基于小样本短句意图识别算法研究

王炳翔 [1]
WANG Bingxiang

摘 要 随着 AI 技术的飞速发展，智能问答系统已逐渐融入了公众生活。针对面向中小企业智能问答系统开发所涉及的小样本短句意图识别问题，文章提出了一种应用 AC 自动机搭配人工规则的基于规则模板算法和应用 TextRank 搭配 TF-IDF 优化分类器训练过程而改进的朴素贝叶斯算法，两种算法均可以解决语料样本不充足时对较短语句的意图识别任务。通过实验和分析，表明两种算法都可以为中小企业智能问答系统的研发提供一种合理实现途径，有一定的实践价值。

关键词 小样本；短句；算法；规则；意图识别

doi：10.3969/j.issn.1672-9528.2025.02.038

0 引言

开发面向中小企业的智能问答系统，在进行用户问句意图识别时会遇到这样的实际情况：（1）甲方单位的客户想用尽可能短的表达去获更丰富的回复，即用短问句（字数≤20）咨询业务，有时甚至仅用关键词；（2）甲方单位无法提供咨询问题的基础数据样本，即便有也非常稀少；（3）甲方单位拥有的硬件设施资源很有限，且配置普遍不高；（4）甲方单位要求被开发的系统必须是轻量级、易部署、维护简单、性价比高。

针对上述情况，解决关键即在满足上述需求前提下，找到最合适针对小样本短句意图识别的算法，因此本文提出两种不同形式的算法作为解决方案以供参考。

1 基于规则模板的意图识别算法

1.1 Aho-Corasick automaton 算法

对于 NLP（natural language processing）并非所有任务都要首选大模型。尤其在深度学习飞速发展的当下，一些轻量级算法在不同需求场景的实践项目中仍占有一席之地。例如，AC 算法（aho–corasick automaton）[1]，适用于处理多个模式串匹配一个主串的情况，擅长快速查找特定的关键词，在处理网络文本内容检测、敏感词过滤、病毒名扫描等任务中表现优异。其原理是通过构造一个称为 Trie 的树形结构和 fail 指针来实现，整个算法具体可分为构建 Trie 树（goto 表）、创建 fail 指针、进行搜索（创建 output 表）三部分 [2]。图 1 以“你单位、你公司、公司业务”为模式串举例，构造 AC 自动机形态，其中虚线表示 fail 指针的指向，数组表示模式串的长度。

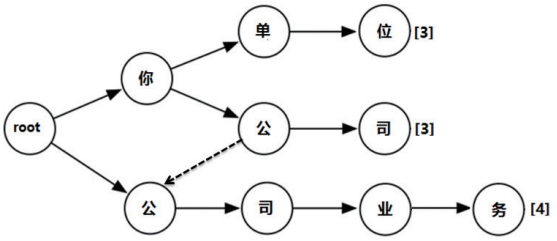

图 1 AC 自动机示意图

1.2 制定规则模板

规则模板是根据问句的结构特点人工设定的一系列规则集，设定依据是具体的业务流程，如果输入的文本经过问句解析后恰好符合某些规则，则给该问句依次打上规则对应的类别标签，并形成标签集。标签即识别出的问句意图，通常一条规则对应一个标签，但一个问句可能会对应多个规则。进行规则匹配的方法有很多，例如可以建立不同类别的规则词典，在每个词典内定义一些关键词，然后用 AC 自动机对照词典进行抓取，算法流程如图 2 所示。

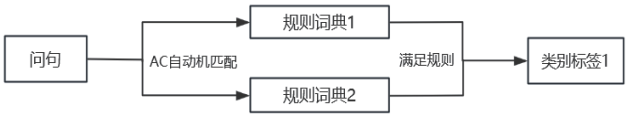

图 2 基于模板分类算法流程图

例如问句：网站维护的具体内容是什么？

规则词典 1 = { 系统集成，网站维护，软件开发…}

规则词典 2 = { 内容，简介，方案，步骤…}

用 AC 自动机对问句进行关键词匹配，该问句满足规则词典 1 和规则词典 2，即被标记为 product_content。

2 改进的朴素贝叶斯意图识别算法

2.1 TextRank 算法

TextRank 是一种基于图的排序算法，主要用于文本关键

1. 西安石油大学计算机学院 陕西西安 710000

词提取和摘要生成，其本质是对 PageRank 算法[3]的改进，用词为节点词来代替网页节点，用词语之间的共现关系建立节点之间的链接，进而构建整个无向有权图网络。例如分词后的文本“我\公司\有\软件开发\业务”，设窗口长度为 2 个单位，按顺序依次向后滑动，则构建为图 3 所示。

