

基于 SBERT - Attention - LDA 与 ML - LSTM 特征融合的烟草问句意图识别方法

朱波¹ 黎魁¹ 邱兰¹ 黎博²

(1. 昆明理工大学机电工程学院, 昆明 650504; 2. 武汉工程大学机电工程学院, 武汉 430205)

摘要: 针对烟草领域中问句意图识别存在的特征稀疏、术语繁多和捕捉文本内部的语义关联困难等问题, 提出了一种基于 SBERT - Attention - LDA (Sentence-bidirectional encoder representational from transformers - Attention mechanism - Latent dirichlet allocation) 与 ML - LSTM (Multi layers - Long short term memory) 特征融合的问句意图识别方法。该方法首先基于 SBERT 预训练模型和 Attention 机制对烟草问句进行动态编码, 转换为富含语义信息的特征向量, 同时利用 LDA 模型建模出问句的主题向量, 捕捉问句中的主题信息; 然后通过更改后的模型级特征融合方法 ML - LSTM 获得具有更为完整、准确问句语义的联合特征表示; 再使用 3 通道的卷积神经网络 (Convolutional neural network, CNN) 提取问句混合语义表示中隐藏特征, 输入到全连接层和 Softmax 函数中实现对问句意图的分类。基于烟草行业权威网站上获取的数据集开展了实验验证, 实验结果表明, 所提方法相比其他几种深度学习结合注意力机制的方法精确率、召回率和 F1 值上有显著提升, 与 BERT 和 ERNIE (Enhanced representation through knowledge integration and embedding) - CNN 模型相比提升明显, F1 值分别提升 2.07、2.88 个百分点。

关键词: 烟草问句分类; 自然语言处理; 特征融合; 自注意力机制

中图分类号: TP183; TP391

文献标识码: A

文章编号: 1000-1298 (2024) 05-0273-09

OSID:



Tobacco Interrogative Intent Recognition Based on SBERT - Attention - LDA and ML - LSTM Feature Fusion

ZHU Bo¹ LI Kui¹ QIU Lan¹ LI Bo²

(1. Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming 650504, China

2. Faculty of Mechanical and Electrical Engineering, Wuhan Engineering University, Wuhan 430205, China)

Abstract: Aiming at the problems of feature sparsity, terminology and difficulty in capturing semantic associations within the text in question intention recognition in the tobacco domain, a feature fusion method based on sentence-bidirectional encoder representational from transformers - Attention mechanism - latent dirichlet allocation (SBERT - Attention - LDA) and multi layers - long short term memory (ML - LSTM) feature fusion was proposed. The method first dynamically encoded the tobacco question based on the SBERT pre-training model combined with the Attention mechanism and converted it into semantic-rich feature vectors, and at the same time, the topic vector of the question was modelled by using the LDA model to capture the topic information in the question; and then the joint feature representation with more complete and accurate question semantics was obtained by using the modified model-level ML - LSTM feature fusion method; and then the three-layer LSTM and ML - LSTM feature fusion method was used to identify the intention of the question. Then a 3-channel convolutional neural network (CNN) was used to extract the hidden features in the hybrid semantic representation of the question and fed them into the fully connected layer and Softmax function to achieve the classification of the question intent. Compared with the enhanced representation through knowledge integration and embedding (BERT and ERNIE) CNN models, the improvement was obvious (the F1 values were improved by 2.07 percentage points and 2.88 percentage points, respectively), which supported the construction of the Q&A system for tobacco websites.

Key words: classification of tobacco questions; natural language processing; feature fusion; self-attention mechanism

收稿日期: 2023-12-26 修回日期: 2024-03-08

基金项目: 中国烟草总公司云南省烟草公司重点项目 (2021530000241012)

作者简介: 朱波 (1978—), 男, 讲师, 博士, 主要从事信息化和机器学习模式识别研究, E-mail: zhubo20110720@163.com

0 引言

随着烟草行业智能化的推进,越来越多的信息化与智能化从业人员参与到烟草行业的智能化建设中,他们通常不具备烟草领域的背景知识,亟需通过搜索引擎或问答系统获得智能化的信息服务。问答系统相比搜索引擎回答问题更准确、更全面、更高效,更加适应问题多样化和复杂化的趋势。然而,烟草领域问句存在口语化严重、文本较短、难以捕捉到文本内部的语义关联等问题,例如“什么是烤烟?”,由于句中缺乏足够的上下文信息,系统无法确定用户具体想了解什么信息,回答可能会涉及到烤烟的品种、加工技术、市场价值等多个方面,导致回答过于冗杂。这些问题给问句意图的准确识别造成挑战,导致问答系统难以准确理解用户意图,无法满足用户的需求。

问句意图识别是一种将用户的语言表达转化为具体目的的过程^[1],通过分析问句的语义内容,确定问句背后所隐含的特定意图,是实现智能问答的关键环节。例如,在烟草自动问答系统中,需要对用户提问的意图进行分类,以明确用户到底是在寻求烟草种植、销售,还是加工等具体问题,跳转到相应的业务模块中,以缩小问题空间,由此提高回答准确率。问句意图识别隶属于文本分类的研究范畴,涉及到对问句文本的特征提取、语义分析等技术。在2010年之前,基于浅层机器学习的文本分类模型^[2]占据主导地位,主要包括:支持向量机^[3]、最大熵模型^[4]、隐马尔可夫模型^[5]、分类回归树^[6]、随机森林^[7]、朴素贝叶斯^[8]等。基于浅层机器学习方法具有针对性强、可解释性好、能够反映词语在相关领域的重要性等优点,在精度和稳定性方面相比于传统基于规则的方法具有明显的优势,但也存在文本上下文语义表示能力弱和依赖人工提取关键特征的缺点,对数据集要求高,对使用者经验要求也高,因而难以获得较好的泛化性。在2010年以后,文本分类逐渐从浅层特征学习向深度学习模式^[9]转变。后者相比前者能够获得更好的泛化性能。学术界针对基于深度学习的文本分类方法开展了大量研究,KIM^[10]在预训练 Word2vec 词向量的基础上,首次提出了基于卷积神经网络的文本分类模型,应用到句子级别的文本分类中,其具有参数少、计算量小、训练速度快等优势,取得了较好的效果。史梦飞等^[11]使用 Bi-LSTM 和 CNN 结合的方式同时获取问句的时序特征和本质特征,可以更好地提取待分类问句中的有效信息,在多个数据集上都取得了不错的效果。为了能够更好利用文本中不同词语的重

