情报学报 ISSN 1000-0135
第35卷第3期293-300，2016年3月
JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 35 No. 3, 293-300 March 2016

doi:10.3772/j.issn.1000-0135.2016.003.008

学术文本的结构功能识别
——基于章节内容的识别[1)]

黄 永 陆 伟 程齐凯

（武汉大学信息管理学院，信息检索与知识挖掘研究所，武汉 430072）

摘要 学术文本的结构功能是对学术文本结构及章节功能的阐述和概括，主要分为引言、相关研究、方法、实验、结论五种。根据研究对象的不同，结构功能识别的方法可分为基于章节标题的识别、基于章节内容的识别和基于段落的识别三个层次。然而，基于章节标题的结构功能识别法存在较多的局限性，如数据集构建困难、含未登录词的标题的识别率较低等。因此本文以章节内容为研究对象，探讨学术文本结构功能识别的第二个层次，并将基于章节内容的结构功能识别问题转化为文本分类问题，在特征选择上，除了传统的词汇特征，还引入词汇的聚类特征，并使用支持向量机作为分类器在基于自然标注的实验数据集上进行了实证研究。实验结果表明相比较于只使用词汇特征，本文所提方法的识别效果有明显提升。

关键词 结构功能 文本分类 词汇特征

The Structure Function Recognition of Academic Text
——Chapter Content Based Recognition

Huang Yong, Lu Wei and Cheng Qikai

(*School of Information Management, Wuhan University, Wuhan* 430072)

Abstract The structure function of the academic text refers to the summarization of academic text structure and section function, mainly dividing into five parts, introduction and related research, method, experiment, and conclusion. Depending on the research object, three different analytical levels can be applied to recognize the structure function, namely title-based, chapter-based and paragraph-based. However, there are many limitations of the title-based method, such as unknown words problem, dataset construction difficultly and so on. This paper studies the chapter content, recognizes the structure function of academic text at the chapter-based level and regards it as a text classification problem. This paper applies the bag-of-word feature and clustering features into support vector machine (SVM), the result is improved significantly.

Keywords structure function, text classification, lexical feature

1 引 言

学术文本的结构功能，是指使用“引言”、“相关研究”、“方法”、“实验”、“结论”这五类标签对学术文本的结构及章节功能进行阐述和概括[1]，它们是对文章思想不同方面的描述，如阐述研究依据、提出研究思想、得出研究结论等。在之前的研究中[1]，

收稿日期：2015年5月5日

作者简介：**黄永**，男，1991年生，博士研究生，主要研究方向：信息检索、数据挖掘。**陆伟**，男，1974年生，博士，教授，主要研究方向：信息检索、知识管理、数据挖掘等，E-mail:weilu@whu.edu.cn。**程齐凯**，男，1989年生，博士研究生，主要研究方向：信息检索、数据挖掘。

1） 本文系国家自然科学基金面上项目“面向词汇功能的学术文本语义识别与知识图谱构建”（项目编号：71473183）；教育部人文社会科学基地重大项目“面向细粒度的网络信息检索模型及框架构建研究”（项目编号：10JJD630014）的研究成果之一。

笔者基于章节标题识别章节结构功能，将基于章节标题的结构功能识别问题转化为序列标注问题，并在人工标注的数据集上取得了较高的准确率。但是这种方法具有很大的局限性，如①训练集构建难，易出现过拟合，可扩展性较差；②训练集中含未登录词的章节标题识别准确率较低；③部分学术论文的内部章节不具顺序性等。为了解决上述问题，本文从章节内容出发，将结构功能识别问题转化为文本分类问题。

词汇特征是文本分类问题中最常用的特征，词汇的出现、共现、共缺等是分类的主要依据[2]。词汇特征也是文本分类问题中最直接明显的特征，它通过一些潜在因素作用于文本分类，例如主题分布、词汇的类型分布等。在词汇特征的基础上，本文首先使用深度学习方法在无标记的学术文本上进行无监督学习得到词的词向量(Word Embedding)[3]，然后利用词向量对词汇进行聚类，最后使用章节内容中聚类类别的比例作为词汇特征的辅助特征来解决基于章节内容的学术文本结构识别问题。在基于自然标注数据集上的实验表明，使用两种混合特征能取得令人满意学术文本结构功能识别效果。

