面向传媒语言语料库的关键词自动抽取研究*

吴继媛 孙淳 侯敏

中国传媒大学 北京 100024

E-mail: wjy__00@163.com, houminxx@263.net

摘 要：本文根据广播电视语言及其关键词的特点，提出通过建立过滤词典过滤垃圾串直接切分短语，并将切分好的短语序列作为关键词候选，进而对其进行各种权重因子的加权计算，最后根据统计结果按序抽取关键词的研究策略，构建了一个名为“传媒语言语料库关键词自动抽取”的软件系统。经过封闭和开放测试，该系统达到了实用性的要求。

关键词：自动抽取，关键词，传媒语言语料库，短语切分，过滤词典

Automatic Keyword Extraction Research for Media-Language Corpus

Wu Jiyuan, Sun Chun, Hou Min

Communication University of China, Beijing 100024

E-mail: wjy__00@163.com, houminxx@263.net

Abstract: According to the features of media language and its keywords, the paper proposed a new research strategy, with which a software system named "Automatic Keyword Extraction System for Media-Language Corpus " was established. The main steps of the strategy are: firstly, segment the words and phrases directly by compiling a dictionary as a filter of the unwanted words and phrases; secondly, select the segmented phrase array as the possible keywords, and then various factors of these keyword candidates are weighted; finally, extract the keywords in descending order based on the previous statistic data. As well, after the close and open tests, this system had reached its practical requirements.

Keywords: automatic extraction, keyword, Media-Language Corpus, phrase segmentation, filter dictionary

1 引言

关键词[1]是指从文章标题和正文中提取出来的能够反映原文主要内容的词。通常所说的关键词有相当一部分是短语，本文中提到的关键词除非特殊说明，都包括关键词和关键短语。关键词抽取是文本挖掘领域使用的术语，是指在非人工干预情况下，由计算机提取出关键词的技术。

目前汉语关键词抽取技术主要有以下几种：基于主题词表方法[3][4][5]、基于词义的方法[6]、基于统计的方法[9][8][10]、基于结构的方法[11]与基于 KeyGraph 算法或 KeyWorld 算法的方法[1][2]等。这些方法的相同点都是先分词，后进行关键词抽取，之后考虑各种特征对抽取出来的候选关键词进行部分合并，来弥补某些关键词或关键短语在分词阶段就被切开的缺陷。

本研究致力于提供一个能达到一定准确率、有实际应用价值的传媒语言语料库关键词自动

*本文得到国家语言资源监测与研究中心课题“[基于传媒语言语料库的关键词提取研究][YZYS08—05]”和国家 863 计划专项课题“面向汉语语音合成的言语语义计算模型研究”（项目编号：2007AA01Z198）之子课题“新闻播报言语数据库的分析与研究”（项目编号：IM0810）支持.

抽取系统。传媒语言语料库始建于2001年，以广播电视语料为主要内容。本文研究对象是传媒语言语料库中单一主题的广播电视节目转写文本。目前该语料库的关键词采取人工标注的方式，不仅耗费了大量人力物力，而且存在一致性难以统一等问题。另外，传媒语言语料库的语料有自己特殊的格式，语言上也有自身的特点，不适合直接利用现有系统来进行关键词自动抽取。

本文第2节通过分析2006年全部已由人工提出关键词的广播电视语料，归纳出广播电视语言的特点及其关键词的特点，进而根据这些特点，提出通过过滤垃圾字串直接切分短语，并将切分好的短语作为关键词候选串的研究策略；第3节对这一策略进行理论认证，构建三部过滤词典，通过程序实现短语切分，验证用垃圾字串过滤实现短语切分的可行性；第4节在短语切分的基础上，结合广播电视语料关键词的特点，运用加权方法，提出计算关键度的模型，并实现关键词的自动抽取；第5节对系统进行测试，开放测试在抽取与标准语料相同数量(各为5个)关键词时准确率和召回率分别为74%和72%，将抽取数量放宽到8个后召回率达到85%；最后对全文总结。