要度信息,Attention 机制^[12]被引入到分类任务中,其可以对文本中的每个词或句子分配不同的权重,从而突出关键信息并减少噪声的干扰。王郝日钦等^[13]通过在卷积神经网络(CNN)上下游卷积块之间建立一条稠密的链接,并结合 Attention 机制,使文本特征提取更具有代表性,在水稻问答系统问句分类上获得更好效果。随着深度学习的发展,ELMo (Embeddings from language modeling)^[14]、GPT (Generative pre-training)^[15]、BERT (Bidirectional transformer for language understanding)^[16]等大量动态语义特征学习模型受到广泛关注,通过在海量样本下预训练出的深度语言模型上实施迁移学习,能够获得具有丰富语义关系的词嵌入映射函数,生成的词向量拥有丰富的先验信息,并且能充分结合上下文语义^[17-18]。基于预训练模型的文本分类研究虽已相对成熟,但模型复杂、参数量大造成模型训练难度大且可解释性差,单纯依靠预训练模型存在未充分考虑文本所处的领域信息(全局信息),造成语义表达不完整的问题。因此学者们开始研究将深度学习模型与隐含狄利克雷分布模型(LDA)相结合来处理自然语言任务。LDA 对主题建模的全局表示可以更好地理解文档集合中的主题结构和主题之间的关系,结合深度学习在文本局部特征提取方面的优势,使词语更加紧密,特征融合后的主题语义性也更强。文献[19]将 LDA 模型与 SBERT 模型相结合,挖掘在线评论中主题的情感信息取得较好效果。但是,目前针对主题信息与深度语义信息相结合的研究还不够系统和全面,特征融合只采用简单的直接拼接方式,没有考虑两者对下游任务的重要度差异。

针对烟草领域问句特点,为实现问句文本在信息稀疏且相关的高维空间中的主题挖掘,提出基于混合神经网络 SBERT-Attention-LDA 与多层次 LSTM 特征融合的问句意图识别模型,该模型一方面通过对 SBERT 微调获得蕴含语义信息的句子嵌入向量,并且通过自注意力机制使 SBERT 模型更加专注于对问句关键部分的提取,另一方面通过 LDA 主题模型对烟草问句语料库进行概率计算生成概率主题向量,并通过模型级特征融合方法 ML-LSTM 获得具有更为完整、准确问句语义的联合特征表示,最后,采用3层通道的 CNN 来提取向量中的隐藏语义特征,以提高烟草问句意图识别准确率,为设计准确、高效的烟草领域问答系统奠定基础。

1 数据处理

1.1 数据采集

本文在烟草在线 (<http://www.tobaccochina>).

com)、烟悦网(<http://bbs.yanyue.cn>)、国家烟草专卖局(<http://www.tobacco.gov.cn>)等网站上采用 Scrapy 爬虫框架共获取有关烟草行业知识的问答语料数据 29 808 条。

1.2 数据预处理

为了尽可能突出文本特征及语义信息,本文首先去除与主题无关的标点、乱码、语气词和特殊符号等,过滤少于 8 个字符的超短文本,以及多于 50 个字符的长文本,然后采用 Python 的 Jieba 分词库对文本进行分词。此外,对于烟草问答数据集中广泛存在的与问句主题强相关的英文术语,如“NRT(尼古丁替代疗法)”、“TCA(烟草控制法案)”等,为突出其文本特征,还进行了标准化操作,将其替换为对应的汉语全称。

1.3 问句标注及划分

预处理得到 27 400 条语料数据,从中随机抽取 10 000 条,利用关键词的词频及共现频次主题聚类后划分类别,聚类效果如图 1 所示。各个主题相隔

较远,说明聚类效果明显,且在主题无相交的条件下,主题一致性(Coherence value,CV)^[20]达到最大。以图 1 中类别 9 为例,在数据集中与该主题最相关的主题词显示有“种植”“种子”“播种”“育苗”“施肥”“病虫害”“成熟度”等,通过总结词汇表达的意义,归纳出该主题,结合实际应用形成烟草问句的 10 个类别,包括烟草危害与健康问题、烟草市场与经济、烟草行业政策法规、烟草生产与加工等。然后进行手工标注,将其中 80% 的数据作为训练集,10% 的数据作为验证集,10% 的数据作为测试集。具体类别和样本数以及经分词后的关键词序列如表 1 所示。

2 模型构建

本文提出的基于 SBERT - Attention - LDA 与特征加权融合的问句意图识别模型框架主要包括双通道混合语义表示模块和基于 3 层通道 CNN 的特征学习预测模块,模型组成单元为文本输入层、特征表