本文的主要贡献有以下几点：

(1)考虑到学术文献章节标题的随意性以及基于章节标题识别方法的局限性，本文提出了从章节内容角度解决结构功能识别这一全新视角；

(2)提出了一个基于文本分类的解决方法，在不考虑章节标题的情况下，仍然能够取得较好的识别效果；

(3)在词汇特征的基础上，本文引入深度学习方法，取得具有统计显著性的效果提升。

文章后续结构如下：第二部分对相关研究进行了调研，第三部分对所提出的方法进行阐述，第四部分对数据集的构建、实验的设计以及最终的实验结果进行了论述，最后总结工作，并对下一步的研究工作做出展望。

2 相关研究

结构功能是在章节层次上对于文章结构和章节功能的一种描述，基于章节内容的结构功能识别是一种文本分类问题。传统的文本分类研究大都使用词汇作为分类特征，研究的核心在于不同的特征选择方法[4]、不同模型[5]的效果比较。除此之外，也有将词汇的潜在主题作为特征用于面向主题分类任务中的相关研究[1]。

深度学习能够在浅层特征中根据不同的任务学习得到深层次的潜在影响因素，因此在各个领域取得了不错的效果[6]。例如在自然语言处理领域中，通过神经语言模型[7]训练得到的使用多维浮点数来表示词汇的各个方面的语义特性的词向量，不断地被用于各种自然语言处理任务[8]，如分词[9]、词性标注[10]、命名实体识别[8]、语义角色标注[8]等。也有其他深度学习方法在文本分类中的应用，如文献[11]使用约束波兹曼机(RBM)模型堆叠构成深度致信网络(DBN)对词汇特征进行深层次的特征提取，并将深层次特征传入支持向量机(SVM)中进行分类；或是使用RBM构建新的文档主题模型[12]，用来改善文本的分类效果。这些文本分类研究使用新的算法解决文本分类问题，使用深度学习方法完成了对于浅层词汇特征之下的深层次特征的抽取，可见对于文本分类的潜在影响因素的有效分析将有助于提升分类效果。

本文在词汇特征的基础上，加入基于word2vec词向量的聚类特征，用于基于章节内容的结构功能识别，在实验中取得了令人满意的效果。

3 方法描述

3.1 结构功能识别框架

在学术文本中，句子是有意义的最小结构单元[2]。句子围绕文章的主题，构成章节、段落，对文章的方法、实验、结论进行描述；章节构成文章，传递作者的观点、思想、知识；文章累积而成期刊、领域，反映整个领域的动态和热点。在这样的嵌套层次结构中，高层次结构单元对低层次结构单元施加约束，但都可以通过观察词汇分布在各层次上进行分析和研究。结构功能识别是一种章节层次的面向结构的文本分类，词汇特征是重要的分类依据。基于词汇特征的文本分类框架，如图1所示左图所示。假设数据集中词汇表中有V个词汇，则每一个段落可以表示成为词汇的词频向量，也是图中的c_{ij}，其中$i \in \{1,2,\cdots,m\}$，$j \in \{1,2,\cdots,V\}$，章节也是其包含段落的累积，使用特征选择方法得到最有用的词汇特征$(F_1,F_2,\cdots,F_n)$，最终使用文本分类算法，学习特征与结构功能类别的映射。