2 广播电视语料及其关键词特点分析

广播电视语言大多来自于书面又有别于书面语言，表现为口头语言又有别于原生态口语。大众传播和有声传播的需要，使得口语化成为其主要特点，体现为：句子通常较短，句内成分常有省略，并存在较多的虚词，以及代词、副词和相对时间词。广播电视语言，无论对话类节目、独白类节目，或者居于两者之间的综合类节目，都属于口语体文本，只不过口语化程度不同。

考察2006年全部已提出关键词的广播电视语料，包括180个栏目，共10599个文本，由人工提取出关键词共34093个，用北京大学开发、中国传媒大学进一步完善的分词标注系统BBIbst对其进行分词标注：(1) 人工提取出的34093个关键词中，未被切开的有23482条，占68.88%，可以认为是“词”；被切开的有10611条，占31.12%，可以认为是“短语”。可以近似地得出结论：广播电视语料文本的关键词是由2/3的词和1/3的短语所构成。(2) 34093个关键词中含有名词标记“n”(地名、组织名、人名等专名也归入其中) 和时间词标记“t”的有25662条，占75.27%。含有“n”标记的短语中，名词是该关键短语的中心词的占绝大多数；只有极少量的动宾短语，如“看/v 病/n、偷/v 税/n”。因此，名词短语（含单个名词）占关键词总数3/4以上。进一步调查名词性关键词的组成，专指性名词有6399条，占24.94%，近四分之一。另外，广播电视语言中也会出现大量的新词语，造成未登录词比较多。

由此可见，广播电视文本关键词短语形式多，未登录词多，不适合采用目前广泛流行的先分词、再组合的关键词提取策略。

3 短语切分系统的构建

3.1 基于短语的关键词自动抽取的理论分析

既然广播电视文本不适合采用先分词后组合的策略，那如何将字符串分割，获得适合关键词形式的候选呢？广播电视语言的口语化特点给了我们启示：既然它句子短，虚词多，那么是否可以利用那些在关键词中不会出现的垃圾字串作为分割符号对文本进行分割呢？我们随机选择了一段文本进行观察：

她叫陈青，今年54岁，是福建省老年医院的一名针灸技师。这是陈青的女儿君君，已经五岁多的君君今天特别高兴，因为从今天起，她就可以上幼儿园了。陈青不是君君的亲生母亲，4年前她才第一次见到君君。是什么原因让这对母女走到了一起，在她们背后又有着怎样一段故事呢。（中央电视台《家庭》栏目2006年1月11日）

对提取关键词来说，这个段落中有用的信息有“陈青 福建省老年医院 针灸技师 陈青 女儿 君君 君君 高兴 上幼儿园 亲生母亲 君君 原因 母女 背后 故事”。文本中有些词是不可能成为关键词的，如代词（她、她们、这）、数量词（一段、第一次、54岁、对）、助词（了、着、的）、语气词（呢）、相对时间词（指必须联系上下文才能知道具体时间的词语，如：4年前、今年、今天、同时）、副词（不、可以、特别、才、在）、疑问词（什么、怎样）、使令动词（让）、感官动词（见到）、连词（因为）。如果在文本中滤去这些不可能成为关键词的词语，就自然形成了一个包含关键词候选的短语序列。其中最关键的是要精心构建过滤词典。因为语言现象十分复杂，过滤词典既要保证能把文本切成适当的短语序列，又不能丢失有用的、可能成为关键词的词串。

3.2 短语切分系统的实现

3.2.1 三部过滤词典的建立

停用词典 stoplist：停用词指不能出现在关键词中的词语及符号。包括文章中经常出现的标点符号、虚词以及没有实际意义的名词、动词，如“提出、但愿”等等。目前停用词典包含停用词1083条，是《2006年广播电视语料15000高频词表》和BBIbst分词软件基本词表中的不可能作为关键词的词语的并集。