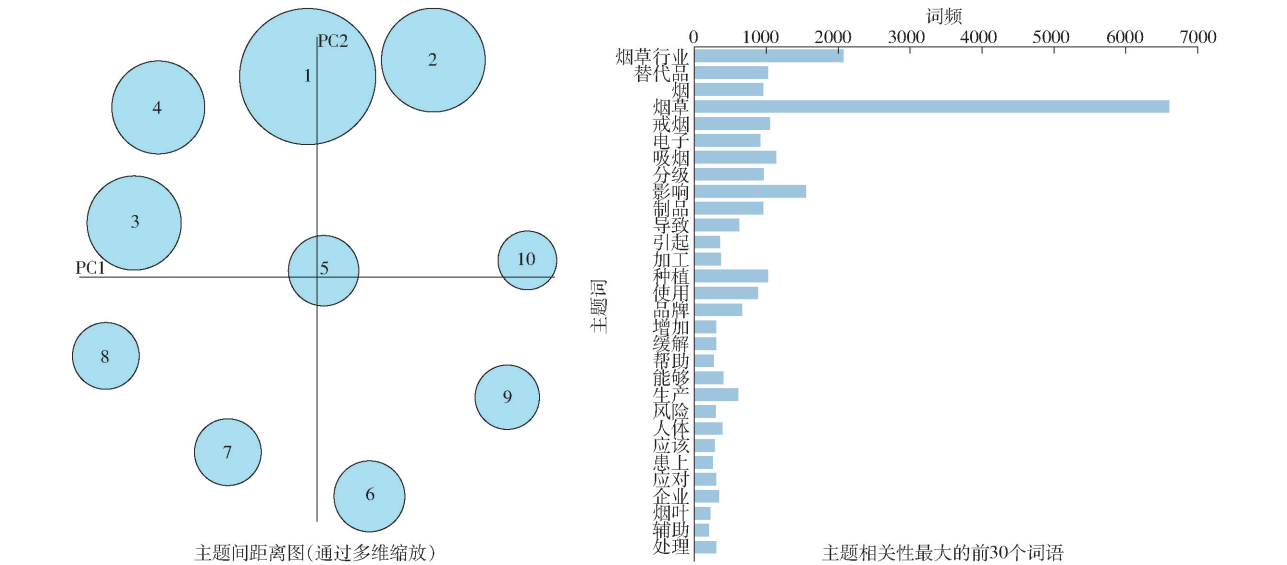


图 1 聚类效果

Fig. 1 Clustering effect diagram

表 1 烟草问句数据集类别划分和样例

Tab. 1 Classification and sampling of tobacco survey dataset

类别	数量/条	类别样例	分词预处理后的样例词序列
烟草种类和品牌问题	2 681	烟草行业中存在一些品牌之间的合作和联盟?	烟草/行业/存在/品牌/之间/合作/联盟
烟草危害与健康	3 911	我如何处理因戒烟而引起的身体剧痛和失眠?	处理/戒烟/引起/身体/剧痛/失眠
烟草制品用途与效果	1 216	烟草制品是否具有抗炎作用?	烟草/制品/具有/抗炎/作用
烟草生产与加工	3 068	如何进行烤烟的干燥和烘烤?	进行/烤烟/干燥/烘烤
戒烟与辅助戒烟	3 097	可以进行深呼吸练习来辅助戒烟吗?	进行/深呼吸/练习/辅助/戒烟
烟草行业政策法规	2 814	TCA 的效果如何评估?	烟草控制法案/效果/评估
烟草市场与经济	2 925	客户满意度对企业经济效益的影响如何评估?	客户/满意度/企业/经济/效益/影响
电子烟与烟草替代品	2 436	电子烟的使用会对人体的免疫系统产生影响?	电子烟/人体/免疫系统/产生/影响
烟草种植问题	3 094	烟草的生长过程中需要进行病虫害的监测?	烟草/生长/过程/需要/进行/病虫害/监测
烟叶分级问题	2 258	烟叶分级是否涉及烟草的香气成分检测?	烟叶/分级/涉及/烟草/香气/成分/检测

示层、特征融合层、多尺度卷积层、分类层,如图 2 所示。首先,将文本序列作为输入,序列长度设置为固定值 50。保留到前 50 个词,对于长度小于 50 的文本,使用特殊字符“PAD”将文本长度补齐。其次,由 LDA 构建词袋模型^[21],将每个问句表示为一个向量,使用词频-逆文档频率(TF-IDF)方法计算权重,再使用构建好的词袋模型训练 LDA,通过迭代计算得到每个词属于每个主题的概率以及每个问句含有的主题概率分布。对于每个问句,LDA 模型会给出其主题分布概率,经过微调训练后的 SBERT 模型使用预训练的上下文语义信息来生成句子的 512 维度特征向量,这些特征向量捕捉了句子之间的语义相似性。然后,使用 Attention 计算每个维度的权重,给句向量中的每个维度赋予不同的重要性,得到一个根据上下文语义特征进行加权的更准确、全面的句子表示,然后通过更改后的模型级特征融合方法(ML-LSTM)获得具有更为完整、准确问句语义的联合特征表示。其次,通过设置不同大小和不同数量的卷积核,经过卷积、池化、拼接等操作后提取文本混合语义表示中的隐藏特征。最后,利用模型学习到的特征与标签之间的映射关系,实现问句意图标签的预测。

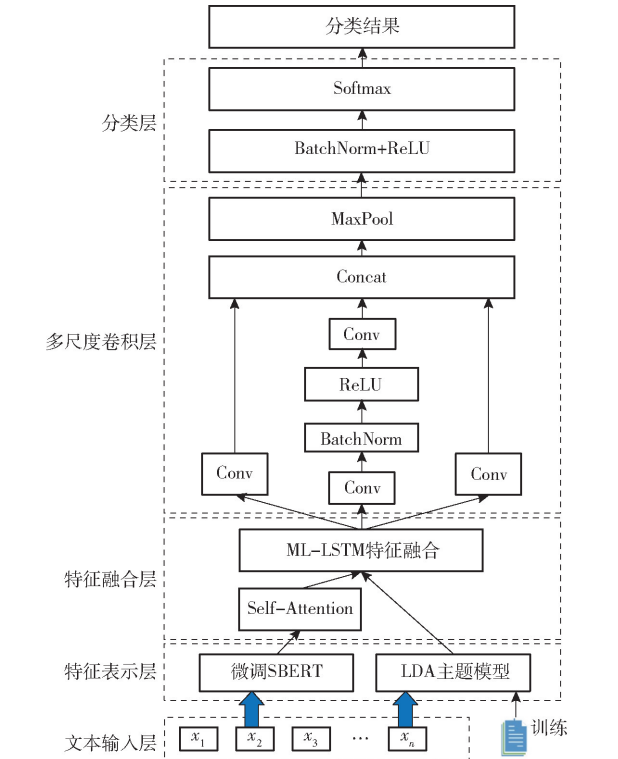


图 2 模型框架
Fig. 2 Model framework

2.1 SBERT 提取语义特征向量

Sentence-BERT (SBERT)^[22] 是普适知识处理实验室在 2019 年为了解决 BERT 生成的句向量不

利于计算句子相似度问题而提出的句子向量计算模型。SBERT 实际是对 BERT 的训练方式进行了改进,通过 Siamese and Triplet Network (孪生网络和三胞胎网络)更新权重参数,引入 Pooling 操作来构建定长的句子嵌入。如图 3 所示的孪生网络中,SBERT 用同一个 BERT 对 2 个输入句子 (Sentence A 和 Sentence B) 进行编码,通过池化后得到句子向量 u 和 v ,并计算出 u, v 的向量差 $|u - v|$,和 u, v 进行拼接后乘以权重矩阵,再通过 Softmax 运算进行分类。SBERT 能够更好地捕捉句子间的关系,获取具有语义特征的句向量。

