计算机科学
COMPUTER SCIENCE

基于字频差算法与左切分词库构建的专利文献组件名称识别方法

孔嘉斌，吕剑文，刘江南，杜文轩

引用本文

孔嘉斌，吕剑文，刘江南，杜文轩.基于字频差算法与左切分词库构建的专利文献组件名称识别方法[J]．计算机科学, 2023, 50(7): 229-236.

KONG Jiabin, LYU Jianwen, LIU Jiangnan, DU Wenxuan. Recognition Method of Component Names in Patent Documents Based on the Algorithm of Word Frequency Difference and Library of Left-segmentation Words [J]. Computer Science, 2023, 50(7): 229-236.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220500068

基于字频差算法与左切分词库构建的专利文献组件名称识别方法

孔嘉斌　吕剑文　刘江南　杜文轩
湖南大学汽车车身先进设计制造国家重点实验室　长沙 410082
(jbkong@hnu.edu.cn)

摘　要　机械专利文献蕴含着海量以组件名称为信息单元的领域知识信息,组件名称用词灵活多变,具有独特、复杂和生僻等特点,难以被计算机准确识别,成为专利知识挖掘的一大阻碍。为了提出组件名称的高效识别方法,剖析并提炼专利文本语句中的组件名称构词特征;从组件名称相关的外部用词入手,通过标识附图标记,识别其左侧的名称字符,自动从文本中检索候选名称,并构建组件候选名称集合;提出了字频差算法,过滤候选名称集合的冗余字符;提出了动态构建左切分词库算法,进一步剔除未能被过滤的冗余字符;通过交叉实验测试和分析识别过程中字频差先验阈值、词频阈值和字频差阈值的选取对识别效果的影响,形成一种面向机械领域中文专利的组件名称识别三段式综合方法。最后通过对实验结果的对比分析,验证了该方法的有效性与高效性。

关键词:专利文本;冗余字符;附图标记;字频差;左切分词
中图法分类号　TH122;TP182

Recognition Method of Component Names in Patent Documents Based on the Algorithm of Word Frequency Difference and Library of Left-segmentation Words

KONG Jiabin,LYU Jianwen,LIU Jiangnan and DU Wenxuan
State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body,Changsha 410082,China

Abstract　Mechanical patent literature contains a large amount of domain knowledge where component names exist as information units. Being flexible and changeable, the word formatting of component name represents the characteristics of uniqueness, complexity and lesser-known expressions. The challenge of accurate recognition of component names by computers becomes an obstacle to patent knowledge mining. In order to propose an efficient method to recognize component names, the features of word formation in patent text statements are analyzed and extracted. Starting with external words related to component names, characters on the left side of the appended drawing reference signs(ADRS) are identified. Accordingly, candidate names are automatically retrieved from texts, and the set of candidate names are constructed. An algorithm of word frequency difference is proposed to filter redundant characters in the set of candidate names. By building left-segmentation library(LSL) dynamically, redundant characters which are not filtered are further eliminated. Based on cross-over experiment, the influence of character frequency difference prior threshold(CFDV-Ⅰ), word frequency threshold(LSWF) and character frequency difference threshold(CFDV-Ⅱ) on recognition result is tested and analyzed. Furthermore, a three-stage comprehensive method for recognizing component names from patent documents in mechanical field is proposed. Finally, the method has been proved to be effective and efficient by comparing the results of experiments.

Keywords　Patent text, Redundant characters, Appended drawing reference signs, Word frequency difference, Left-segmentation words

1　引言

机械领域专利文本中包含大量用于描述机械系统技术方案的组件单元,其名称属于自然语言的一种命名实体。专利中的组件名称多为合成词汇,其命名习惯因人而异,导致词组结构多变、术语复杂独特,难以被计算机准确识别。在专利知识挖掘过程中,命名实体名称的错误识别致使约 60%的文本分词错误[1-2],进一步导致语义关系的抽取错误。如何提高自动识别专利文本组件名称的效果,仍然是一个具有挑战性的问题[3-5]。

