计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(11):45−53 [doi: 10.15888/j.cnki.csa.007152]

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

基于 Bert-Condition-CNN 的中文微博立场检测①

王安君, 黄凯凯, 陆黎明

(上海师范大学 信息与机电工程学院, 上海 201400)
通讯作者: 陆黎明, E-mail: lulimingshnu@163.com

摘　要: 微博立场检测是判断一段微博文本针对某一目标话题所表达的观点态度是支持、中立或反对. 随着社交媒体的发展, 从海量的微博数据中挖掘其蕴含的立场信息成为一项重要的研究课题. 但是现有的方法往往将其视作情感分类任务, 没有对目标话题和微博文本之间的关系特征进行分析, 在基于深度学习的分类框架上, 扩展并提出了基于 Bert-Condition-CNN 的立场检测模型, 首先为提高话题在文本中的覆盖率, 对微博文本进行了主题短语的提取构成话题集; 然后使用 Bert 预训练模型获取文本的句向量, 并通过构建话题集和微博文本句向量之间的关系矩阵 Condition 层来体现两个文本序列的关系特征; 最后使用 CNN 对 Condition 层进行特征提取, 分析不同话题对立场信息的影响并实现对立场标签的预测. 该模型在自然语言处理与中文计算会议 (NLPCC2016) 的数据集中取得了较好的效果, 通过主题短语扩展后的 Condition 层有效地提升了立场检测的准确度.

关键词: 立场检测; 主题短语; 关系矩阵; 句向量

引用格式: 王安君,黄凯凯,陆黎明.基于 Bert-Condition-CNN 的中文微博立场检测.计算机系统应用,2019,28(11):45–53. http://www.c-s-a.org.cn/1003-3254/7152.html

Stance Detection in Chinese Microblogs via Bert-Condition-CNN Model

WANG An-Jun, HUANG Kai-Kai, LU Li-Ming

(College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 201400, China)

Abstract: Stance detection task aims to automatically determine whether a Weibo text is in favor of the given target, against the given target, or neither. Mining the stance information about a given target is an emerging problem. Based on the success of deep learning in classifying, this study proposed a Bert-Condition-CNN model to predict the stance label. Firstly, noted that the given target may not be present in the Weibo text, so we extracted the topic phrases from Weibo corpus as the given target supplement. Then, we used Bert language model to accept the text representation vector and calculated a Condition matrix whose entries represent the relationship between Weibo text and topic phrases. Finally, a convolutional neural network was utilized to capture the stance features from Condition matrix. Experimental results on NLPCC2016 datasets demonstrate the model has achieved a sound effect of stance detection.

Key words: stance detection; topic phrase; condition matrix; text representation

1　引言

近年来针对微博数据的情感分析引起了广泛的关注[1], 同时也促进了立场检测研究的兴起与发展. 立场检测可以看作是针对特定目标话题进行的对情感分析任务的改进. 2016 年 Mohanmmad 等[2]构建了基于 Twitter 数据的立场检测英文数据集, 并用于 SemEval-2016 会议的 Task 6: 立场检测 (Stance Detection). 随后, Xu 等[3]从 Mohammad 等人的工作中受到启发, 构建了

① 收稿时间: 2019-04-12; 修改时间: 2019-05-08; 采用时间: 2019-05-16; csa 在线出版时间: 2019-11-06

面向中文微博的立场检测数据集, 并将其用于 2016 年的自然语言处理与中文计算会议 (Natural Language Processing and Chinese Computing, NLPCC) 发表的任务中.

立场检测任务通过自然语言处理技术, 分析出当前微博文本内容对目标话题的立场倾向是“支持”、“反对”还是“中立”. 看似与情感分析相似, 但情感分析侧重的是一段文本中情感特征的极性, 而立场检测是要根据给定的目标话题来判断文本的立场, 在很多情况下是无法仅仅从文本的情感极性来判断其立场分类的. 例如, “最反感这些拉客的! 还有在机动车道上行驶的!”, 这条微博不考虑任何目标话题时, 它的情感极性是消极的, 但是当针对“深圳禁摩限电”这个话题时, 这条微博的立场应为“支持”. 由于立场检测任务比情感分析多出一步很重要的目标话题特征, 所以在模型的设计与使用上自然也要与情感分析有所不同. 早期的立场检测研究中, 往往直接忽略掉目标话题而只是对微博的文本内容进行类似情感分析的处理. 而在当前的立场检测研究中, 将目标话题与微博文本内容以不同方式拼接到一起然后进行分类, 这些方法都没有对目标话题同微博文本之间的关系特征进行分析.