不同的结构功能中，其包含的不同类型词汇的比例是不同的，因此本文希望通过聚类方法对相同

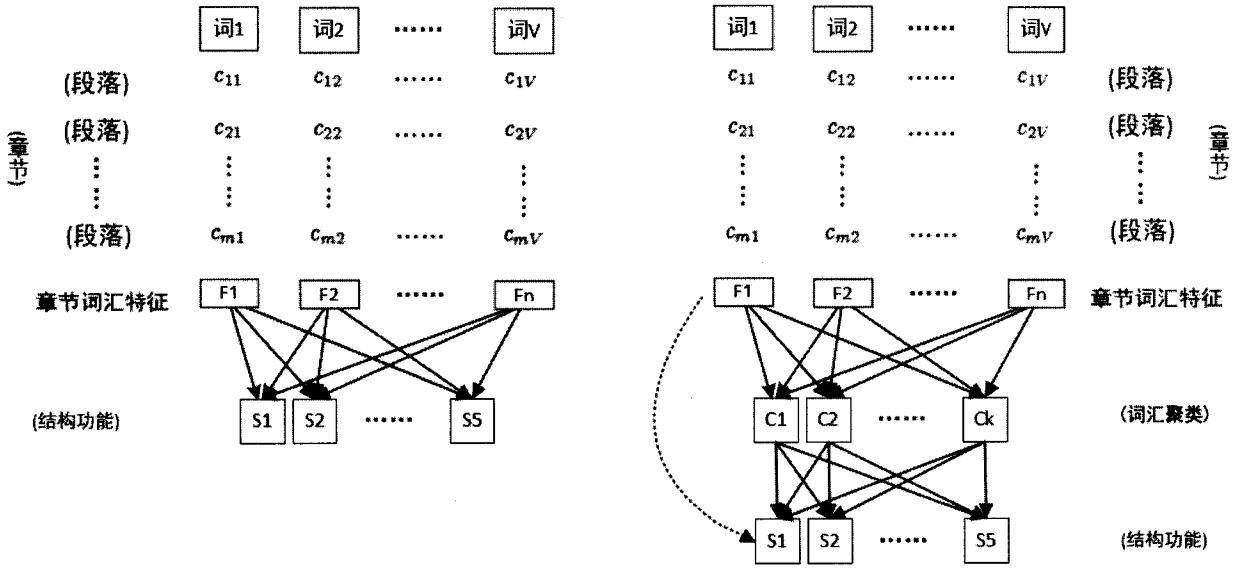
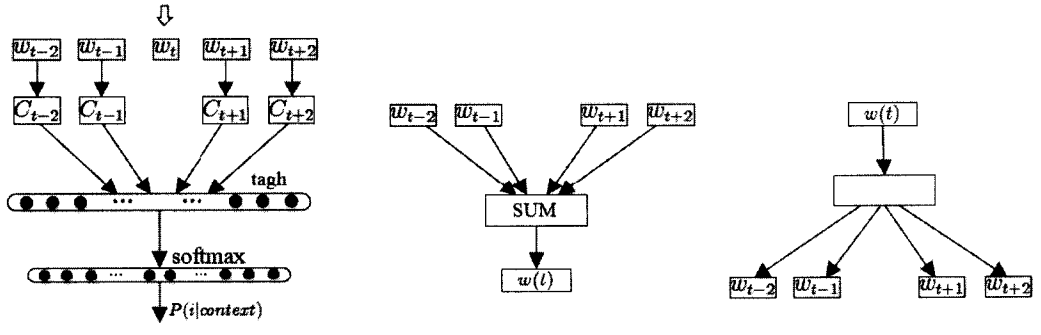


图 1　结构功能识别框架



(a) NNLM　　(b) CBOW　　(c) Skip-gram

图 2　语言模型训练框架

类型的词汇进行聚类，将聚类的类别在各章节中的分布作为深层次的特征。相比较于基于词汇特征的文本分类流程，本文提出分类框架如图 1 的右图所示，在词汇特征的基础上使用聚类方法对词汇类型进行聚类，使用聚类特征和词汇特征混合进行最终的文本分类。右边框架图中虚线箭头表示词汇特征与聚类特征同时作用于最终的分类任务。