条件词典 leftlist：有条件过滤的单字集合，是对停用词典的补充。一些放入停用词典的单字词，还可能与其他字组合成词，这些词语有可能成为关键词。如，将规则但！=|但丁|放在条件词典中，意为当“但”后面是“但丁”时，不作为过滤符号。目前条件词典包含这样的单字93个。

量词词典 quanlist：量词单独处理而不放入上两个词典，是因为量词之前一般会有数词、数字或指示代词，这些词一般也不能成为关键词，此外，在量词前面的数词或数字情况比较复杂，需要单独处理。量词词典中共有量词159个，取商务印书馆《简明汉日词典》附录中的《常用量词表》和《2006年广播电视语料词表》中的量词的并集。

3.2.2 短语切分系统的流程

短语切分是通过对停用词等的过滤达到将正文切分成短语的目的(题目中的短语提取不采用这种方式)。短语切分系统主要包含顺序操作的三个部分：数量短语过滤、停用词过滤、条件词过滤。之所以采用这个顺序，是多次调试，经验获得的结果。

4 基于短语的关键词自动抽取系统的实现

4.1 系统的框架

本系统分为四个模块：预处理模块、切分过滤模块、加权计算模块、后期处理模块。在前三个模块处理后，生成预处理文本PreTxt，去头信息处理后，生成净文本CleanTxt。（见图4-1）

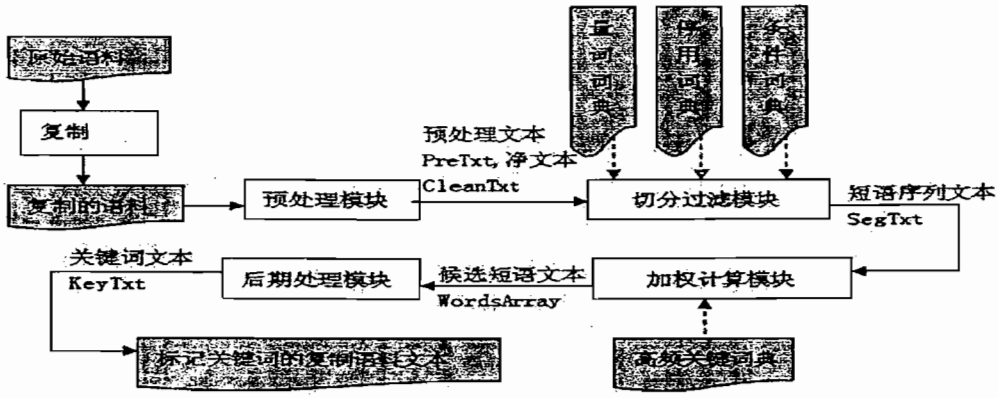

图 4-1 基于短语的关键词自动抽取系统框架图

4.2 系统的模块

4.2.1 预处理模块

预处理模块分为错误处理、半角转全角、去空白行、去头信息四个子模块。

4.2.2 过滤切分模块

过滤切分模块分为题目切分和正文过滤两个子模块，前者采用高频字串提取候选词语的方式，后者采用短语切分模块处理。二者采用不同的切分方式，是因为题目比较短，如果也用短语切分模块处理，会影响系统的速度；此外，短语切分模块采用规则过滤的方式，有错切分的情况，题目的高频字串提取候选词语也可以作为一个有益的补充。

4.2.3 计算加权模块

目前对于关键词权重的计算还没有一个公认的公式，基本上都采用经验公式的方法。我们对下述几类加权因子分别进行考虑，最后综合，再根据实验结果对计算参数进行适当调节。

a. 覆盖因子 Coverage(w)：主要考查的是词频信息。在同一篇文献中出现次数较多的词较为重要。但由于文本长度与出现的次数有着明显的对应关系，且广播电视语料文本长度差异很大，所以将频率公式定义为：

$$Coverage\ (w)=\frac{\text{词语出现次数}\times\text{词语长度}}{\text{正文长度}}$$，其中词语长度也是变量之一。

b. 位置因子 Position(w)：据表 4-1，对话类节目文本比独白类和综合类节目文本的关键词更多地依赖位置信息（如题目、首段、末段等），更少地依赖在正文中的出现频率。因此，对不同栏目形式，只采用一种权重计算公式是不合理的。本系统对独白和综合形式的文本标题、首段、末段分别设定的权值是：6、3、2，对对话形式的文本相应位置分别设定权值：7、4、3。