到稿日期:2022-05-07　返修日期:2022-10-23
基金项目:国家科技部创新方法专项资助项目(2019IM050100);湖南省自然科学基金(2018JJ2039)
This work was supported by the Innovation methods work Special Projects of Science and Technology of China(2019IM050100) and Natural Science Foundation of Hunan Province,China(2018JJ2039).
通信作者:刘江南(Liujiangnan@hnu.edu.cn)

命名实体的识别方法主要归纳为3种类型：基于规则的方法、基于统计的方法以及两种方法的组合[6-7]。目前，对于某些特定领域的命名实体识别，学者们多采用基于统计的方法。例如，在生物医学领域[8-10]、社交媒体领域[11-13]、化学药物领域[14-16]和军事领域[17-18]等，通过标注领域内的命名实体，并结合统计模型如条件随机场(Conditional Random Field，CRF)、长短期记忆网络(Long Short-Term Memory，LSTM)、支持向量机(Support Vector Machine，SVM)、注意力机制(Attention Mechanism)等对标注数据进行训练，得到适用于该领域的命名实体识别方法。如果特定领域文本的训练语料匮乏，那么命名实体的识别效果就会差强人意。

机械领域的专利文献中，为突出技术特征的创造性，组件实体的命名常有结合相关交叉学科的术语和短语等。针对如何识别这些内部构词规律复杂的实体名称，学者们从不同角度进行了探索。对于英文专利，Wang等[19]通过建立英文专利的命名实体词性标记规律和外在边界词性标记规律，利用非确定有限状态自动机建立所要表示短语的词性模式集合，通过模式匹配识别组件名称；Fantoni等[20]以数字字符为识别英文专利中实体词语的标志词，当数字字符后第一个单词为“of”时，将识别数字字符左侧的数个单词作为实体候选词，建立正则表达式识别命名实体；Alex等[21]使用数字附图标记获取候选实体，基于机器学习方法对候选实体进行分类，自动从专利文本中识别实体；Chen等[22]基于Bi LSTM-CRF模型，分别对专利文献中的17种实体类型的名称进行识别，结果显示零部件类型的实体名称识别效果相对突出。由于中英文语言的差异，上述方法难以在中文专利中复现。

针对机械中文专利文本的特征，学者们采用规则与统计相结合的方法对实体名称进行识别。Li等[23]通过对指定类型的命名实体词语进行词性标注，统计其内部词性规则，利用双向长短期记忆神经网络和条件随机场概率模型对机电产品9类命名实体进行了识别；Wang等[24]使用CRF构建了冶金领域字符角色标注的中文专利术语识别模型；Yu等[25]基于左侧通用词的分割，从专利文本中识别候选词，通过计算目标候选词的相似度对候选词进行排序进而识别候选词；Chen等[26]以专利实体新词内部的双向条件概率统计数据为基础，结合词边界规则实现专利实体新词的识别；Li等[27]通过对BERT-LSTM-CRF模型的改进实现了实体名称识别。机械中文专利文本具有一定的领域专业性和语言复杂性，缺少成熟的标注语料，而且在训练过程中难以进行可视观测以及故障排除[28]，仍存在不能有效识别领域实体名称的问题。

为此，本文聚焦组件名称在专利文本语句中的表达特征，从组件名称相关的外部用词着手，识别其右侧的附图标记，初步检索候选名称字符串，再过滤其中的冗余字符；进一步引入左切分词限制其左侧字符，优化组合识别过程中涉及的多个阈值，逐步收敛提高组件名称的识别精度。

2 专利文本中组件名称的表达特征分析

专利文献属于集技术、经济和法律于一体的半结构化特种文献，关于技术方案的具体描述，个体间存在较大差异，组件名称的构词形式尤为多样。因此，本文从专利文献撰写要求和具体文本着手，发掘隐藏在差异化组件名称中的表达特征。