3.2　词汇聚类

不同结构功能的章节包含的词汇类型比例是不同的。词汇的类型一般是由词汇所在上下文决定的，词汇类型的划分可认为是具有相同上下文的词汇聚类，最基本的问题是词汇的表示。共现词汇的向量空间模型是最常用的词汇表示方法，但是其前提假设是词汇之间相互独立，且与顺序无关，而词汇在句子中的词汇类型与其在句子中的位置、角色紧密相关，显然使用空间向量模型表示词汇进行聚类是不可行的。词向量是一种根据词汇上下文训练出的分布式表示方式，一般是通过神经语言模型进行训练，模型的框架结构一般具有以下几种，如图 2 所示。图(a)是最常用的神经语言模型的框架[7]，图(b)是连续词袋模型[13]，去掉了最耗时的非线性隐藏层并且所有词共享隐藏层。图(c)是 Skip - gram 模型[14]，与连续词袋模型相反，通过对邻近词的预测来进行模型训练。本文使用图(c)的框架对词向量进行训练，不仅缩减了传统模型中非线性的神经元运算消耗的时间，而且具有很高的准确率。该训练过程独立于整个识别框架，可以预先完成。

神经语言模型通过使用词汇的上下文信息，将词汇映射到多维空间中去。具有类似的上下文的词汇在空间中的距离越近，也就是相同类型的词汇在空间中的距离越近。本文使用 K-means 聚类方法对

相同类型的词汇聚类,同时计算不同类型词汇在不同章节中所占的比例,完成词汇聚类。

3.3 使用SVM进行分类

支持向量机(SVM)在文本分类任务中有非常好的分类效果。由于文本分类的特征维度高,训练数据大,线性SVM可以克服该缺点,且分类效果与其他核函数的SVM相差不大,所以本文使用LIBLINEAR作为分类器。LIBLINEAR是从由Lin等[15]开发的LIBSVM独立出来的用于文本分类的线性SVM工具。

完成章节的词汇聚类之后,可得到不同类型词汇在章节中所占的比例,随后以词汇频次和不同类型词汇比例为特征进行结构功能分类,其中词汇频次是指词汇在章节中出现的次数。

4 实验及结果

4.1 数据集构建

结构功能的训练数据集可以根据文章的章节标题自动构建。章节标题可以认为是作者对于章节结构功能的标注,使用以下标题(表1)与学术文本中的章节标题进行完全匹配,并将匹配得到的章节标注为对应的结构功能。

表1 使用章节标题筛选结构功能

结构功能	对应的章节标题
引言	introduction
相关研究	related work, literature review, background
方法	method, methodology, model
实验	experiment, result, data
结论	conclusion, conclusion and discussion, discussion

本文对ScienceDirect中2000~2013年计算机领域128本期刊26万篇论文全文进行抽取,最终得到约27万条样本,其中引言有13万条,相关研究和方法各2万多条,实验4万多条,结论约6万条,各个结构功能类别的样本分布十分不平衡。因此在抽取得到的样本中对每种结构功能随机抽取了5000条训练样本,构建出规模为25000的平衡数据集。

4.2 特征选择

特征选择过程主要包括词汇的特征选择和词汇聚类的类别数量的确定两个过程。

对于词汇特征,首先将所有词汇作为特征进行特征抽取,然后使用信息增益进行特征选择。具体步骤如下:

(1)预处理:将数据集中所有词进行转为小写,去除所有标点符号、数字,使用PorterStemmer进行词干提取,去除停用词。

(2)统计词频:统计每个词在数据集中出现的次数,将词频大于10的词汇作为词汇特征,得到一个大小为23 122的词汇表。

(3)特征抽取:根据得到的词汇表对章节内容进行特征抽取,考虑到因章节长度带来的影响,以每一个词汇在章节中出现的频次与章节中所有词出现的频次的比值作为词汇特征值。

(4)特征选择:本文使用的是信息增益,选择信息增益最大的K个词作为词汇特征。以下为不同K值对分类实验结果的影响。

图3显示了不同的K值对于分类实验的效果的影响,纵坐标为五折交叉检验所得准确率。由图中可以明显看出,在K值取4000时,准确率达到第一个高峰,随着K值的增大,准确率没有明显上升。所以本文选择信息增益最大的4000个词汇作为词汇特征。

对于词汇的聚类特征,处理步骤如下:

