要先将其进行断句处理; “stance”是立场标签, 共有 3 个分类: “FAVOR”代表支持、“AGAINST”代表反对、“NONE”代表中立. 针对 5 个不同的目标话题, 其立场标签的分布情况如表 2 所示.

表 2 NLPCC 训练集数据分布

目标话题	FAVOR	AGAINST	NONE
深圳禁摩限电	160	300	126
春节放鞭炮	250	250	100
iPhone SE	245	209	146
俄罗斯在叙利亚的反恐行动	250	250	100
开放二胎	260	240	240

4.2 数据预处理

由于微博文本中的数据较为口语化, 并包含很多表情符号、繁体字、URL 链接、多次标点符号重复等情况. 这些情况都会对文本分析产生很大的噪声影响, 因此本文在预处理部分进行了语料清洗的工作, 主要包括: 清除了冗余的标点符号和链接, 将繁体字转为简体等, 如表 3 所示.

表 3 数据预处理对比

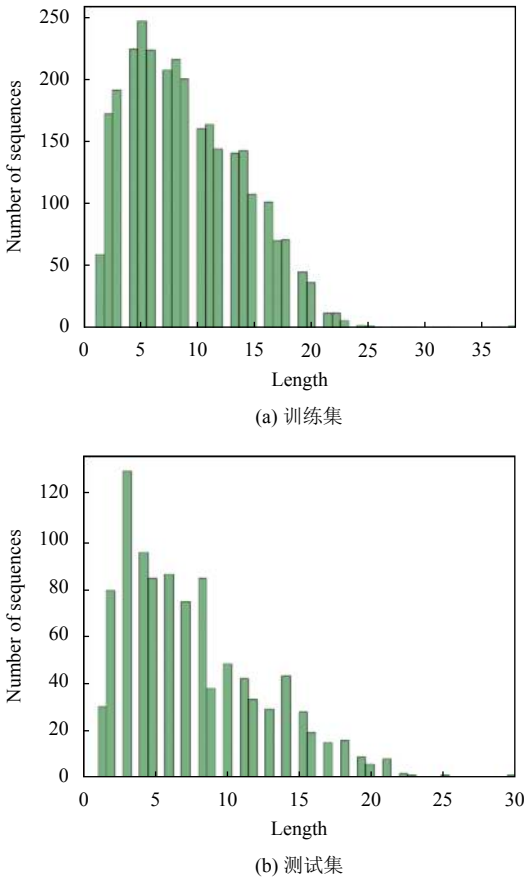
原微博文本	【今年过年, 您放鞭炮了吗?】“爆竹声中一岁除, 春风送暖入屠苏(*^▽^*). ”鞭炮声响起, 空气的污浊味也随之浓烈起来..... 除夕夜, “南京环保”发布微博呼吁大家尽量少放或不放烟花爆竹, 这个春节里, 您放鞭炮了吗?我们究竟还要不要放鞭炮呢?您怎么看?也欢迎在评论中告诉我们您的想法. http://t.cn/zYIUZpT
清洗后的微博文本	【今年过年, 您放鞭炮了吗?】“爆竹声中一岁除, 春风送暖入屠苏. ”鞭炮声响起, 空气的污浊味也随之浓烈起来…除夕夜, “南京环保”发布微博呼吁大家尽量少放或不放烟花爆竹, 这个春节里, 您放鞭炮了吗?我们究竟还要不要放鞭炮呢?您怎么看?也欢迎在评论中告诉我们您的想法.

Bert-Condition-CNN 模型的输入是基于句子级别的, 但因为微博文本的内容普遍较长, 所以需要在预处理部分将微博文本内容进行断句处理. 本文在实验中将微博文本中出现的“, ”、“?”、“!”, “、”和“. ”标点符号作为断句标识符对文本内容进行断句分割. 断句后训练集和测试集中微博文本的长度 (包含句子的个数) 分布情况如图 5 所示. 由图可见训练集和测试集文本长度的分布大体上是一致的, 且大部分数据的长度是集中在 0~25 之间, 因此为保证在计算 Condition 层时, 微博文本内容的长度一致. 所以在预处理部分将微博文本的长度固定为 25, 对长度不足 25 的数据进行“[PAD]”符号的补齐, 长度大于 25 的数据进行截断处理.

4.3 评价指标

分类器的主要评价指标有准确率 (Accuracy)、精确率 (Percision)、召回率 (Recall) 和 F 值 (F-score). 准确率是指分类正确的样本占总样本个数的比例, 精确率是指分类正确的正样本占分类器预测为正样本个数的比例, 召回率是指分类正确的正样本占真正的正样本个数的比例. 为平衡精确率和召回率之间的关系, 以免出现由于数据类别分布不均衡导致两个分数之间相差过大, 无法充分反映分类器的效果, 通常在分类任务中, 引入两者的调和平均值, F 度量值作为分类的评价指标, 其计算公式如式 (7) 所示.

$$F = \frac{2P \times R}{P + R} \tag{7}$$



(a) 训练集

(b) 测试集

图 5 训练集和测试集的微博长度

在 NLPCC 任务中, 官方给出的评价指标是使用 F_{Faver} 和 F_{Against} 的平均值作为最终评价指标. 其中 F_{Faver} 是“支持”标签的 F 度量, F_{Against} 是“反对”标签的 F 度量. 其计算公式如下:

$$F_{avg}=\frac{F_{Favor}+F_{Against}}{2} \quad (8)$$

4.4 参数设置

实验中涉及的网络模型参数如表 4 所示. 使用 Relu 作为卷积层的激活函数. 实验采用 4.1 节中介绍的数据集, 其中 3000 为训练集, 1000 为测试集. 将训练集中 20% 的数据抽出作为验证集使用, 迭代次数 *epoch*=150, 选取在验证集上得到最好效果的模型作为最终模型在测试集上进行测试.