2.1 现行专利法对专利文书中组件名称的撰写要求解读

根据中国现行专利法，对机械领域技术方案撰写相关的法规条文进行解读。分析《专利法实施细则(2010)》的第二章“专利的申请”、《专利审查指南(2010)》的第二章“说明书和权利要求书”等相关条款要求，结合具体的机械专利文本，归纳出组件名称的普遍特征如下。

特征1 同一组件名称至少两次出现在权利要求书和具体实施方式中。

特征2 具体实施方式中的组件名称必须附带附图标记，在权利要求书中可不附带。

特征3 组件名称的左侧多为相对固定的援引词，其规律更容易通过统计分析的方法得到。

2.2 专利语句中的组件名称构词特征分析

虽然《专利法实施细则(2010)》对专利说明书提出了章节结构、规范用语等基本要求，但是为了充分展现机械技术方案的新颖性，组件名称的命名依然复杂多样。基于组件名称普遍特征，进一步分析机械领域专利文本权利要求书和具体实施方式章节中的大量语句，对比复杂组件名称和语句的词性序列，提炼组件名称的书面表达特征。如图1的例句所示，采用JIEBA分词工具对组件名称“第一液压伸缩杆件”与干扰例句“所述多个轴承嵌入箱体中”进行分词并标注词性，其中“13”和“1”为附图标记。

名称分词		第一	液压	伸缩	杆件	13
词性序列		m	n	v	n	m
干扰示例	所述	多个	轴承	嵌入	箱体	1 中
词性序列	b	m	n	v	n	m f

图1 复杂组件名称与干扰语句的分词标注结果

Fig. 1 Segmentation results of complex component names and interference statements

分析发现，复杂组件名称的构词中，普遍存在并列短语、偏正短语和领域术语等功能性限定词汇，而且这些词汇的词性序列也出现在其他语句中。因此，通过大量构建实体内部词性序列的识别方法，其难度较大且容易引入无关信息。

相比组件名称内部的构词特点，其外部用词的潜在规律更容易被发掘。一方面，根据在具体实施方式文本中组件名称(System Component Name，SCN)与附图标记(Appended Drawing Reference Signs，ADRS)依次组合的书面描述特征，结合具体的专利文本内容分析发现，附图标记因紧随组件名称之后，且具有明显区别于汉字符号的特点，可作为组件名称识别的右边界词。另一方面，位于组件名称左侧的词语在专利文本中相比命名组件名称更为简单，如“所述”“包括”“有”“一个”等，亦可作为组件名称识别的重要边界词之一，类似地，将此类词语称为左切分词(Left-Segmentation Word，LSW)。

通过分析大量专利文献的具体实施方式文本还发现，组件名称和附图标记在专利语句中的表达形式主要有以下3种：

(1)组件名称后接西文括号内+附图标记；

(2)组件名称后接中文括号内+附图标记；

(3)组件名称后接附图标记。

根据上述规律，构建专利语句关于组件名称的表达特征图，如图 2 所示。可见，附图标记这一重要特征有助于在文本中标记定位组件名称。

形式①	...	OTW	SCN	(	ADRS	)	OTW	PUS
形式②	...	OTW	SCN	(	ADRS	)	OTW	PUS
形式③	...	OTW	SCN		ADRS		OTW	PUS

图 2 专利语句关于组件名称的表达特征

Fig. 2 Characteristics of component namesin statements

3 组件候选名称集合的构建与过滤冗余字符的字频差算法

根据图 2 所示的模型，利用附图标记 ADRS 作为组件名称的右边界，向左检索组件名称。组件名称一般在专利全文中多次出现，同一个附图标记定位在多个不同位置。由于缺少组件名称左侧切分词的限定，检索字符长度不同，将获得多个候选名称字符串。通过构建组件候选名称集合，再通过算法过滤组件候选名称的冗余字符，初步实现自动从文本中识别组件名称。