本文提出了一个基于 Bert-Condition-CNN 的立场检测模型, 首先对微博文本集进行主题短语的提取以扩大话题信息在微博文本中的覆盖率; 然后使用 BERT 获取扩充后的话题集和微博文本的句向量, 通过构建两个文本序列间的 Condition 矩阵来提取话题信息和微博文本间的关系特征; 最后使用 CNN 对关系矩阵 Condition 层进行立场信息的判断.

2 相关工作

针对立场检测任务, 目前国内外的研究人员采用的方法主要有基于特征工程的机器学习方法和基于神经网络的深度学习方法.

2.1 基于特征工程的机器学习方法

Zheng 等[4]将微博文本中的情感词和主题词作为特征词进行提取, 然后通过 Word2Vec 对特征词进行词向量的训练, 将词向量取平均作为文本的特征输入到 SVM 分类器中进行立场分类. 实验表明只使用情感词作为特征时, 对立场的分类并不理想, 情绪并不能准确地反映作者的立场倾向, 而加入主题词的特征选取效果更好. Dian 等[5]探究了文本的多种特征融合对立场检测的影响, 分别有基于词频统计的词袋特征、基于同义词典的词袋特征、词与立场标签的共现关系特征、文本的 Word2Vec 的字向量和词向量, 对这些特征的不同组合方式, 分别使用 SVM、随机森林和决策树对进行立场分类, 实验表明词与立场标签的共现关系同 Word2Vec 的字、词向量的组合对立场分类的结果改善最为明显.

2.2 基于深度学习的方法

相比基于特征工程的机器学习方法而言, 深度学习的优势在于不用进行复杂的人工特征抽取, 而是通过将文本内容全部映射为向量, 然后使用多层的神经网络与标签之间进行拟合自动学习文本的特征. 目前现有的立场检测研究中, 基于深度学习的工作主要是通过将目标话题信息以不同的方式添加到微博文本内容中和通过修改神经网络结构这两种方法来提升立场检测效果.

Wei 等[6]使用基于 Yoon Kim[7]的卷积神经网络对微博文本进行分类, 它使用了一种对模型投票的机制来融合训练中产生的各模型结果, 每一个 epoch 训练结束后都会迭代一些测试集数据对标签进行预测, 最终测试集的结果是将所有的 epoch 迭代完, 每条数据选择被预测次数最多的标签作为最终结果, 但是它只针对微博文本进行了特征提取和分类, 而忽略了目标话题在立场检测中的作用. 针对这个问题, Augenstein 等[8]提出了一个 Bidirectional Conditional Encoding 模型将目标话题与微博文本进行拼接, 通过使用 BiLSTM (Bidirectional Long Short-Term Memory) 将目标话题细胞状态层的输出作为微博文本 BiLSTM 的细胞状态层的初始值, 从而实现两个文本序列的拼接, 而隐层状态的 BiLSTM 对目标话题和微博文本的编码则是相互独立的. 为了加强模型针对目标话题对立场检测的影响, Bai 等[9]提出了一种基于注意力的 BiLSTM-CNN 模型对中文微博立场进行检测, 首先使用 BiLSTM 和卷积神经网络 CNN 分别获取文本的全局特征和局部卷积特征; 然后使用基于注意力 (Attention) 的权重矩阵将文本的 BiLSTM 输出加入到 CNN 的输出中; 将最终获取到的 CNN 的句子表示输入 Softmax 层进行分类. 在基于注意力机制的方法上, Yue 等[10]提出了基于两段注意力机制的立场检测模型, 首先使用 Word2Vec 进行词向量表示; 然后对微博文本的词向量和目标话题的词向量进行 Attention 计算, 使用 BiLSTM 对微博文本进行特

征提取, 对提取到的特征再次与目标话题进行 Attention 计算, 将最后得到的结果使用 Softmax 进行分类.

根据对现有研究的分析和对比, 如何充分发挥话题信息在立场检测任务中的作用是本文研究的重点.