(1)预处理:对于计算机领域的128本期刊的正文进行分句、词干提取、小写转换、去除标点符号处理。

(2)词向量训练:使用word2vec工具[16],在计算机领域的128本期刊的正文上进行词向量训练。

(3)词汇聚类与统计:基于训练得到的词向量,使用k-means将相同类型的词汇进行聚类,得到N个类别,并计算出章节中不同类型词汇所占比例。以下为类别数N对于分类效果的影响:

如图4中所示,纵坐标为不同类别数N生成的聚类特征的五折交叉检验的准确率。可以明显看出,在$N<750$时准确率逐步上升,在$N=750$时,准确率达到最高,在之后准确率上下波动。在本文之后的试验中,聚类类别数量设置为750。

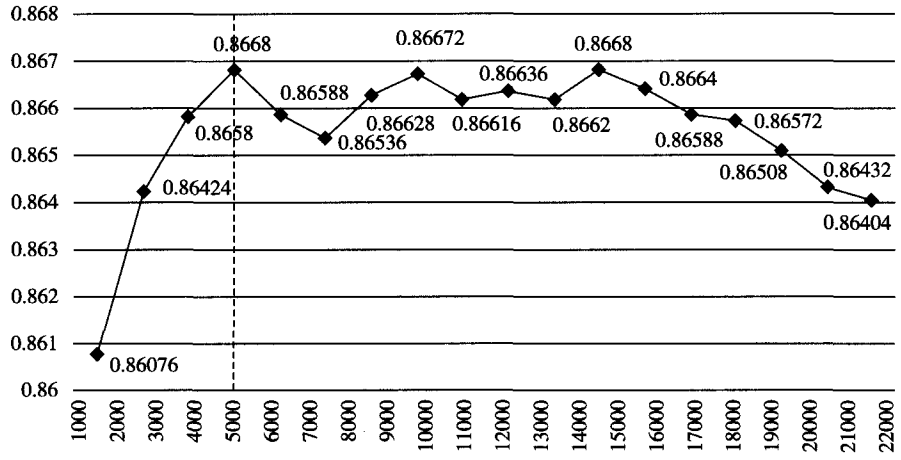


图 3　不同的 K 值对分类准确率的影响

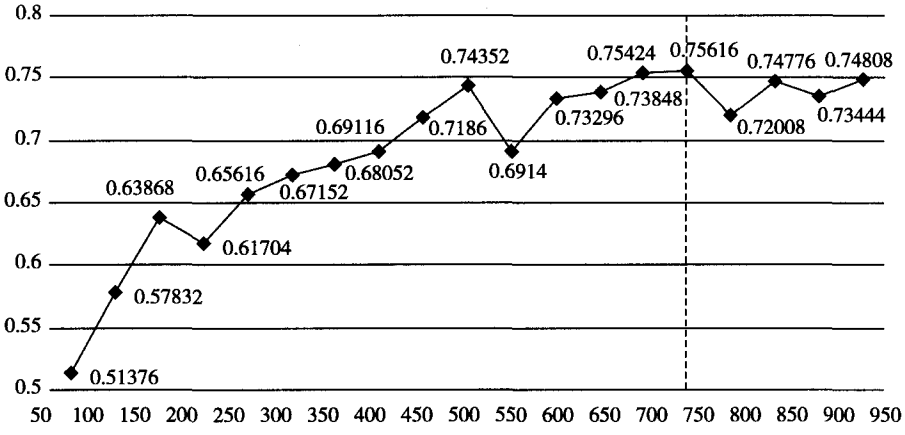


图 4　类别数 N 对分类效果的影响

4.3　基于词汇特征的分类

词汇特征作为文本分类中最常用的特征，本文首先探讨词汇特征对于实验分类的影响。使用上一节中信息增益最大的 4000 词作为词汇特征，并探讨归一化对于实验结果的影响。设置一下两组对比试验：

（一）4000Words_TF + LIBLINEAR

（二）4000Words_TF_normlized + LIBLINEAR

上述两组实验中，（一）直接使用 4000 个词在各个章节中出现的频次作为特征值，（二）则是在特征词的基础上使用章节长度对词频进行归一化处理。将两组特征传入 LIBLINEAR 中在数据上进行五折交叉检验，得到以下结果：