3.1 附图标记识别与组件候选名称集合构建

(1)附图标记的识别算法

根据组件名称信息的表达特征可知，附图标记在文本中紧随组件名称的右侧。通过统计和分析专利说明书的附图标记可知，附图标记主要由数字、字母、特殊符号组成，其构成形式有纯数字、数字与符号组合、数字与字母组合以及 3 种符号混合。

由于专利说明书中还存在对技术参数属性的描述，其类似的表达方式对附图标记的识别产生干扰。为避免仅采用正则表达式识别的信息冗余问题，提出了过滤非附图标记字符的补充算法。

Step1 识别专利具体实施方式文本中所有的非中文字符字段；

Step2 删除上述每个字段中最后一个字符为标点符号的符号，包括“，。：；、?!”；

Step3 删除英文字母个数超过 2 个的字符串；

Step4 记录字符串信息。

(2)组件候选名称集合的构建

针对相同的附图标记在专利文本中不同位置获得的多个候选名称字符串，引用集合的概念进行表达。将具有相同附图标记的所有候选名称字符串作为该附图标记的子集。候选名称字符串的每个文本符号都作为其子集。于是，构建组件候选名称集合($CSCG_i$)：

$$CSCG_i=\begin{bmatrix} ADRS_i:\{PNC_{i1},PNC_{i2},\cdots,PNC_{ij}\} \\ PNC_{i1}:\{\cdots,CH_{i13},CH_{i12},CH_{i11}\} \\ PNC_{i2}:\{\cdots,CH_{i23},CH_{i22},CH_{i21}\} \\ \cdots \\ PNC_{ij}:\{\cdots,CH_{ij3},CH_{ij2},CH_{ij1}\} \end{bmatrix} \tag{1}$$

其中，$ADRS_i$ 为附图标记；PNC_{ij} 为候选名称字符串；CH_{ijk} 为字符符号。

3.2 字频差计算引导的冗余字符过滤

借助附图标记识别的候选名称字符串中，掺杂着许多冗余字符。如图 3 所示的一个实例中，附图标记“32”对应的组件名称为“螺纹滑动轴”，集合中存在许多诸如“所述”“从而能够利用”等冗余字符。

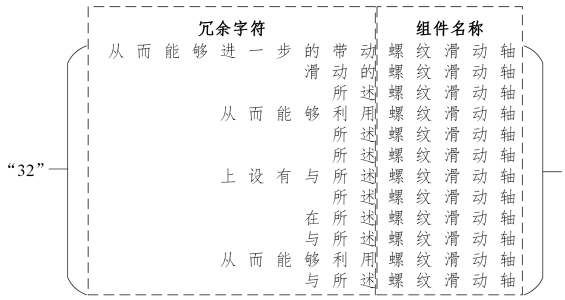

图 3 组件候选名称集合实例

Fig. 3 Instance of CSCG

分析组件候选名称集合，发现从右往左每一个字符所在的字符位置出现的次数并不一致。出现在组件名称的字符，在该名称位置上的频率明显高于出现在冗余字符当中的频率。为了直观体现这种差异，提出了式(2)、式(3)，计算候选名称字符串从右往左第 k 个字符位置上的字频差 WFm_k。

$$WF_{jk}=\frac{CH\_N_{jk}}{CH\_Z_k} \tag{2}$$

$$WFm_k=\begin{cases} WF_k\ \max-WF_k\ \min, & WF_k\ \max\neq 1 \\ WF_k\ \max, & WF_k\ \max=1 \end{cases} \tag{3}$$