3 本文工作

本文的主要工作是设计完成了基于 BERT–Condition-CNN 的中文微博立场检测模型. 首先, 为增大话题信息在微博文本中的覆盖率, 本文结合 LDA 和点互信息, 在数据处理部分对微博文本进行主题短语的提取, 将目标话题进行扩充构成话题集; 然后进行网络模型的构建, 使用 Bert 获取话题集和微博文本的句向量 (分别用 U、V 表示), 并构建两个句向量矩阵的 Condition 层 C, 以计算目标话题和微博文本的关系特征; 最后使用 CNN 学习得到最终的立场信息输入 Softmax 层得到立场标签. 模型的流程图如图 1 所示, 本章将对模型中各个部分的具体实现步骤进行介绍.

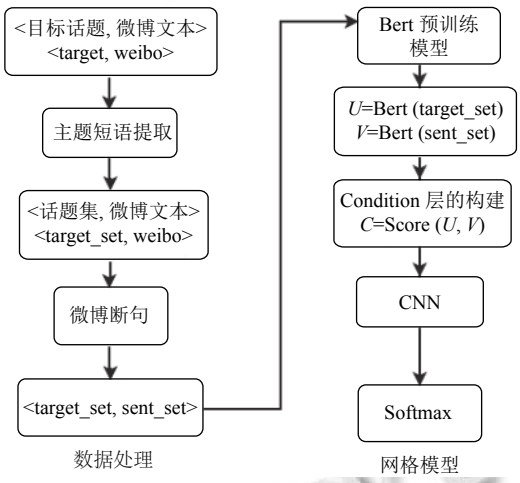


图 1 Bert-Condition-CNN 流程图

3.1 主题短语提取

本文的主题短语提取采用的是基于 *n*-grams 的识别技术[11], 首先对语料进行 *n*-grams 词组集合的构建 (这里的 *n*-grams 词组是指由相邻的 *n* 个词组成的词组序列); 然后将 *n*-grams 词组集合中包含低频词和标点符号的无意义词组序列进行删除构成主题短语候选集; 最后对候选集中的词组进行打分, 主要考虑两个方面: 主题关联度和短语质量. 主题关联度是指词组序列中包含主题词的比例num_{keys}/n(其中 num_{keys} 是词组中含主题词的个数; n 是词组的长度), 本文使用 LDA 主题模型对微博文本进行主题词的提取; 短语质量是指词组中相邻词之间的点互信息和, 点互信息 (Pointwise Mutual Information, *PMI*) 通常被用于计算两个词之间的关联度[12], 其计算公式如下:

$$PMI(word_1,word_2)=\log_2\frac{P(word_1,word_2)}{P(word_1)P(word_2)} \quad (1)$$

以词组在语料中出现的频率 *freq* 为权重, 最终短语的得分计算方法为:

$$score=freq\times\frac{num_{\text{keys}}}{n}\times\sum_{w_i,w_j\in phrase}PMI\left(w_i,w_j\right) \quad (2)$$

设置短语得分的阈值为 s, 当词组的 *score* 大于等于 s 时, 认为该词组可合并为主题短语. 如果词组的 *score* 小于 s, 则其为普通词组序列. 主题短语提取的算法步骤如算法 1.

算法 1. 主题短语提取算法

1. 将微博语料进行分词处理并进行词频统计;
2. 使用 LDA 对语料进行主题词的提取, 设定主题个数为 K;
3. 构建语料的 *n*-grams 词组集合, 删除其中包含步骤 1 中统计出的低频词和包含标点符号的词组序列, 构成主题短语候选集 D;
4. 对步骤 3 构建的候选集 D 中的词组进行主题关联度和短语质量的打分, 将短语得分大于阈值 s 的词组作为主题短语进行提取.

上述算法通过主题关联度过滤掉不包含主题词或包含主题词比例较少的词组, 通过计算词间的 *PMI* 值来判定词组合并为短语是否合理, 最后通过频率筛选掉具有高主题关联度和高短语质量但出现次数不多的词组.表 1 为 NLPCC 语料中 5 个目标话题的主题短语提取结果、实验中低频词的阈值为 3、主题个数 K=200、短语得分阈值 s 为 0.0019, 从每个话题的结果中选取 5 个作为最终的主题短语.

