeling with recurrent neural networks. Proceedings of 2013 Conference on Empirical Methods in Natural Language Processing. Washington, DC, USA. 2013. 1044–1054.

8 邵良杉, 周玉. 基于语义规则与 RNN 模型的在线评论情感分类研究. 中文信息学报, 2019, 33(6): 124–131. [doi: 10.3969/j.issn.1003-0077.2019.06.018]

9 刘峰, 高赛, 于碧辉, 等. 基于 Multi-head Attention 和 Bi-LSTM 的实体关系分类. 计算机系统应用, 2019, 28(6): 118–124. [doi: 10.15888/j.cnki.csa.006944]

10 Nguyen HT, Le Nguyen M. An ensemble method with sentiment features and clustering support. Neurocomputing, 2019, 370: 155–165. [doi: 10.1016/j.neucom.2019.08.071]

11 蓝雯飞, 徐蔚, 汪敦志, 等. 基于 LSTM-Attention 的中文新闻文本分类. 中南民族大学学报 (自然科学版), 2018, 37(3): 129–133.

12 万圣贤, 兰艳艳, 郭嘉丰, 等. 用于文本分类的局部化双向长短时记忆. 中文信息学报, 2017, 31(3): 62–68.

13 高成亮, 徐华, 高凯. 结合词性信息的基于注意力机制的双向 LSTM 的中文文本分类. 河北科技大学学报, 2018, 39(5): 447–454. [doi: 10.7535/hbkd.2018yx05010]

14 彭敏, 杨绍雄, 朱佳晖. 基于双向 LSTM 语义强化的主题建模. 中文信息学报, 2018, 32(4): 40–49. [doi: 10.3969/j.issn.1003-0077.2018.04.005]

15 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, CA, USA. 2017. 6000–6010.

16 Zhou P, Shi W, Tian J, *et al*. Attention-based bidirectional long short-term memory networks for relation classification. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany. 2016.

17 李卫疆, 李涛, 漆芳. 基于多特征自注意力 BLSTM 的中文实体关系抽取. 中文信息学报, 2019, 33(10): 47–56, 72. [doi: 10.3969/j.issn.1003-0077.2019.10.006]