其中，CH_N_{jk} 为第 k 个字符位置上某字符出现的次数；CH_Z_k 为该位置上的字符总数；WF_{jk} 为该位置上某字符的字频值。

基于字频差(WFm_k)的计算方法，提出过滤组件候选名称集合冗余字符的算法，具体如下：

(1)从右往左计算第 k 个字符位置上的字频差；

(2)当 WFm_k 为 1 时，保留该字符到原字符的位置；

(3)当 WFm_k 大于或等于某一阈值时，保留该字符位置上字频值最大的字符到原字符的位置；

(4)当 WFm_k 小于某一字频差阈值时，不保留该字符位置上的任何字符并停止遍历循环，写出已保留的字符，作为组件名称的识别结果。

上述算法有效剔除了存在于组件名称左侧的冗余字符，并且，字频差阈值的选取与冗余字符的剔除效果直接相关，组件名称的识别效果与字频差阈值之间的关系得以可视观测。

4 左切分词的获取及其词库动态构建方法

由于缺乏对组件名称左侧用词的限制，利用字频差过滤的组件名称仍可能存在冗余字符。为此，需要获取左切分词，并归纳左切分词的词频规律；提出动态构建左切分词库的方法，进一步在文本中检索候选名称字符串时限制其左侧的冗余字符，以提高组件名称的识别精度。

4.1 左切分词的获取与词频统计分析

通过字频差过滤组件候选名称集合冗余字符，可能仍有某些经常出现的词汇难以被过滤，如“所述”“并且”“包括”等。在某专利分析案例中，表 1 列出了正确识别组件名称的情况，表 2 列出了错误识别的情况。通过对比分析发现，在附图标记为“5”的集合中的冗余字符都不一样，通过字频差的计算可以轻易实现剔除；而在附图标记为“2”的集合中的冗余字符均为“所述”，难以有效通过字频差过滤。

表 1 正确识别的组件名称集合

Table 1 Sets of correctly identified component names

附图标记	识别结果	正确结果	组件候选名称集合
5	料液分离箱	料液分离箱	置于料液分离箱 和料液分离箱 积碳一起落入料液分离箱 磨料留在料液分离箱

表 2 错误识别的组件名称集合

Table 2 Sets of misidentified component names

附图标记	识别结果	正确结果	组件候选名称集合
2	所述弹性夹套	弹性夹套	所述弹性夹套 所述弹性夹套 所述弹性夹套 所述弹性夹套

在专利文本中，组件名称左侧的词汇多种多样，其中的一些词汇有可能作为某组件名称冗余字符，又同时存在于某组件名称中。在构建左切分词库时，应尽量避免采用可能存在组件名称中的词汇，并尽可能覆盖所有可能出现的左切分词。

为了获取并统计左切分词的词频规律，利用字频差过滤冗余字符方法，预识别组件候选名称(Candidate Name，CN)，在文本中获取其左侧的第一个词语作为左切分词。从 100 篇专利中获取组件候选名称左侧的第一个词汇共 3 759 个，统计这些词汇出现的频率结果如图 4 所示。

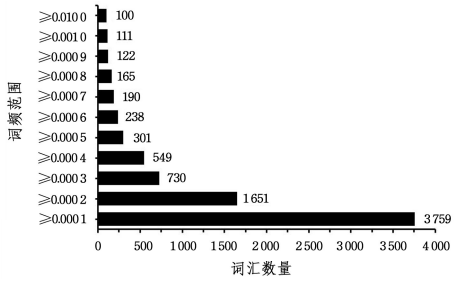

图 4 100 篇专利的左切分词统计数据图

Fig. 4 Statistics of 100 patents with left segmentation words

通过分析发现，词频越高的词汇，其词汇数量较少，词频最高的前 5 个词语分别为“和”“所述”“与”“有”“包括”。此类词语一般用于引导语句的展开，不随主题发生变化，较少出现在组件名称的构词组成部分。由此得知，采取上述方法获取的左切分词，其限制冗余字符的效果与词频具有较大的相关性。

4.2 组件名称识别过程的词库动态构建算法

左切分词属于多次重复出现的冗余字符，若能够在检索组件名称时提前识别并加以限制，将进一步提高组件名称的识别精度。因此，通过构建左切分词库的方法，利用字符匹配的方式对组件名称左侧的冗余字符进行限制。

为有效限制多样化专利文本中组件名称左侧的冗余字符，词库需要在识别组件名称时自动更新。在获取文献全文中所有的左切分词后，构造两个动态词库用于构建词库时实现迭代更新：