表 1 NLPCC 中主题短语的提取结果

目标话题	主题短语
春节放鞭炮	环卫工回家、雾霾、低碳、市环保局、燃放烟花爆竹
iPhone SE	中国市场、电池续航、开发者大会、外观侵权、1200 万像素摄像头
开放二胎	取消晚婚假、女性回归家庭、女性就业、延长退休年龄、人口老龄化
俄罗斯在叙利亚的反恐行动	极端组织、战斗民族、大国博弈、胜利阵线、武装分子
深圳禁摩限电	只可自行车送外卖、深圳禁摩限电对准餐饮业、电动自行车、非法运营、整治行动

3.2 Bert 句向量

2018 年 Google AI 团队发布了一种新的语言模型 Bert[13], Bert 一经推出, 给自然语言处理中的预训练模型带来了突破性的发展, 在许多自然语言处理任务上取得了 state-of-the-art 的成绩. Bert 是一种多层双向的 Transformer 编码器, 其结构如图 2 所示 (图中 T_m 模块为 Transformer 中的 Encoder 部分).

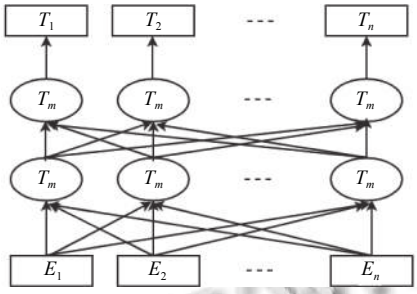


图 2 Bert 结构图

Bert 的预训练过程使用的是两个非监督任务: Masked LM(掩码语言模型) 和 Next Sentence Prediction (下一句话预测). 第一个任务是使用 Masked LM 实现了双向语言模型的预训练, 不同于 Word2Vec 等其他语言模型需要对输入序列中所有词进行预测, Masked LM 是在输入数据中随机选取 15% 的词进行 masked 操作, 通过上下文的词去预测这 15% 的词, 以避免下文的词对当前词的影响, 从而实现了真正意义上的“双向”. 这 15% 被 masked 的词中, 有 80% 是用“[MASK]”符号进行替代, 10% 用语料中随机抽出的词进行替代, 剩余的 10% 保留原有词不进行转变. Bert 的第二个任务是 Next Sentence Prediciton, 用来判断两句话 (A, B) 是否为上下句关系的二分类任务. 训练数据中 50% 的 (A, B) 数据是真实上下句作为正例, 剩余的 50% 的 (A, B) 中的 B 是随机抽取的作为负例进行训练. 该任务的最终预训练结果可以达到 97%~98% 的准确率. Bert 预训练模型也可作为 fine-tuning 用于改善序列对分类的效果, 用于 QA (判断两句话是否为问答对) 和 NLI(自然语言推理) 任务等.

在本文的实验中, 使用了 Google 发布的 Bert 中文的预训练模型“BERT-Base, Chinese”. 该模型采用了 12 层的 Transformer, 输出大小为 768 的维度向量, multi-head Attention 的参数为 12, 模型总参数大小为 110 MB, 共包含约 2 万的中文简体字和繁体字, 含有部分英文单词和数字. 将模型载入后, 可以直接输出训练好的字向量或句向量. 本文使用该模型获取句向量并将其作为后续网络模型的输入.

3.3 Condition 计算层

使用 3.1 节中抽取出的主题短语对目标话题进行扩充得到话题集 (*targets*), 在<话题集, 微博文本>的数据中, 微博文本可以对应到更多的话题相关信息. 将扩充后的话题集和微博文本使用句子序列的方式进行表示: 话题集 $targets=\{target_1, \cdots, target_n\}$、微博文本 $weibo=\{sent_1, \cdots, sent_m\}$, 其中 n、m 分别代表话题集中包含话题的个数和微博文本中包含的句子个数. 不同与第一章中介绍的用 Attention 将话题以不同的权重加到微博文本中的计算方法, 本文提出的方法是对话题集和微博文本进行关系矩阵的计算. 如图 3 所示, 在对 targets 和 weibo 进行关系矩阵的计算前, 先将 $target_i$ 和 $sent_j$($0<=i<=n$, $0<=j<=m$) 通过 *Bert* 预训练模型输出为句向量, 记 $u_i=Bert(target_i)$, $v_j=Bert(sent_j)$, 得到的关系矩阵称为 Condition, 其中 $c_{ij}=score(u_i, v_j)$.