通过表 2 可以看出，实验一基于词汇特征的分类可以取得不错的分类效果，在每一个结构功能的准确率都能在 80% 以上，其中“实验”类别的准确率达到了 90%。在经过归一化之后只有个别指标上有微弱提升，在多数指标上都有下降。整体上长度归一化降低了词汇特征的分类效果。

表 2　基于词汇特征的分类实验结果

	实验一			实验二		
	P	R	F	P	R	F
引言	0.8100	0.8314	0.8206	0.8277	0.8078	0.8176
相关研究	0.8598	0.8634	0.8616	0.8458	0.8708	0.8581
方法	0.8596	0.8976	0.8782	0.8505	0.8964	0.8728
实验	0.9102	0.9102	0.9102	0.9039	0.9010	0.9024
结论	0.8986	0.8314	0.8637	0.8934	0.8430	0.8675
总计	0.86764	0.8668	0.86686	0.86426	0.8638	0.86368

4.4　基于词汇聚类特征的分类

基于章节内容的结构功能识别是一种面向结构的文本分类，本文认为识别章节中不同类型词汇所占比例是有意义的，因此使用 word2vec 计算每一个词汇的词向量，然后使用 *K*-means 根据词向量进行

聚类,使用各类别词在章节中的分布代表不同词汇类型所占比例。因此本部分实验引入 word2vec 之后,设置如下两组实验:

(三)750Clusters_CF + LIBLINEAR

(四)750Clusters_CF_normlized + LIBLINEAR

其中实验(三)使用 750 个类簇的在章节中的频次作为特征值,在实验(四)种同样使用章节长度对每一个类别的特征值进行归一化,两组特征进行五折交叉检验得到以下结果:

表 3 词汇的聚类特征的分类结果

	实验三			实验四		
	P	*R*	*F*	*P*	*R*	*F*
引言	0.6637	0.7452	0.7021	0.7177	0.6426	0.6781
相关研究	0.7563	0.7884	0.772	0.7428	0.8096	0.7747
方法	0.7259	0.7742	0.7492	0.7413	0.7610	0.7510
实验	0.8088	0.8502	0.8290	0.8314	0.8434	0.8374
结论	0.8686	0.6228	0.7255	0.7644	0.7442	0.7542
总计	0.76466	0.75616	0.75556	0.75952	0.76016	0.75908

从表 3 中可以看出,词汇聚类特征在各种结构功能的准确率都能高于 70%,其中"实验"类的准确率达到 83%。从 *F* 值来看,词汇聚类特征经过归一化之后分类效果有明显提升,这与基于词汇特征的分类的实验结果相反。整体上词汇聚类特征的分类准确率低于基于词汇特征的分类准确率,但是仍能看出其能够为结构功能分类带来一定的信息量。

4.5 混合特征

由上述两节的实验结果可以看出,词汇聚类特征的分类效果不如使用词汇特征的分类效果,并且长度归一化对于两种特征的效果也不同,因此本节主要考虑将两种特征混合进行结构功能试验,并设置了以下四组对比试验:

(五)4000 Words _TF + 750 Clusters _CF + LIBLINEAR

(六)4000 Words _TF + 750 Clusters _CF _normlized + LIBLINEAR

(七)4000 Words _TF_normlized + 750 Clusters_CF + LIBLINEAR

(八)4000 Words _TF_ normlized +750 Clusters_CF_ normlized + LIBLINEAR

其中(五)将两种特征未经过长度归一化的特征串联,(六)将(五)中的词汇聚类特征进行长度归一化,(七)将(五)中的词汇特征值进行章节长度归一化,(八)将两种特征都进行归一化处理。使用上述四组混合特征在数据集上进行五折交叉检验,得到结果如下:

表 4 中可以看出,上述四组试验中,使用非归一化词汇特征 + 归一化词汇聚类特征(即实验六)在实验效果上相较于其他特征组合有全面提升。其各种结构功能的识别准确率都在 86% 之上,"实验"类的准确率达到了 92%。从召回率来看,"实