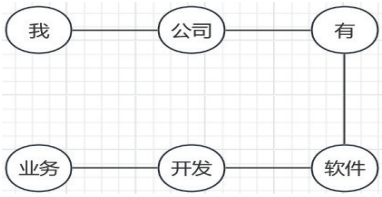

图 3 TextTank 关系示意图

TextRank 算法节点权重的数学公式（1）是基于 PageRank 公式的改进。

$$W(V_i)=(1-d)+d\times\sum_{j\in \mathrm{In}(V_i)}\frac{W_{ji}}{\sum_{V_k\in \mathrm{Out}(V_j)}W_{jk}}W(V_j) \quad (1)$$

式中：d 是阻尼系数，一般设置为 0.85，表示图中某节点指向其他任意点的概率；W_{ji} 表示权重；$\mathrm{In}(V_i)$ 与 $\mathrm{Out}(V_j)$ 表示节点集合。

与其他关键词提取算法相比，TextRank 的一个显著特点就是不需要基于某个现成的语料库进行样本训练，其算法原理简单，无须监督，计算便捷，易部署和维护，关键词提取效率高，但结果会受到文本预处理影响（如分词或停用词选取）。在 NLP 任务中，TextRank 结果表现和分词工具（如 Jieba、HanLP 等）有异曲同工之效，特别是处理短句时，特征提取能力更强。下面以 Jieba 分词工具为例，对比一下两种方式的特征向量处理结果：

包含 254 个中文字符语料库中的问句：“能简单介绍一下你们单位的基本情况吗？”。

用 Jieba 分词工具处理结果为：能\简单\介绍\一下\你们\单位\的\基本\情况\吗\？

BOW[4] 向量维度为 79；其中零元素占比为 91.14%。

用 TextRank 处理结果为：简单\介绍\单位\一下\基本\你们\情况

BOW 向量维度为 66；其中零元素占比为 89.39%。

通过实验结果可以观察到，同样在没有停用词表参与的前提下，TextRank 不仅对问句特征的捕捉更精准，而且还可以降低文本向量的维度，这有利于简化向量运算复杂度以及节省存储空间。

2.2 朴素贝叶斯分类算法

朴素贝叶斯分类器（naive bayesian，NB），其公式为：

$$h_{\mathrm{nb}}(x)=\arg\max_{c\in y}P(\mathrm{c})\prod_{i=1}^{d}P(x_i|c) \quad (2)$$

简单易理解，分类效率稳定，支持增量式训练，能处理多分类问题，尤其是在处理小样本分类任务时有较好表现。但由于联合概率运算满足乘法交换律，因此在处理有关文本语序问题时该算法表现不佳，所以要处理文本是短句且对基于语序的语义理解要求不高的任务，NB 算法无疑是很好的选择，否则需要考虑基于深度学习的算法。

若某个样本属性值没有出现在训练集中，即发生概率值为零的情况，必然会导致最终分类的准确度。为避免这种情况，估计概率值时通常使用“拉普拉斯修正”进行平滑处理，即令 N 表示训练集 D 中可能的类别总数，N_i 表示第 i 个属性可能的取值数，则式（2）变量可修正为：

$$\hat{P}(c)=\frac{|D_c|+1}{|D|+N} \quad (3)$$

$$\hat{P}(x_i|\mathrm{c})=\frac{|D_{c,x_i}|+1}{|D_c|+N_i} \quad (4)$$

实验表明在训练集变大时，普拉斯修正过程引入先验的影响也会逐渐变得可忽略[5]，会使估值渐趋向真实概率值。

2.3 算法的改进

识别句子意图本质就是对句子进行分类，针对短句首选朴素贝叶斯算法。现对该算法分类器的常用训练过程进行局部改进，即用 TextRank 替换分词工具（如 Jieba）对训练样本进行预处理，用 TF-IDF 算法替换 BOW 模型进行词频统计，改进的分类器训练流程如图 4 所示。