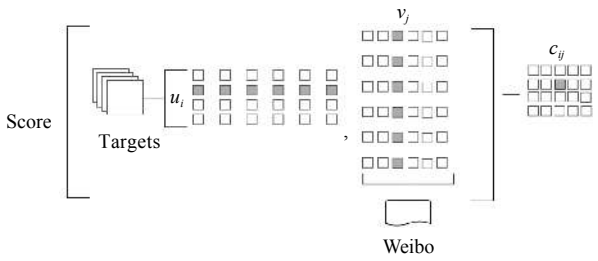


图 3 Condition 计算层的构建

Condition 层的作用可以看作是对原本计算<*target*, *weibo*>的立场检测任务分解为计算每一对<$target_i$, $sent_j$>序列组合的立场检测. 通常认为立场检测任务分为以下两个步骤: 一是判断 $sent_j$ 是否是围绕 $target_i$ 进行展开评论的, 即两个文本序列之间是否存在蕴含关系; 二是 $sent_j$ 针对 $target_i$ 的立场是支持、中立还是反对的. 若 $sent_j$ 与 $target_i$ 不存在蕴含关系时, 则其立场为中立. 在本节的 Condition 计算层中进行的主要工作是通过计算 $sent_j$ 与 $target_i$ 的关系得分 $score\ (u_i, v_j)$, 从而判断两个序列是否在蕴含关系. 由于 u_i, v_j 均是句向量 (句向量的维度为 d), 所以在计算 $score\ (u_i, v_j)$ 时, 参考向量之间的距离计算, 本文设计了归一化的欧几里

得距离、余弦距离和向量点乘的 3 种方法:

(1) 归一化的欧几里得距离

欧几里得距离是向量中常用的距离定义, 两点的距离越大, 欧几里得距离越大, 由此代表的两个句向量之间的关系就越小. 因此欧几里得距离与 u_i, v_j 的关系成反比. 故在本文的实验中采取式 (3) 所示的归一化处理, 使得 $score(u_i, v_j)$ 与 u_i, v_j 的关系形成正比.

$$score\left(u_i,v_j\right)=1/\left(1-\sqrt{\sum_{l=1}^{d}\left(u_{il}-V_{jl}\right)^2}\right) \tag{3}$$

(2) 余弦距离

余弦距离计算的是两个向量所形成夹角的余弦值, 如式 (4) 所示, 值越大说明两个句向量的夹角越小, 两个向量的关系就越大.

$$score\left(u_i,v_j\right)=\frac{\sum_{l=1}^{d}u_{lk}v_{lk}}{\sqrt{\sum_{l=1}^{d}u_{il}^2}\sqrt{\sum_{l=1}^{d}v_{jl}^2}} \tag{4}$$

(3) 向量点乘

向量点乘的计算同余弦距离相比, 不仅可以体现两个向量之间的夹角, 还反映了向量 u_i 在向量 v_j 上的映射大小, 计算公式如式 (5) 所示.

$$score\left(u_i,v_j\right)=\sum_{l=1}^{d}u_{lk}v_{lk} \tag{5}$$

在这 3 种计算向量关系的方法中, 欧几里得距离是通过计算空间距离来反映向量之间的关系; 余弦距离是通过计算空间中两个向量之间的夹角余弦值来反映向量之间的关系. 点乘计算不但反映了向量间的夹角, 而且其计算复杂度和空间复杂度都相对较低, 因此在深度学习中, 通常使用点乘来计算两个向量之间的关系. 通过 3 种方法计算得到的关于话题集和微博文本之间的关系矩阵 Condition 层, 反映了微博文本和话题集的蕴含关系. 在后续特征提取的计算中, 以 Condition 层作为输入进行分类.

3.4 CNN 特征提取层

CNN 特征提取层的输入是 3.3 节中的 Condition 计算层, 该特征矩阵为话题集与微博文本之间的关系矩阵, 其中涵盖了话题 $target_i$ 和文本 $sent_j$ 这一对文本序列中存在的蕴含关系和所持立场信息. 本节内容针对 Condition 层对所有<$target_i$, $sent_j$>序列对计算得到的 C_{ij} 进行特征融合并分类, 通过二维卷积计算相邻序列对<u_i, v_j>的关系特征对最终立场分类影响的权重, 计算公式如式 (6) 所示.