表 4 混合特征的分类结果

		引言	相关研究	方法	实验	结论	总计
实验五	*P*	0.8133	0.8616	0.8610	0.9095	0.9008	0.86924
	R	0.8318	0.8640	0.8970	0.9124	0.8372	0.86848
	*F*1	0.8224	0.8628	0.8786	0.9109	0.8678	0.8685
实验六	*P*	0.8788	0.8761	0.8821	0.9288	0.8695	0.88706
	R	0.8196	0.8772	0.8780	0.9058	0.9526	0.88664
	*F*1	0.8482	0.8767	0.8800	0.9172	0.9091	0.88624
实验七	*P*	0.8535	0.8588	0.8669	0.9216	0.9026	0.88068
	R	0.8248	0.8760	0.8908	0.9024	0.9124	0.88128
	*F*1	0.8389	0.8673	0.8787	0.9119	0.9093	0.88122
实验八	*P*	0.8339	0.8487	0.8479	0.9097	0.8985	0.86774
	R	0.8082	0.8720	0.8998	0.9026	0.8518	0.86688
	*F*1	0.8208	0.8602	0.8731	0.9052	0.8745	0.86676

验”、“结论”的召回率都在90%之上，其中“结论”的召回率达到了95%。从F值来看，实验六结果中每种结构功能的F值相较于其他组合有明显提升。

4.6 错误分析

本节以上述八组试验中效果最好的实验六作为分析对象，分析错误的主要原因。对实验六的五折交叉检验结果进行分析，得到其具体的分类矩阵。如表5所示，其中每一行表示每一种结构功能分别被识别为的类别的个数，每一种类别的总数为5000。每一列，表示从各个类别中分为该类别的数量。

因此从行来看，“引言”更多的被错分为“相关研究”和“结论”两种结构功能，而“相关研究”和“结论”也同样更多的被错分为了“引言”。“方法”的错分样本中，错分为“实验”的样本比例最大。同样“实验”的错分样本中，错分为“方法”的样本比例最大。

表5 分类矩阵表

类别	引言	相关研究	方法	实验	结论
引言	4098	424	133	27	318
相关研究	294	4386	150	36	134
方法	126	108	4390	227	149
实验	42	33	282	4529	114
结论	103	55	22	57	4763

同样可以从侧面看出，从内容与结构角度来看引言、相关研究和结论三种结构功能更相似，方法和实验两种结构功能的更相似。

4.7 讨 论

词汇特征在结构功能识别中取得了不错的效果，词汇聚类特征虽然同样有效，但是整体效果不如词汇特征。这样的结果与笔者调研的多篇文献中的结论相反[17,18]，这些文献使用word2vec的聚类结果作为特征在文本分类实验中取得了比词汇特征更好的效果，这样的结果给笔者造成极大的困扰，最终本文的实验结果证明使用word2vec的聚类结果作为特征进行文本分类效果不如词汇特征，本文结果与word2vec的作者mikolov最新实验[19]、LeCun的实验[20]的结论相同，LuCun认为出现这种情况的原因是词汇特征是随不同任务而改变的，但是词向量总体上来说是恒定的[17]。

在本文中集中探讨了词汇特征和词汇的聚类特征两种特征对于结构功能识别的影响，并探讨了章节归一化的作用。除词汇特征及聚类特征外，章节内容中仍然存在其他特征如章节的位置信息，章节中引文信息等都可以用来辅助结构功能识别，同样其他的特征值的归一化方法也值得探讨。

5 总结与展望

本文为解决基于标题的结构功能识别的局限性，以章节内容为研究对象，将基于章节内容的结构功能识别问题其转化为文本分类问题，只是用词汇特征本文$F1$值就达到了86%，词汇特征在本类问题的表现十分的好，同时以词汇与归一化词汇聚类特征达到了88.7%的$F1$值。同样本文仍然存在一定的不足，本文提出了两类特征包括词汇特征以及基于词向量的聚类特征，仍然存在其他类型特征如上下文特征如位置特征等对于分类有所帮助；在使用词汇特征以及聚类特征中，本文使用TF以及CF作为特征值，不同的特征值取值方法如TFIDF等以及归一化方法同样会给实验结果带来差异，这些并没有全面涉及，这些不同的因素的影响仍然会使下一步探索的关键内容。