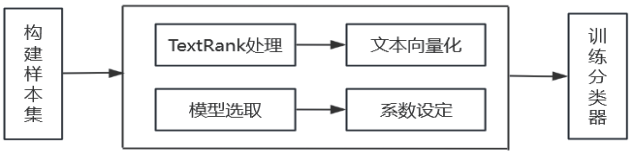

图 4 贝叶斯分类器训练流程图

2.3.1 构建样本数据集

由于甲方公司没有现成的样本集，所以需要手工构建问句样本，并全部进行标记。以售前业务为例，如图 5 所示，其中 X_i（i=1,2,3…）为样本变量 question，Y_i（i=1,2,3…）为针对 question 标记的类型，desc 为对类型的具体描述。

	x：question	y:type		desc
x1	你们单位在哪里	1	y1	公司地址
x2	你们单位的公司地址在哪里	1	y1	公司地址
x12	你们单位主要做什么	2	y2	公司产品
x13	贵单位有什么产品	2	y2	公司产品
x22	你们公司有座机吗	3	y3	公司电话
x23	你们单位的销售电话是多少	3	y3	公司电话
x31	你们公司的网站地址是什么	4	y4	公司网站
x32	贵公司的官网地址是	4	y4	公司网站
	……			

图 5 部分训练样本集示例

2.3.2 设置训练参量

（1）使用 TextRank 算法对训练样本进行分词处理。

（2）使用 TF-IDF 算法将处理后文本向量化，它在统计词频方面要优于 BOW。

（3）使用 sklearn 包中的 MultinomialNB 模型，该模型适用于标记为多分类任务的离散型特征变量。

（4）拉普拉斯（Laplace）平滑系数 alpha 设定，通常 alpha 设置为 1.0。alpha 越小，迭代次数越多，精度越高，当 0＜alpha＜1 时，Laplace 就会过度为 Lidstone 平滑 [6]。

2.3.3 训练分类器

根据构建好的样本和相关参量训练贝叶斯分类器，核心代码（Python）与注释示例如下：

```
def NBclassifier (self, fileName,stopWordfileName,
question): # 定义分类器
    fileDatas = pd.read_excel(fileName)# 加载训练数据
x_train = fileDatas.apply(self.ByTextRank)
    # 使用 TextRank 算法提取特征并分词
    vector = self.createTf_idfVec(stopWordfileName)
    # 使用 TF-IDF 算法文本向量化，并加入停用词库
    x_train_vec = vector.fit_transform(x_train)
nb = MultinomialNB(alpha = 1.0 ) # 贝叶斯模型选取
与平滑系数设定
y_train = fileDatas.type  # 获取标记
    nb.fit(x_train_vec, y_train)  # 数据拟合
    …
    result = nb.predict(x_vec)  # 预测验证数据
    return result    # 得到训练好的分类器
```

3 算法对比与评价

度量和评价算法的性能，要在统一量纲下设置合适的评价指标。在分类任务中，通常利用精度，如式（5）所示，

$$\mathrm{acc}(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(x_i)=y_i) \quad (5)$$

和错误率 (1- acc(*f*; *D*)) 作为模型的性能度量，式中 *D* 为样例集。精度可以衡量预测结果中有多少是正确的，但在数据类别分布极不均衡的情况下，不能客观地评价模型的性能，例如测试集中有 100 个测试样本，99 个是正样本，1 个是负样本，假设某模型预测整个测试集都是正样本，虽然数值达到了 99% 的精度，但实际上却很难评估这个模型确切的分类能力。为了优化评价指标，基于二分类问题，引入分类结果的混淆矩阵，根据混淆矩阵，可以定义出准确率（即查准率，可以反映误检程度）、召回率（即查全率，可以评估漏检程度）和 F_1 值（准确率和召回率的调和平均）等三个评价指标 [5]。