$$s_{i,j}=f\left(\sum_m\sum_n C(m,n)K(i-m,j-n)+b\right) \tag{6}$$

式中, $K(i–m, j–n)$ 为卷积核权重参数, b 为偏置项, f 为非线性激活函数, 通常为 Relu、Sigmoid 或 Tanh. 卷积后的特征矩阵 S 要经过最大池化层的处理, 池化层可以看作是一种降采样方式, 最大池化就是选取当前池化窗口中最大的数值作为特征, 可有效缩减特征矩阵的大小, 缩小模型参数数量, 从而加快计算速度, 有利于减少模型的过拟合问题. 将池化后的特征向量使用全连接进行特征融合, 然后进行 Softmax 算法对其进行分类. 全连接层和 Softmax 层的主要任务是将最终获取到的特征信息进行融合, 获取特征向量对于每个立场标签的得分, 并输出<*targets*, *weibo*>的最终立场标签. 本文采用 Softmax 层是概率转换层, 将输入的向量以概率形式表示, 完成对立场标签的预测.

4 实验结果与分析

4.1 数据集

本文使用的数据集是 NLPCC 在 2016 年发布的任务 4: “中文微博立场检测任务”中所提供的公开数据集. 该数据集中共包含 4000 条已标注立场类别标签的中文微博数据, 其中 3000 条为训练集, 1000 条为测试集, 如图 4 所示.

<id>	766
<target>	春节放鞭炮
<weibo>	今年哈尔滨雾霾很严重, 我作为一名少先队员积极号召大家少放鞭炮，从自我做起, 建议父母不放鞭炮. “春节不放鞭, 赶走雾霾天”
<stance>	AGAINST

图 4 立场检测任务数据

数据以“<id><target><weibo><stance>”的格式给出, 其中“target”为目标话题, 共有 5 个, 分别是: “iPhone SE”、“春节放鞭炮”、“俄罗斯在叙利亚的反恐行动”、“开放二胎”和“深圳禁摩限电”; “weibo”为微博文本内容, 一般文本长度较大, 因此在进行实验前需

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

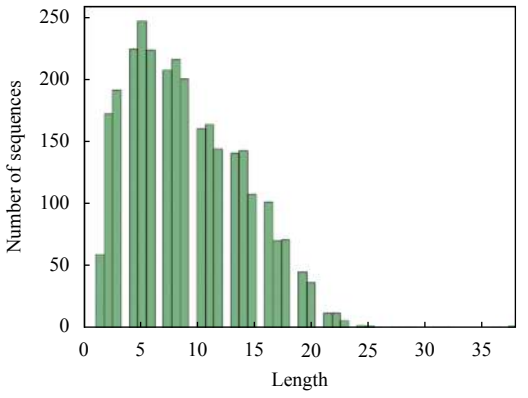
原微博文本	【今年过年, 您放鞭炮了吗?】“爆竹声中一岁除, 春风送暖入屠苏(*^▽^*). ”鞭炮声响起, 空气的污浊味也随之浓烈起来..... 除夕夜, “南京环保”发布微博呼吁大家尽量少放或不放烟花爆竹, 这个春节里, 您放鞭炮了吗?我们究竟还要不要放鞭炮呢?您怎么看?也欢迎在评论中告诉我们您的想法. http://t.cn/zYIUZpT
清洗后的微博文本	【今年过年, 您放鞭炮了吗?】“爆竹声中一岁除, 春风送暖入屠苏. ”鞭炮声响起, 空气的污浊味也随之浓烈起来…除夕夜, “南京环保”发布微博呼吁大家尽量少放或不放烟花爆竹, 这个春节里, 您放鞭炮了吗?我们究竟还要不要放鞭炮呢?您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

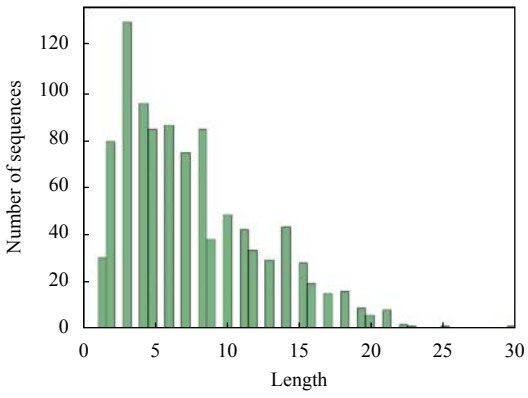
4.3 评价指标

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.