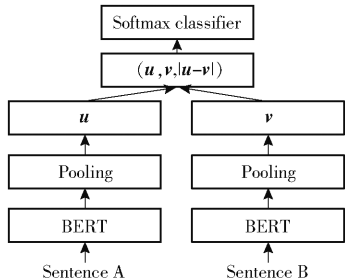


图 3 SBERT 的分类目标函数
Fig. 3 Classification objective function of SBERT

SBERT 采用 3 种不同的池化策略来生成句子的表示,Mean pooling (均值池化)把模型输出的均值向量作为句子的表示;Max pooling (最大池化)即选择每个维度上的最大值,得到整个句子的表示;CLS pooling (CLS 池化)使用模型中 [CLS] 标记的隐藏状态作为句子的表示。实验结果显示,在无监督环境下,获得句子向量最好的方式是 MEAN 策略^[23]。

经过微调训练后的模型使用 Self-Attention 进行权重赋值并不断学习更新,将重要特征信息赋予较大的权重,这些向量通过学习得到的权重矩阵与输入元素线性变换得到。输入序列 $Y = [Y_1 \ Y_2 \ \dots \ Y_h]$,权重矩阵为 W_Q, W_K, W_V 。输入元素查询向量、键向量、值向量计算公式为

$$Q_i = W_Q Y_i \tag{1}$$

$$K_i = W_K Y_i \tag{2}$$

$$V_i = W_V Y_i \tag{3}$$

式中 Q_i, K_i, V_i ——第 i 个元素查询向量、键向量、值向量

对于每一对元素 Y_i 和 Y_r ,计算一个注意力分数来表示 Y_i 对 Y_r 的注意力程度。注意力分数 S_c 计算公式为

$$S_c(Q_i, K_r) = \frac{Q_i \cdot K_r}{\sqrt{d_k}} \tag{4}$$

式中 d_k ——键向量维度
 K_r ——第 r 个元素的键向量

通过 Softmax 函数使注意力分数转换为 0 ~ 1 之间的数值,且和为 1,进而得出注意力权重 w_i 为

$$w_i = \text{Softmax} S_c(Q_i, K_r) \quad (5)$$

最后,将每个元素的值向量与其对应的注意力权重相乘,然后求和输出 Z_i, Z_i 计算式为

$$Z_i = \sum_{r=1}^n w_{ir} V_r \quad (6)$$

式中 Z_i ——第 i 个元素的输出向量

w_{ir} ——位置 i 与位置 r 之间的注意力权重

V_r ——第 r 个元素的值向量

这些权重被用来组合输入的词向量,生成一个新的上下文相关的词向量,使模型更好地理解每个词在具体上下文中的含义,最后将这些输出拼接在一起,形成最终的输出 Z, Z 计算式为

$$Z = W_o \text{Contact}(Y_1, Y_2, \dots, Y_h) \quad (7)$$

式中 W_o ——输出权重矩阵

Contact——拼接操作

Contact 拼接过程如图 4 所示。

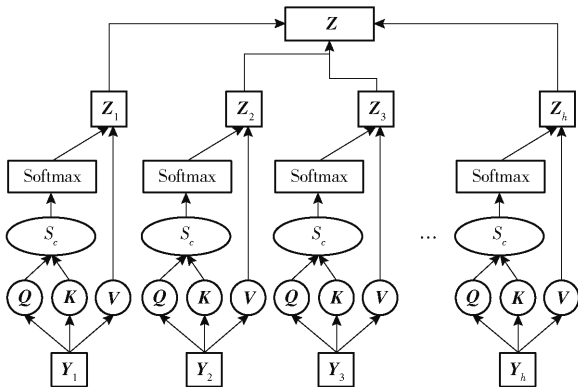


图 4 基于 Attention 的权重更新

Fig. 4 Attention-based weight updates

2.2 LDA 提取概率主题向量

采用潜在狄利克雷分布 (Latent dirichlet allocation, LDA) 获取问句文本的主题特征向量。LDA 模型能够从大量的文本数据中抽取主题特征,帮助理解文本的内容和语义,将文本数据转换为潜在的低维特征表示^[24]。然而,由于烟草问句文本具有上下文信息不足、文本规范性不高等特征,使得文本在文档层面的词语共现时存在稀疏性较高的问题,因此,简单地将主题模型应用于文本分类的效果不佳。为此本文将 SBERT 与 LDA 相结合,即加入文本句子层面的语义相关性信息,更好地理解不同语境下烟草问句文本的语义特征,提高主题的连贯性及细粒度划分的精准度。