表 4-1 关键词按栏目形式统计结果

栏目形式		关键词长	题目出现频率	首段出现次数	末段出现次数	首次出现位置	末次出现位置	是否高频关键词	正文出现频率
独白类	统计个数	12335	12335	12335	12335	12335	12335	12335	12335
	平均值	3.346	.26	.858	.67	.2120	.5144	.30	.0079
	标准差	2.6689	.444	2.2998	2.612	.29865	.41078	.460	.01035
	最小值	1.0	0	.0	0	.00	.00	0	.00

	最大值	27.0	2	155.0	155	1.00	1.00	1	.11
对话类	统计个数	6283	6283	6283	6283	6283	6283	6283	6283
	平均值	3.082	0.39	1.002	0.58	0.1552	0.6065	0.34	0.0075
	标准差	1.6683	0.493	1.3407	0.948	0.27149	0.40853	0.473	0.00973
	最小值	1.0	0	.0	0	.00	.00	0	.00
	最大值	29.0	2	19.0	18	1.00	1.00	1	0.08
综合类	统计个数	15475	15475	15475	15475	15475	15475	15475	15475
	平均值	2.844	0.28	0.746	0.61	0.1903	0.5739	0.35	0.0078
	标准差	1.4702	0.457	0.9828	1.026	0.29422	0.41774	0.476	0.01029
	最小值	1.0	0	.0	0	.00	.00	0	.00
	最大值	26.0	2	16.0	29	1.00	1.00	1	0.09

c. 标记因子 InPunc(*w*)：此处标记指书名号和引号。书名号和引号内的词语，比普通词语更可能成为关键词，可以认为是出现在特殊标记里。出现在标记里的词语，该项因子取权值为 1。

d. 词长因子 LenWeight(*w*)[7]：前人研究表明，关键词长一般在 2-8 字，因汉语中双音节词占优势，关键词更多集中在 2 字和 4 字词语。据统计结果，独白类 1-8 字词成为关键词的可能性的比值分别是：1.1：56.6：17.4：17.0：3.3：2.3：0.9：0.5。最初，关键词词长为 2 时的权重设为 3，这引起了三字四字词和其二字子串同时出现时，大量取二字子串的问题，如“孟广斌”和“广斌”、“三国演义”和“三国”。反复调试后调整为 1.5，很好地解决了这个问题。对话和综合类的关键词长分布表与独白类的大致相同，故采用同一公式：

$$\text{LenWeight}(w)=\begin{cases}0.6 & \text{当}w\text{字长为}1\\ 1.5 & \text{当}w\text{字长为}2\\ 1 & \text{当}w\text{字长为}3\\ 1 & \text{当}w\text{字长为}4\\ 0.25 & \text{当}w\text{字长为}5\\ 0.2 & \text{当}w\text{字长为}6\\ 0.1 & \text{当}w\text{字长为}7\\ 0.1 & \text{当}w\text{字长为}8\\ 0.05 & \text{其他情况}\end{cases}$$

e. 首位置因子 FirstLoc(*w*)：广播电视节目有时会有引导语，对话类节目在开场时会有嘉宾介绍等内容，所以，除了在位置因子中进行首末段及题目的中出现的参数设置，也需要考虑到候选词语首次出现的位置，作为关键词权重的一个参考。该因子取 0-1 之间的浮点数，是 1 与候选词语的首次出现的位置与文本长度的比值的差值。