对于多分类问题，可以将问题分解为多个二分类子问题来解决，即对于任意类别，本类为正例，非本类均为反例。短问句除去主谓宾等基础字符外，尽管问题类别具有多样性，但语料内容并不丰富，对于同一个意图而言，样本无外乎通过句式重构、同义转换等方式进行扩充，例如“现在几点了？”可以调整语序为“几点了现在？”，或用“时间、时候”等词做同义词替换。所以对于字符很少的句子，少量样本理论上是可以满足模型对文本重要特征的学习。下面进行基于小样本分类算法的性能实验，这里以甲方实际业务中具有代表性的 8 种问题类别为例，进行分类器的训练，其样本分布和表示方式如表 1 所示。

表 1 分类样本数据类型示例表

y	类型	类型表示	样本示例	数量
1	地址	address	你们单位在哪里	10
2	产品	product	贵单位有什么产品	10
3	电话	phone	你单位有电话吗	10
4	网址	website	你公司的网站是什么	10
5	资质	qualific	你公司有什么资质	10
6	操作	act_desc	平台课程退费如何操作	30
7	处理	act_deal	平台不能打印证书怎么办	30
8	问候	other_talk	现在几点了	10

依据总样本集（图 5 例），本文按 7∶3 的比例划分训练集和测试集进行实验，为了更好地反应出模型的泛化能力，测试集特选取了分布均匀且特征典型的数据。算法模型的性能度量，可以采用精度、准确率、召回率以及 F_1 值作为评价指标，现将朴素贝叶斯算法（Jieba+BOW）和改进的朴素贝叶斯算法（TextRank+TF-IDF）以及基于规则模版的算法使用统一的测试数据进行比较，实验结果如表 3 所示。

表 2 分类算性能指标对比

算法	精度	准确率	召回率	F1 值
朴素贝叶斯算法	83.33%	82.76%	85.71%	84.21%
改进的朴素贝叶斯算法	87.04%	86.21%	89.29%	87.72%
基于规则模版的算法	98.15%	/	/	/

通过上面指标数据可知，改进的 NB 算法在性能上比改进前有所提高，且 F_1 值均围绕在 85%，说明两种分类器都可以解决基本的分类需求。基于统计学习的算法分类能力强依赖于训练集，虽然该样本数据是由手工构建，不可避免地会带有主观偏向性，但有如此表现对于处理这种短句的意图识别任务还是完全够用的。同时，基于规则模板算法，只要规则存在，其精度都会在 98% 以上，而且该算法可以在同一个句子捕捉到符合条件的多个类别标签，例如问句“请问你们公司的法人、地址、电话都是什么？”，算法则会构建标签集并组合成最终的答案，这在答案容忍性上有较好效果。

综上所述，基于规则模板的算法可维护性的灵活度更强，对于问句意图判断，不需要构造训练集，应对新任务仅添加新规则即可。反观贝叶斯分类算法如果去处理一句多问的情况，则必须先收集足量的样本，再通过训练才能获取对应的类别，想凭借原始代码的复用减少工作量，显然是办不到的。所以通过上述分析，结合实际的业务需求和语料特点，在问句意图识别算法的选取上，基于规则模板的算法性价比和适用性会更强一些。当然，如果后期在进行问答系统开发的时

（下转第 167 页）

4.4 对比实验

实验中，将本文提出的方法作为方法 1，基于改进 KNN 算法作为方法 2，基于改进卷积神经网络方法作为方法 3，对比结果如表 2 所示。

表 2 不同方法在文本自动分类中的性能对比

类别名称	分类精度			Kappa 统计量			汉明损失		
	方法 1	方法 2	方法 3	方法 1	方法 2	方法 3	方法 1	方法 2	方法 3
新闻	0.81	0.68	0.72	0.56	0.39	0.45	0.14	0.22	0.19
科技	0.89	0.76	0.85	0.78	0.61	0.72	0.06	0.11	0.08
娱乐	0.74	0.59	0.63	0.42	0.27	0.31	0.21	0.29	0.26
体育	0.91	0.82	0.88	0.83	0.70	0.77	0.04	0.09	0.07
教育	0.61	0.55	0.48	0.25	0.18	0.09	0.31	0.34	0.41
健康	0.84	0.78	0.71	0.73	0.60	0.54	0.09	0.12	0.18
法律	0.93	0.87	0.90	0.86	0.79	0.81	0.02	0.04	0.05