LDA 模型本质上是一种贝叶斯概率模型,采用概率估计算法处理文本,在生成文档-主题-词分布之后,计算得到主题特征向量。使用 LDA 主题模型

计算得到文档-主题分布 ($p(z|d)$)、主题-词分布 ($p(w|z)$),然后为每个主题选取概率最大的前 H 的词,计算每个词在其主题下的归一化权重,计算式为

$$\theta_{w_{t,i}} = \frac{p(w_i|z_t)}{\sum_j p(w_j|z_t)} \quad (8)$$

式中 $\theta_{w_{t,i}}$ ——词 i 在主题 t 下的归一化权重

$p(w_j|z_t)$ ——主题 t 中第 j 个词的概率

通过式 (8) 计算得到每个词在主题下的归一化权重,结合词向量通过加权求和得到主题向量 V_{topic_t} 为

$$V_{topic_t} = \sum_{i=1}^H \theta_{w_{t,i}} V_{w_i} \quad (9)$$

式中 V_{w_i} ——词 i 对应的词嵌入向量

LDA 模型计算得到每一个潜在主题与词项的概率表示,对文本提取之后会获得该文本在 10 个主题上概率构成的 10 维向量。

2.3 ML-LSTM 模型级特征融合

SBERT 提取的具有上下文语义信息的特征向量经过具有灵活感知域的 Self-Attention 机制自动学习不同维度特征之间的相关性和重要性,通过计算注意力分数赋予元素不同的权重,并根据输入数据的不同特点动态调整,得到更准确的语义特征信息。与此同时,LDA 主题模型将每个句子向量映射到其概率分布上,来提取句中的潜在主题特征。如图 5 所示,采用多层网络结构逐层提取和抽象不同尺度特征,将富含上下文语义信息特征和主题特征进行融合,提高特征表达能力。

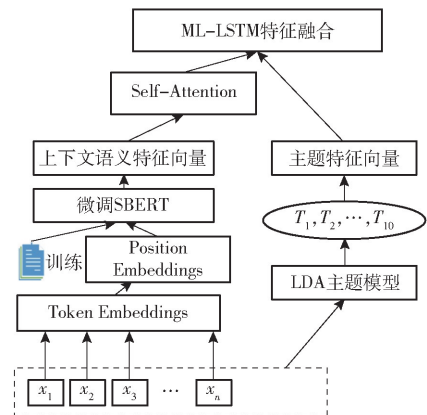


图 5 特征融合流程图

Fig. 5 Feature fusion flowchart

引入 ML-LSTM 模型融合方法来获得语义特征和主题特征的联合特征表示^[25]。ML-LSTM 将多层网络结构与传统的 LSTM 模型相结合,通过充分考虑话语之间的关系在学习过程中处理不同尺度特征融合问题。融合过程如图 6 所示。

将经过 Self-Attention 机制更新权重后获得的更准确的 512 维语义特征向量输入到第 1 层 LSTM

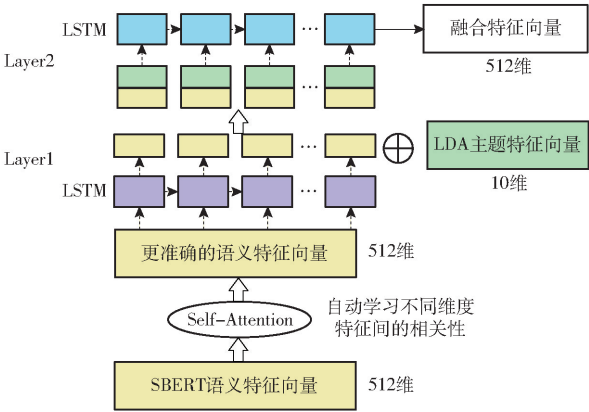


图6 特征融合过程

Fig.6 Generation process of joint feature

(Layer1 隐藏层为 64 维),得到隐藏层状态,然后将 10 维的主题特征向量与 Layer1 隐藏层状态相拼接得到 74 维向量,再输入到第 2 层 LSTM(Layer2 隐藏层为 512 维),得到 512 维联合特征表示。

2.4 基于多尺度 CNN 的特征学习与预测

SBERT 句向量与 LDA 主题向量融合后,得到具有浓缩信息的低维表示,再采用 3 通道 CNN 卷积神经网络^[26]来提取连接后向量的深层次隐藏特征。其结构由输入层、卷积层、池化层和全连接层组成。

2.4.1 输入层

问句意图学习预测模型以基于 SBERT - LDA 特征融合表示的混合语义向量 $V_{SBERT+LDA}$ 为输入,该混合语义包含了主题信息(文本所在领域的全局语义信息)与基于 SBERT 的问句特征向量(结合上下文语义信息)。

2.4.2 多尺度卷积层

卷积层(Convolutional layer)可提取混合语义表示中与类别相关的不同粒度大小的语义特征信息。本文针对烟草领域问句的特点,参考 Inception V1 卷积模块^[27]的思想,设计了如图 7 所示的 Text Inception 卷积模块。该结构通过不同尺度的卷积核学习问句中不同尺度的信息,将这些多尺度特征进行拼接来获得问句混合语义表示中隐藏的特征。Text Inception 有 3 个卷积通道和 3 种尺寸的卷积核,大小分别设置为 1×512 、 3×512 、 5×512 ,通过卷积核感受野的不同,得到不同尺度的特征。其中,第 1 个卷积通道和最后 1 个卷积通道只做 1 层卷积,以避免深层卷积可能导致的信息流失;第 2 个卷积通道设置为 2 层卷积,来获得深层的多尺度高维特征,并在其中使用 BatchNorm^[28]和 ReLU^[29]来加速神经网络的训练速度。Text Inception 模块中所有卷积核的步长均为 1;第 1 通道卷积核尺寸为 1×512 ,数量为 256,padding 为 0;第 2 通道中第 1 层卷积核尺寸为 3×512 ,数量为 256,padding 为 1,第 2

层卷积核尺寸为 5×512 ,数量为 256,padding 为 2;第 3 通道卷积核尺寸为 3×512 ,数量为 256,padding 为 1。将 3 个通道的特征拼接在一起可输出一个 768 维的问句向量。

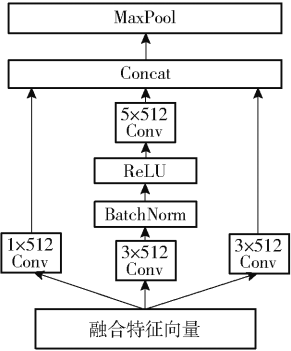


图7 3层通道卷积模块

Fig.7 3-layer channel convolution module

2.4.3 池化层

池化层(Pooling layer)通过对输入特征图中的局部区域进行池化(Pooling)操作,从中提取最显著的特征并降低空间维度。

2.4.4 全连接层及损失函数

在经过多层卷积、池化后,使用全连接层(Fully connected layer)进行特征拼接,并将拼接后的特征向量映射到神经元,再通过 Softmax 函数将该特征转换为不同意图类别的概率,概率计算式为

$$f(k_j|q,w) = \frac{\exp(h_i)}{\sum_{i=1}^n \exp(h_i)}$$

(10)