f. 末位置因子 LastLoc(*w*)：候选词语末次出现位置与文本长度的比值。

g. 高频因子 InHKey (*w*)：尽管关键词是为了体现文章主题，即一篇文章与另一篇的差异性而出现的，但又不得不承认，有些词语比其他词语更可能成为关键词。如“爱心”、“安全”、“营救”这些词比其他词具有更强的概括性，更能体现文章主题。于是，我们将做过 5 次以上关键词的词语在经过人工选择（删除热点事件、热点人物、专题名、栏目名）后，编成高频关键词词典 freqkeylist，并对词语是否在该词典中出现设置一个权重，取 0 或 1。

权重的计算总公式：Weight(*w*)=LenWeight(*w*)×Coverage(*w*)×Score(*w*)，其中，Score(*w*) = Position(*w*)+ InHKey (*w*) + InPunc(*w*) + FirstLoc(*w*) +LastLoc(*w*)。

短语序列文本 SegTxt 中所有的词语加权计算后，将结果写进候选词语文本 WordArray 中。

4.2.4 后期处理模块

后期处理模块是取候选词语文本 WordArray 中排在前 20 位的候选词语，根据一定的规则，对它们之间互相包含的情况进行处理，如“赵小燕”和“小燕”在文本中同指一人，只留下“赵小燕”作为关键词，但要将“小燕”的权值加在“赵小燕”上。重新排序后，写入关键词文本 KeyTxt。

此模块中，我们引入了传媒语言语料库流行语提取中的“贡献词”的概念。贡献词是指本身不做关键词，而将它的权重或部分权重加到与之相关词语的权重上的词。被它贡献权重的词语称为“被贡献词”。

5 系统测试及分析

封闭测试：语料采用的是传媒语言语料库中 2006 年已标注了关键词的的 50 篇文本。语料内容涉及独白、对话、综合三个门类的 16 个栏目，每篇标注关键词 5 条。实验结果显示，如果同样抽取 5 条关键词，测试结果准确率和召回率分别是 74.80%和 69.60%。

分析关键词提取效果不好的文本，有如下几个原因：（1）对话段落多，且每段对话前都有说话人（主持人、嘉宾）的名字，所以这几个人名频次特别高。如果对对话文本中说话人做屏蔽处理，可解决该问题。（2）小标题的存在。关键词只在小标题中出现过一次或两次因频次低而未被提取出。（3）文本过短，题目中的关键词没有提出来。题目切分模块中，只提取在正文中至少出现三次的高频词串。如果对题目切分模块按照文件长度进行相应调整，可以解决这个问题。（4）过滤词典不够完善。（5）人名中的单字被过滤掉。为了弥补这一缺憾，题目中的短语切分可采用在正文中查找最长高频字符串的方法。此外，人名还存在小名、昵称的问题。

开放测试：选取 2009 年未被标注关键词的语料，随机抽取了来自 18 个栏目的 20 篇单一主题、有题目的文本，由人工每篇提取出 5 条关键词，又经多人反复核对，作为评测标准。然后用机器每篇提取 5 条关键词，将二者进行比对，计算准确率与召回率。又将提取结果放宽到 8 个，再计算召回率。由于篇幅限制，下面只列举了部分结果，见表 5-1：

表 5-1 关键词自动抽取开放测试结果（部分）

机器关键词（括号中为第 6 到 8 个关键词）	人工关键词	准确率	召回率	召回率（放宽到 8 个词）	来源
华连英；家庭暴力；丈夫；杜佰春；选择（母亲；悲剧；杀死）	华连英；丈夫；家庭暴力；法律；人身保护裁定	80%	60%	60%	今日说法 2009.3.25
方队；训练；阅兵；领队；女民兵（队员；卢宛峰；教官）	国庆阅兵；女兵方队；卢宛峰；训练；领队	80%	80%	100%	军事纪实 2009.10.8
怀孕；周老；医院；检查；湖南省（恶心呕吐 ；医生；媒体）	周老太；怀孕；媒体；检查；高血压	60%	60%	80%	解密 2009.6.25
房地产；中国经济；投资；价格；泡沫（担心；隐患；大的）	中国经济；隐患；房地产；泡沫；投资	80%	80%	100%	财经点对点 2009.10.25