未来的研究将在两个方面展开：首先，在实验中笔者发现基于章节内容的结构功能分类所用到的词汇及分类函数与传统的面向主题的文本分类存在较大差异，具体存在哪些差异以及原因将是未来研究的重点；其次，本文解决了结构功能识别问题三个层次[1]中的第二个层次——基于章节内容的识别，而第三个层次——基于段落的识别，将在下一步工作中进行探索。

参 考 文 献

[1] 陆伟，黄永，程齐凯，等. 学术文本的结构功能识别——功能框架及基于章节标题的识别[J]. 情报学报，2014(9)：979-985.

[2] Leydesdorff L. The Challenge of Scientometrics: The Development, Measurement, and Self-organization of Scientific Communications [M]. Boca Raton Universal-Publishers, 2001.

[3] Hinton G E. Learning distributed representations of concepts [C]//Proceedings of the eighth annual conference of the cognitive science society. 1986, 1: 12.

[4] Yang Y, Pedersen J O. A comparative study on feature

selection in text categorization[C]//ICML. 1997, 97: 412-420.

[5] Forman G. An extensive empirical study of feature selection metrics for text classification[J]. The Journal of Machine Learning Research, 2003, 3: 1289-1305.

[6] Bengio Y. Learning deep architectures for AI[J]. Foundations and trends © in Machine Learning, 2009, 2(1): 1-127.

[7] Bengio Y, Schwenk H, Senécal J S, et al. Neural probabilistic language models[M]//Innovations in Machine Learning. Springer Berlin Heidelberg, 2006: 137-186.

[8] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011, 12: 2493-2537.

[9] Wu K, Gao Z, Peng C, et al. Text Window Denoising Autoencoder: Building Deep Architecture for Chinese Word Segmentation[M]//Natural Language Processing and Chinese Computing. Springer Berlin Heidelberg, 2013: 1-12.

[10] Zheng X, Chen H, Xu T. Deep Learning for Chinese Word Segmentation and POS Tagging[C]//EMNLP. 2013: 647-657.

[11] Liu T. A Novel Text Classification Approach Based on Deep Belief Network[M]//Neural Information Processing. Theory and Algorithms. Springer Berlin Heidelberg, 2010: 314-321.

[12] Larochelle H, Bengio Y. Classification using discriminative restricted Boltzmann machines[C]//Proceedings of the 25th international conference on Machine learning. ACM, 2008: 536-543.

[13] Schwenk H. Continuous space language models[J]. Computer Speech & Language, 2007, 21(3): 492-518.

[14] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv Preprint arXiv: 1301.3781, 2013.

[15] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification[J]. The Journal of Machine Learning Research, 2008, 9: 1871-1874.

[16] Poultney C, Chopra S, Cun Y L. Efficient learning of sparse representations with an energy-based model[C]// Advances in neural information processing systems. 2006: 1137-1144.

[17] Zhang D, Xu H, Su Z, et al. Chinese comments sentiment classification based on word2vec and SVM perf[J]. Expert Systems with Applications, 2015, 42(4): 1857-1863.

[18] Su Z, Xu H, Zhang D, et al. Chinese sentiment classification using a neural network tool—Word2vec [C]//Multisensor Fusion and Information Integration for Intelligent Systems (MFI), 2014 International Conference on. IEEE, 2014: 1-6.

[19] Zhang X, LeCun Y. Text Understanding from Scratch [EB/OL]. http://arxiv.org/abs/1502.01710v4. 2015-09-08.

[20] Le Q V, Mikolov T. Distributed representations of sentences and documents[J]. arXiv Preprint arXiv: 1405.4053, 2014.

(责任编辑 赵 康)