- 式中 j ——问句分类标签
 q ——问句
 w ——模型的训练参数
 k_j ——类别标签
 h ——模型的隐藏层特征
 n ——隐藏层特征的数量

模型采用交叉熵损失函数计算网络损失值,通过反向传播进行迭代训练,最后利用自适应矩估计梯度优化算法^[30](Adaptive moment estimation, Adam)进行优化,损失值 L 计算公式为

$$L = -\frac{1}{N} \sum_i \sum_{j=1}^D y_{ij} \ln e_{ij}$$

(11)

- 式中 D ——问句分类的标签数量
 N ——训练集中的样本数量
 y_{ij} ——真实问句意图标签
 e_{ij} ——模型预测的问句意图标签

3 实验与结果分析

3.1 实验环境

实验软硬件环境如下:操作系统为 64 位

Windows 10 家庭版,处理器为 AMD Ryzen 7 5800U, 8 个物理核心和 16 个线程可用于并行处理任务,基准时钟频率为 1.9 GHz,最高可提升至 4.4 GHz。语言环境为 Python 3.9, SBERT 建模环境为 Pytorch 3.9, LDA 建模环境为 Ubuntu 16.04。

3.2 性能评价指标

为了对提出的基于 SBERT - Attention - LDA 特征融合的问候意图识别方法的性能进行客观评价,选择精确率 (Precision, P)、召回率 (Recall, R)、F1 值 (F1 score) 作为评价指标。

3.3 消融实验

基于 SBERT - Attention - LDA 与多层次 LSTM 特征融合的问候意图识别模型包含 2 个消融因素,即特征融合结构与 3 通道 CNN 神经网络结构。为了验证本文模型的有效性,消融实验在烟草问句数据集上构建 4 个网络模型,4 个模型在烟草问句数据集上的问候意图识别实验结果如表 2 所示。

表 2 消融实验结果			
模型	精确率	召回率	F1 值
SBERT - LDA	92.34	91.90	92.12
CNN	92.76	92.66	92.71
SBERT - Attention - LDA	93.42	92.54	92.97
SBERT - Attention - LDA - CNN	94.81	94.36	94.58

由表 2 可知, SBERT - Attention - LDA - CNN 的准确率为 94.81%, 分析可知, 将网络训练过程中不同层次的特征信息进行提取处理, 分类准确度有所提升, 但没有达到最优的效果。提取不同层次的特征信息, 但不对其进行处理, 加权融合后的分类准确度不升反降。当对不同层次的特征信息进行处理并加权融合后, 分类精度明显超过基准网络, 并且达到了最优精度。可见, 本文问候意图识别模型中的多尺度 CNN 神经网络结构和特征融合结构单独使用, 均不能达到最优分类效果, 只有将两者相结合, 在烟草问句数据集上的分类精度才能达到最优。

3.4 对比实验

为验证本文模型的有效性, 在同一实验环境下选择多个模型进行对比实验, 为了确保对比的公平性, 对每个模型进行了参数优化设置, 具体如下:

(1)BERT:采用基于词级别和字符级别的混合编码,在 BERT 模型的最后一个编码器层的输出上应用全局平均池化,通过激活函数 ReLU 和权重矩阵的线性变换生成分类标签的分数,最后通过 Softmax 函数进行分类。

(2)TextCNN:经典的文本卷积分类算法,实验

采用 Word2vec 作为词向量表示问句,采用尺寸 (2,3,4) 的卷积核进行局部特征提取。

(3)LDA - CNN - Attention:使用 LDA 模型计算得到主题向量矩阵输入,不考虑逆主题空间频率,采用尺寸 (3,4,5) 的卷积核进行局部特征提取,通过全连接层和 Softmax 层实现问句分类。

(4)ERNIE - CNN:模型采用 ERNIE 对问句进行语义编码,然后使用尺寸为 (2,3,4) 的卷积核进行局部特征提取,使用 Adam 优化器进行模型优化,学习率设置为 0.001。

(5)Bi - GRU - Attention:采用 2 层 Bi - GRU 可提高模型的表达和建模能力,设置 Dropout 参数为 0.3,提高抗模型抗过拟合风险,然后通过注意力机制将 2 个方向上的表示进行加权融合。

(6)Self - Attention:Self - Attention 模型中每个词与文本中所有词均进行 Attention 计算来捕获长距离依赖关系;使用 2 层 Self - Attention, Dropout 参数设置为 0.3。

3.5 实验结果与评价

3.5.1 烟草问句测试集结果

将所提模型与对比的 6 种模型在训练集上进行训练。所提模型的训练批量大小 (batch_size) 为 32, 当迭代周期 (epoch) 为 14 时,出现过拟合趋势,遂停止训练。其他模型也都以相同的训练批量进行训练,并利用早停机机制监测,当出现过拟合时及时停止训练。再将训练好的各个模型作用在测试集上,得出对应的精确率、召回率和 F1 值,如表 3 所示。

表 3 不同方法对比实验结果			
模型	精确率	召回率	F1 值
BERT	92.22	92.55	92.51
TextCNN	90.20	90.35	90.25
LDA - CNN - Attention	89.54	87.65	88.62
ERNIE - CNN	92.32	91.90	91.70
Bi - GRU - Attention	86.56	86.08	86.33
Self - Attention	90.13	89.26	89.68
本文模型	94.81	94.36	94.58

根据表 3 可知, 与注意力机制模型 (Self - Attention)、深度学习模型 (TextCNN)、两者相结合模型 (Bi - GRU - Attention), 以及引入 LDA 特征的深度学习模型 (LDA - CNN - Attention) 相比, 本文模型在精确率、召回率和 F1 值上均具有较为显著的优势。Self - Attention 模型主要使用自注意力机制, 能够充分考虑问句中的词与词之间的依赖关系, 但在处理较长的问句时可能会面临长距离依赖捕捉的难题, 从而影响准确率。Bi - GRU - Attention 模型使