故事；风集团；总经理；企业；汽车（摩根；投资；的刘成强）	故事；刘成强；时风集团；管理理念；投资	60%	60%	80%	风云鲁商 2009. 4. 26
网络；聊天；网友；男人；诈骗（生活；谭伊娜；现金）	互联网；激情聊天；巨额现金；网络诈骗；净化网络	60%	60%	80%	记者调查 2009. 4. 24
历史；石悦；明朝；明月；工作（大学；世界；喜欢）	石悦；明朝那些事儿；历史；明朝；当年明月	80%	80%	80%	面对面 2009. 4. 11
…	…	…	…	…	…
平均值	——	78%	69%	85%	——

测试结果显示，抽取 5 条关键词时，准确率和召回率分别为 78%和 69%，而将数量放宽到 8 个，召回率则达到 85%。开放测试结果比封闭测试结果要好，可能与封闭测试的语料中关键词提取只是语料采集人一人所作，不如开放测试语料是经多人核对，质量更高有关。这也许从另一个侧面说明关键词自动抽取比人工标注在某种程度上更客观，一致性更强。

上述封闭测试和开放测试的实验证明，用这种以垃圾串过滤来进行短语切分的方式对传媒语言语料库的关键词进行抽取，是可行的，具有较高的实用价值。

6 结论

本文基于传媒语言语料库中广播电视文本的特点，考察了基于短语切分的关键词自动抽取方式的可行性。综合考虑了广播电视文本中关键词的特点以及文本的结构，设计了关键词候选词语的权重模型。在此基础上，构建了一个面向传媒语料库的关键词自动抽取系统。经过测评实验证明，达到了较高的准确率和召回率，具有实用价值。

目前系统还有改进的空间：过滤词典需要进一步完善；如果系统能结合专有名词识别和同指词排查，应该能取得更好的效果。

参 考 文 献

[1] Ohsawa Y,Benson N E,Yachida M. KeyGraph:automatic indexing by co-occurrence graph based on building construction metaphor research and technology advances in digital libraries[C] IEEE International Forum on proceeding, 22-24 April 1998: 12-18.

[2] Y.Matsuo, Y.Ohsawa and M.Ishizuka. KeyWorld: Extracting Keywords from a Document as a Small World. Proc.Discovery Science(DS'2001), pp. 271-281. Washington D.C. USA, Nov. 2001.

[3] 陈琳贤. 医学主题词及文献主题标引[J]. 广西医学, 1989, (03).

[4] 韩客松, 王永成. 一种用于主题提取的非线性加权方法[J]情报学报, 2000, (06).

[5] 韩客松, 王永成. Web页面中文文本的主题自动提取研究情报学, 2001. 4第20卷2期page218-223.

[6] 李有梅. 基于词义的关键词抽取方法的研究情报理论与实践, 2000第23卷2期page81-83.

[7] 李素建, 王厚峰, 俞士汶, 辛乘胜. 关键词自动标引的最大熵模型应用研究[J]计算机学报, 2004, (09).

[8] 杨文峰, 李星. 基于PAT TREE统计语言模型与关键词自动提取[J]. 计算机工程与应用, 2001, (15).

[9] 张海燕, 陈治平, 童调生. 基于2-grams短语标引的关键词自动抽取, 绍兴文理学院学报, 2002, 22(3):53-54.

[10] 张清军, 朱才连. 基于统计的中文文本主题自动提取研究[J]四川大学学报(工程科学版), 2004, (03).

[11] 郑家恒, 卢娇丽. 关键词抽取方法的研究[J]. 计算机工程, 2005, 18(9):194-196.