表 4 模型参数

参数	取值
句向量维度	786
话题集长度	6
微博文本长度	25
卷积核大小	2×2
卷积核个数	128
激活函数	Relu
池化策略	max pooling
batch size	32
epoch	150

4.5 实验结果与分析

为了验证本文提出的基于 Condition-CNN 的模型在中文微博立场检测任务上的有效性. 本节进行了如下实验对比.

如表 5 所示, 首先对比了采用拼接法将目标话题和微博文本连在一起 (Concat) 和使用本文提出的 Condition 层对话题集和微博文本进行关系矩阵构建的两种方法的效果. 同时给出了 Bai[9]中提到的 BiLSTM-CNN-ATT 在相同数据集上的表现结果.

表 5 Condition 层的实验结果

算法	F_{Favor}	$F_{Against}$	F_{Avg}
BiLSTM-CNN-ATT	**0.663**	0.495	0.597
Bert-Concat-CNN	0.606	0.624	0.615
Bert-Condition-CNN	0.650	**0.711**	**0.681**

在本次对比中, Concat 和 Condition 的实验中均使用了 *Bert* 预训练模型输出句向量. 通过这两种对话题和微博文本的不同组成方式的实验结果对比表明, 基于 Condition 计算层进行话题和微博文本关系构建的方式对立场检测任务的效果有着明显的提升. 表中 BiLSTM-CNN-ATT 的模型是基于注意力的混合网络模型, BiLSTM-CNN-ATT 的 F_{Favor} 值取得了最高分, 但其分类结果不均衡的现象导致了最终的 F_{Avg} 值的降低. 通过 Concat 方法和 BiLSTM-CNN-ATT 的对比, 可以看到, *Bert* 作为句向量的语义特征抽取能力是优于 RNN 和 CNN 的甚至是优于将 RNN、CNN、Attention 拼接组合起来的效果.

表 6 中对比了 3.3 节中给出的 3 种 Condition 层计算的方法, 分别是基于欧几里得距离 (Euclidean)、余弦距离 (cosine) 和点乘计算 (dot) 的. 实验结果显示基于点乘计算的效果最佳, 并且相对于另外两个计算方式, 点乘的计算复杂度也相对较低, 因此在后续的实验中采用 Condition 计算方式都是采用点乘的方法, 包括在表 5 中的 Condition 计算也是使用的点乘.

表 6 Condition 的 3 种计算方式

方法	F_{Favor}	$F_{Against}$	F_{Avg}
Condiiton-euclidean	0.623	0.655	0.639
Condition-cosine	0.614	0.671	0.643
Condition-dot	0.650	**0.711**	**0.681**

为了方便对模型结构进行验证对比, 上述两个对比实验在进行训练及测试的时候针对的是数据集中所有的数据, 并未做话题的区分. 但实际上, 从实验数据的角度出发, 5 个目标话题是相互独立的, 因此将 5 个话题的数据分开进行单独训练会得到更好的效果. 如表 7 所示, 将话题分开单独训练的结果同 Dian[5]和 Yue[10]的 ATA 模型进行对比. 其中 Dian 的工作是基于不同特征融合的机器学习模型, 经过实验对比, 对不同目标话题采取了不同的特征组合方式. 该工作在 2016 年 NLPCC 的任务中取得了第一名的成绩. Yue 的 ATA 模型是基于深度学习的模型, 采用两段注意力机制将目标话题和微博文本进行组合. 该表中仅使用了 F_{Avg} 进行对比.

表 7 5 个话题分开单独训练结果

话题	ATA	Dian	Bert-Condition-CNN
iPhone SE	0.600	0.615	**0.631**
深圳禁摩限电	0.807	0.782	**0.800**
俄罗斯在叙利亚反恐行动	0.563	0.620	**0.636**
开放二胎	0.818	0.847	**0.849**
春节放鞭炮	0.801	0.776	**0.803**

从实验对比结果中可以看出, 基于 Bert-Condition-CNN 的模型在 5 个话题的立场检测中, F_{Avg} 均取得了最高的分值. 在话题“深圳禁摩限电”、“开放二胎”和

“春节放鞭炮”中 F_{Avg} 都取得了 0.8 以上的分数. 在话题“春节放鞭炮”和“开放二胎”的任务上以微弱的形式胜出; 在话题“俄罗斯在叙利亚反恐行动”、“深圳禁摩限电”和“iPhone SE”中取得了 1%~3% 的提升. 在同 ATA 模型的对比中, 进一步验证了 Condition 层对立场检测任务的提升.