用了双向 GRU 和 Attention 机制,虽然能够捕捉问句中的上下文信息,但在某些情况下可能无法充分捕捉问句的语义特征,导致准确率相对较低。

与单一预训练模型 (BERT、ERNIE - CNN) 相比,本文模型性能也有明显提升,F1 值分别提高 2.07、2.88 个百分点。BERT 模型和 ERNIE - CNN 模型都能够较好地捕捉问句中的词汇和上下文语义关系,但是信息的来源都局限于问句文本,无法从问句出现的概率角度获取更多的辅助分类信息。而本文所提模型同时关注了 2 个信息源,一方面通过 SBERT 对问句文本进行句向量编码时能够获取语义信息更加突出的特征,另一方面使用 LDA 模型获取的主题信息强化问句类型信息,而基于缩放自注意力机制的权重计算对两部分的特征向量实现合理组合,进而提升了模型识别能力。

3.5.2 模型效率

7 种模型参数量以及在烟草数据集上的训练和测试响应时间对比结果如表 4 所示。

表 4 训练和测试响应时间

Tab.4 Training and testing response time

模型	训练时间/s	测试时间/s	参数量
BERT	6 156.28	15.22	$1.162\,4\times10^8$
TextCNN	161.90	2.30	$3.321\,7\times10^6$
LDA - CNN - Attention	201.07	5.35	$3.872\,3\times10^7$
ERNIE - CNN	6 631.16	18.77	$1.280\,3\times10^8$
Bi - GRU - Attention	213.54	4.39	$3.197\,6\times10^7$
Self - Attention	156.12	2.08	$2.137\,7\times10^7$
本文模型	6 918.93	20.62	$1.332\,5\times10^8$

由表 4 可知,Self - Attention 模型和 TextCNN 模型的训练和测试响应时间皆较低,其中 TextCNN 模型的参数量较小,收敛速度快,Self - Attention 模型主要计算序列数据中的自身注意力,结构相对简单、Bi - GRU - Attention 模型需要执行前向传播和反向传播,LDA - CNN - Attention 模型具有更复杂的结构,参数量比 Bi - GRU - Attention 模型增加 21.09%,但两模型参数量都比 Self - Attention 模型大,导致两模型训练和测试响应时间相比略长。BERT、ERNIE - CNN 和本文模型皆具有预训练模型,进行动态向量的表征,因此参数量较大,分类模型的训练和测试响应时间相对较长。本文模型在训练和测试响应时间上多于对比模型,但烟草问句的识别效果明显更好。

3.5.3 问句类别

为了更好地展示各个模型的具体差异,本研究分别计算不同模型在各个类别下对应的 F1 值,如图 8 所示。实验结果显示,相对于其他类别,各模型在烟草种类和品牌问题与烟叶分级问题上表现略差,由前文主题聚类效果可知,2 个主题气泡距离较近,说明主题间差异性较小,较难区分。此外,除了烟叶分级问题,本文模型在其他类别上分类的 F1 值均最高,表明该模型整体分类效果优于其他对比模型,该模型在烟草市场与经济问题、烟草生产与加工、烟草种植问题中的 F1 值明显优于其他模型,说明本文模型在数据量较少的情况下仍能充分学习到文本的语义信息、提取问句特征进行有效分类。但是,整体来看,针对数据量较小的问句类型,几种模型的 F1 值均明显低于数据量大的问句类型,说明数据类别的不平衡在一定程度上影响了深度学习模型的分类效果。

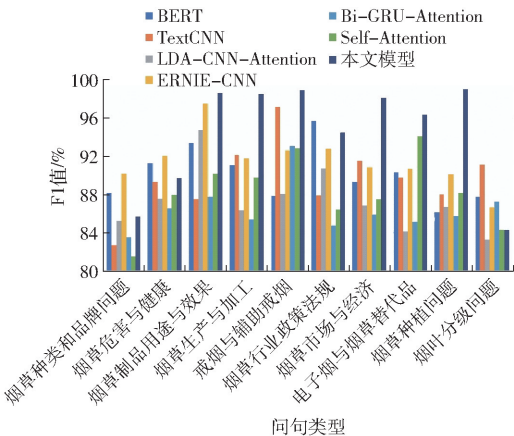


图 8 7 种模型对不同类别问句分类的 F1 值对比
Fig.8 Comparison of F1 score of seven kinds of models for different types of question classification

4 结束语

提出了一种基于 SBERT - Attention - LDA 与 ML - LSTM 特征融合的烟草问句意图识别的深度学习模型。该模型将烟草问句文本经 SBERT 模型进行动态词向量编码后得到的特征向量和经 LDA 提取的主题概率向量通过 ML - LSTM 进行特征融合,再通过多尺度卷积神经网络做进一步的特征提取后进行分类识别。验证实验表明,模型能够有效应对烟草问句文本较短、特征稀疏、隐含语义信息较难学习等问题,在测试集中本文模型 F1 值达到 94.58%,相比基线模型显著提高了烟草问句意图识别性能。

参 考 文 献

[1] 苏雪峰. 结合注意力和双向 LSTM 的开放域问句分类研究[J]. 办公自动化,2023,28(3):6 - 10.
SU Xuefeng. Research on open-domain interrogative sentence classification combining attention and bidirectional LSTM[J].