根据表 2，本文提出的文本分类方法相较于其他方法展现出显著优势。在新闻、科技、体育及法律等类别上，其分类精度与 Kappa 统计量均达高水平，表明分类效果稳定且准确。从汉明损失看，本文方法损失值低，进一步验证其优越性。该方法采用遗传算法优化 SVM 参数，自动搜索最佳配置，显著提升分类性能，使模型表现更稳定可靠。

5 结语

本文提出了一种创新的文本自动分类方法，该方法结合了遗传算法与支持向量机的优势。借助遗传算法的全局搜索和优化能力，成功解决了 SVM 参数调优的复杂问题，从而显著提升了文本分类的准确性。此方法在处理大规模文本数据时表现优异，为文本自动分类领域带来了新的视角和技术革新。展望未来，将进一步深化遗传算法与 SVM 结合机制的研究，探索更多创新性优化策略，以期提升分类模型的性能和稳定性。

参考文献：

[1] 潘国炀 . 基于改进 KNN 算法的档案信息文本自动分类方法研究 [J]. 信息与电脑 (理论版), 2024, 36(4): 71-73.
[2] 刘影 , 余进 , 陈莉 . 基于改进卷积神经网络的多标签文本自动化分类研究 [J]. 自动化与仪器仪表 ,2023(11):62-66.
[3] 李淑红 , 邓明明 , 孙社兵 , 等 . 基于注意力机制和 CNN-BiLSTM 模型的在线协作讨论交互文本自动分类 [J]. 现代信息科技 , 2023,7(13):26-31.
[4] 索南多杰 , 官却多杰 , 拉玛杰 , 等 . 基于深度学习的藏文文本自动分类研究 [J]. 青海科技 ,2023,30(3):192-196.
[5] 刘蕾 , 田鑫宇 , 朱大洲 . 基于 SSA-SVM 的营养健康信息文本分类研究 [J]. 计算机时代 ,2023(6):82-86.
[6] 徐涯昕 , 何泽恩 , 徐绪堪 . 基于 CNN-BiLSTM 网络的数控机床故障文本自动分类 [J]. 计算机与现代化 ,2023(4):7-14.
[7] 陈玉天 , 陈洋 , 梁恒瑞 , 等 . 基于 TI-LSTM 的文本自动分类算法及应用 [J]. 长春理工大学学报 (自然科学版), 2023, 46(1):130-136.

【作者简介】

胡翔（1980—），女，安徽马鞍山人，硕士，讲师，研究方向：计算数学。

（收稿日期：2024-10-12）

（上接第 163 页）

候，完全可以基于规则模板的算法为主，贝叶斯算法为辅助（作为补充算法来解决某些特殊的语义理解问题），以保障系统整体的健壮性。

4 结语

本文为解决基于小样本短句意图识别问题，设计了基于规则模板的意图识别算法和通过 TextRank 优化分类器训练流程而改进的朴素贝叶斯意图识别算法。通过实验分析算法性能，结果表明改进后的朴素贝叶斯算法确实得到了一定程度的性能优化；基于规则模板的算法在不需要样本训练的前提下，搭建灵活性更强，代码维护性和复用性更高，更适合中小企业应用。两种算法都可以为中小企业智能问答系统的研发提供一种实现途径，有一定的实践和参考价值。

参考文献：

[1] AHO A V, CORASICK M J. Efficient string matching: an aid to bibliographic searach [J]. Communications of the ACM, 1975, 18(6): 333-340.
[2] 何晗 . 自然语言处理入门 [M]. 北京 : 人民邮电出版社 ,2019.
[3] 冯建周 . 自然语言处理 [M]. 北京 : 中国水利水电出版社 , 2022.
[4] 刘挺 , 秦兵 , 赵军 , 等 . 自然语言处理 [M]. 北京 : 高等教育出版社 , 2022.
[5] 周志华 . 机器学习 [M]. 北京 : 清华大学出版社 , 2016.
[6] JAMES H M, DANIEL J. 语音与语言处理自然语言处理、计算语言学和语音识别导论 [M]. 北京 : 人民邮电出版社 , 2010.

【作者简介】

王炳翔（1988—），男，陕西西安人，硕士研究生，高级工程师，研究方向：系统规划与管理、知识图谱、自然语言处理等。

（收稿日期：2024-11-05）