对于分类结果较差的两个话题“俄罗斯在叙利亚反恐行动”和“iPhone SE”. 这两个话题经主题短语提取后形成的话题集如 3.1 中的表分别为{“极端组织”、“战斗民族”、“大国博弈”、“胜利阵线”、“武装分子”}和{“中国市场”、“电池续航”、“开发者大会”、“外观侵权”、“1200 万像素摄像头”}. 首先这两个话题集在数据中的覆盖率相比于其他话题的覆盖率来讲是较低的, 在通过 Condition 计算层计算时形成的关系矩阵大多较为稀疏. 因此在进行立场检测分类时得到的效果较差.

5 结论与展望

本文的主要工作是基于构建话题和微博文本之间 *Bert* 句向量的 Condition 层, 利用卷积神经网络模型, 实现了对中文微博的立场检测研究, 并给出了一种主题短语提取的方法. 经过实验对比分析, 验证了本文提出的模型 Bert-Condition-CNN 的有效性和在立场检测任务中取得的进步.

首先对微博数据进行分析发现, 单一的目标话题对微博文本数据的覆盖不足, 因此需要对微博文本进行主题短语的提取. 本文提出了基于 LDA 和点互信息提取的方式. 首先从 *n*-grams 词组集合中删去包含低频词和标点符号的无意义词组序列构成主题短语候选集, 然后使用 LDA 对文本进行主题词提取和点互信息计算, 分别用来反映词组的主题相关性和短语质量; 最终将候选集中的词组进行主题相关性和短语质量的打分, 并以在语料中出现的频率为权重, 从而选出主题短语.

其次在对文本进行向量之间的映射时, 使用了 Google 在 2018 年发布的 *Bert* 预训练模型, 直接生成句向量. 通过对话题集和微博文本的句向量进行 Condition 计算, 得到两个文本的关系特征矩阵. 对立场检测的分类是基于 Condition 层进行计算.

最后通过与目前现有研究中取得最好成绩的基于特征融合的机器学习模型和基于深度学习的模型均在相同的数据集上进行了对比, 对本文提出模型的有效性进行了验证.

本文在进行立场检测的实验对比时发现, 在“俄罗斯在叙利亚的反恐行动”和“iPhone SE”两个话题上, 本文提出的基于 Condition-CNN 模型的得分相对于其他三个话题的得分较低. 对实验结果进行分析后发现, 主要是因为针对这两个话题进行的主题短语提取结果中, 得到的结果在微博文本中的立场表现并不十分明显. 因此, 如何提取有利于进行立场检测研究的主题短语还有待改进.

参考文献

1 Pang B, Lee L. Opinion mining and sentiment analysis. Hanover, MA: Now Publishers, 2008. 1–135.

2 Mohammad SM, Kiritchenko S, Sobhani P, *et al*. A dataset for detecting stance in tweets. Proceedings of the LREC'16. France. 2016. 3945–3952.

3 Xu RF, Zhou Y, Wu DY, *et al*. Overview of NLPCC shared task 4: Stance detection in Chinese microblogs. Proceedings of the 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages. Kunming, China. 2016. 907–916.

4 郑海洋, 高俊波, 邱杰, 等. 基于词向量技术与主题词特征的微博立场检测. 计算机系统应用, 2018, 27(9): 118–123. [doi: 10.15888/j.cnki.csa.006498]

5 奠雨洁, 金琴, 吴慧敏. 基于多文本特征融合的中文微博的立场检测. 计算机工程与应用, 2017, 53(21): 77–84. [doi: 10.3778/j.issn.1002-8331.1702-0292]

6 Wei W, Zhang X, Liu XQ, *et al*. pkudblab at SemEval-2016 Task 6: A specific convolutional neural network system for effective stance detection. Proceedings of the International Workshop on Semantic Evaluation, SemEval'16. San Diego, CA, USA. 2016. 384–388.

7 Kim Y. Convolutional neural networks for sentence classification. Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar. 2014. 1746–1751.

8 Augenstein I, Rocktäschel T, Vlachos A, *et al*. Stance detection with bidirectional conditional encoding. Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. Austin, TX, USA. 2016. 876–885.

9 白静, 李霏, 姬东鸿. 基于注意力的 BiLSTM-CNN 中文微博立场检测模型. 计算机应用与软件, 2018, 35(3): 266–274. [doi: 10.3969/j.issn.1000-386x.2018.03.051]

10 岳天驰, 张绍武, 杨亮, 等. 基于两阶段注意力机制的立场检测方法. 广西师范大学学报 (自然科学版), 2019, 37(1): 42–49.

11 Danilevsky M, Wang C, Desai N, *et al*. Automatic construction and ranking of topical keyphrases on collections of short documents. Proceedings of the 2014 SIAM International Conference on Data Mining. Urbana-Champaign, Urbana, IL, USA. 2014. 61801.

12 赵斌. 基于点间互信息的主题优化方法[硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2012.

13 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805, 2018.