- Office Automation,2023,28(3):6-10. (in Chinese)
- [2] 吴华瑞,郭威,邓颖,等. 农业文本语义理解技术综述[J]. 农业机械学报,2022,53(5):1-16.
WU Huarui, GUO Wei, DENG Ying, et al. Review of semantic analysis techniques of agricultural texts[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022,53(5):1-16. (in Chinese)
- [3] NAEIM R, GALINA N. Persian text classification using naive bayes algorithms and support vector machine algorithm[J]. Indonesian Journal of Electrical Engineering and Informatics,2020,8(1):178-188.
- [4] ALAN J. Some reflections on maxent[J]. Civil Engineering and Environmental Systems,2023,40(2):60-71.
- [5] WANG Liming, SU Peilin, HUANG Di. A cross-spectral target tracking method based on a hidden Markov model[J]. Journal of Physics: Conference Series,2023,2419(1):20-31.
- [6] SANDRA A, LI Yuxiang, QIU Tingting, et al. Estimating the probability of peanut allergy development in the avoidance arm of LEAP using classification and regression tree analysis[J]. The Journal of Allergy and Clinical Immunology,2023,151(2):172-173.
- [7] 戴贵洋,蔡秀丽,余晓晗. 融合人类知识的随机森林特征选择方法研究[J]. 计算机技术与发展,2022,32(7):155-160.
DAI Guiyang, QI Xiuli, YU Xiaohan. Research on random forest feature selection method integrating human knowledge[J]. Computer Technology and Development,2022,32(7):155-160. (in Chinese)
- [8] GAN S F, SHAO S, CHEN L, et al. Adapting hidden naive Bayes for text classification[J]. Mathematics,2021,9(19):2377-2378.
- [9] 淦亚婷,安建业,徐雪. 基于深度学习的短文本分类方法研究综述[J]. 计算机工程与应用,2023,59(4):43-53.
GAN Yating, AN Jianye, XU Xue. A research review on short text classification methods based on deep learning[J]. Computer Engineering and Application,2023,59(4):43-53. (in Chinese)
- [10] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv Preprint:1408.5582,2014.
- [11] 史梦飞,杨燕,贺樑,等. 基于 Bi-LSTM 和 CNN 并包含注意力机制的社区问答问句分类方法[J]. 计算机系统应用,2018,27(9):157-162.
SHI Mengfei, YANG Yan, HE Liang, et al. A community Q&A question classification method based on Bi-LSTM and CNN including attention mechanism[J]. Computer System Applications,2018,27(9):157-162. (in Chinese)
- [12] DUAN S F, ZHAO H. Attention is all you need for Chinese word segmentation[J]. CoRR:abs/1910.14537,2019.
- [13] 王郝日钦,吴华瑞,冯帅,等. 基于 Attention_DenseCNN 的水稻问答系统问句分类[J]. 农业机械学报,2021,52(7):237-243.
WANG Haoriquin, WU Huarui, FENG Shuai, et al. Classification technology of rice questions in question answer system based on Attention_DenseCNN[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021,52(7):237-243. (in Chinese)
- [14] MANEL A, CHIRAZ L. BE-BLC: BERT-ELMO-based deep neural network architecture for English named entity recognition task[J]. Procedia Computer Science,2021,192:168-181.
- [15] SHINSHO O. Generation of long- and short-range potentials from atom-molecules to quark-gluon systems by the GPT potential[J]. Journal of Physics Communications,2022,6(1):34-56.
- [16] VASUDEVAN N, FILIP K, JOHN K. Shapley idioms: analysing BERT sentence embeddings for general idiom token identification[J]. Frontiers in Artificial Intelligence,2022,5(1):5006-5009.
- [17] 李林,刁磊,唐詹,等. 基于 BERT_Stacked LSTM 的农业病虫害问句分类方法[J]. 农业机械学报,2021,52(增刊):172-177.
LI Lin, DIAO Lei, TANG Zhan, et al. Question classification method of agricultural diseases and pests based on BERT_Stacked LSTM[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021,52(Supp.):172-177. (in Chinese)
- [18] 郑丽敏,乔振铎,田立军,等. 基于 BERT-LEAM 模型的食品安全法规问题多标签分类[J]. 农业机械学报,2021,52(7):244-250,158.
ZHENG Limin, QIAO Zhenduo, TIAN Lijun, et al. Multi-label classification of food safety regulatory issues based on BERT-LEAM[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021,52(7):244-250,158. (in Chinese)
- [19] 阮光册,黄韵莹. 融合 Sentence-BERT 和 LDA 的评论文本主题识别[J]. 现代情报,2023,43(5):46-53.
RUAN Guangce, HUANG Yunying. Fusion of Sentence-BERT and LDA for topic recognition of commentary text[J]. Modern Intelligence,2023,43(5):46-53. (in Chinese)
- [20] NADEEM A, MUZAKKIR H. Extractive multi-document summarization using relative redundancy and coherence scores[J]. Journal of Intelligent and Fuzzy Systems, 2020, 38(5):1-12.
- [21] 武帅,施奕,杨秀璋,等. 基于社交网络分析和 LDA 主题挖掘的短文本挖掘研究[J]. 现代电子技术,2022,45(20):124-128.
WU Shuai, SHI Yi, YANG Xiuzhang, et al. Research on short text mining based on social network analysis and LDA topic mining[J]. Modern Electronic Technology,2022,45(20):124-128. (in Chinese)
- [22] REIMERS N, GUREVYCH I. Sentence-Bert: sentence embeddings using siamese bert-networks[J]. arXiv Preprint:1908.10084, 2019.
- [23] CHENG H, LIU S, SUN W, et al. A neural topic modeling study integrating SBERT and data augmentation[J]. Applied Sciences,2023,13(7):4589-4595.
- [24] SAHAN J M, ABBAS E I, ABOOD Z M. A facial recognition using a combination of a novel one dimension deep CNN and LDA[J]. Materials Today: Proceedings, 2023, 80 (P3):3594-3599.
- [25] YANG Fan, SANG Yongsheng, CAO Jun, et al. Prediction of gasoline yield in fluid catalytic cracking based on multiple level LSTM[J]. Chemical Engineering Research and Design,2022,185:119-129.
- [26] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015:1-9.
- [27] JAHANDAD K, SAM M S, KAMARDIN K, et al. Offline signature verification using deep learning convolutional neural network (CNN) architectures GoogLeNet Inception-v1 and Inception-v3[J]. Procedia Computer Science,2019,161(C):475-483.
- [28] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//International Conference on Machine Learning,2015:448-456.
- [29] STEFFEN D, SEBASTIAN K. On minimal representations of shallow ReLU networks[J]. Neural Networks,2022,148:121-128.
- [30] KINGMA D P, BA J. Adam: a method for stochastic optimization[J]. arXiv Preprint, arXiv:1412.6980,2014.