(1)初始左切分词库(Segmented-word Library，SL)，用于保存所有组件候选名称左侧的第一个词语；

(2)左分词词库(Left Segmented-word Library，LSL)，用于存储 SL 中词频大于词频阈值($LSWF$)的词语。

为实现左分词词库(LSL)的自动更新，叠加组件候选名称左侧的第一个词语到 SL 中；计算 SL 中所有词语词频后，将 LSL 中的所有词语替换为词频大于词频阈值($LSWF$)的词语。

5 组件名称识别三段式综合方法的阈值确定与实验验证

字频差算法为左切分词的获取提供支持，左切分词进一步提高了组件名称的识别精度，两者相辅相成。由于字频差阈值的选取直接影响组件候选名称集合中冗余字符的过滤效果，而词频阈值直接影响组件名称左侧冗余字符的限制效果，因此需要解决字频差算法与词库动态构建算法中的阈值组合问题，通过实验验证组件名称识别的有效性。

5.1 组件名称识别的三段式综合方法

将字频差过滤组件候选名称集合的算法和动态构建左切分词库的算法有机地结合在一起，形成精确识别组件名称的综合方法——三段式组件名称识别综合方法。其实施步骤分为 3 个过程，基于 Python 语言编写程序实现方法的全流程的程序框图如图 5 所示。

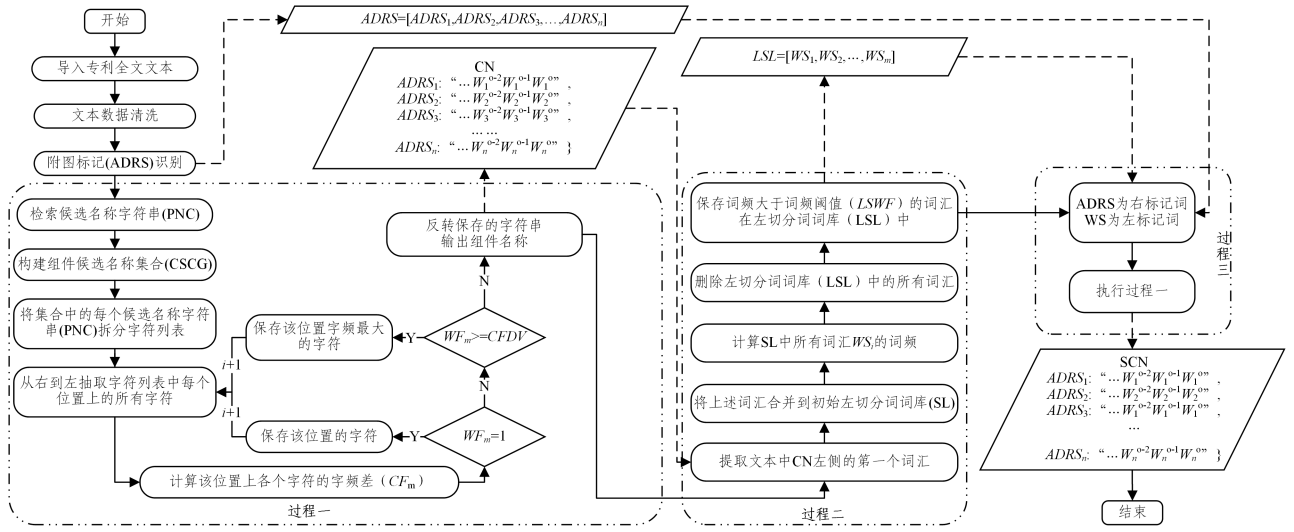

图 5 三段式组件名称识别综合方法编程实现程序框图

Fig. 5 Block diagram of three-stage synthesis method

过程一为组件候选名称 CN 的预识别。通过附图标记检索专利文本中的候选名称字符串，构建组件候选名称集合 $CSCG$，计算 $CSCG_i$ 中的 WFm，以字频差先验阈值（*CFDV*-I）过滤冗余字符，得到组件候选名称 CN。

过程二为左切分词库 LSL 的动态更新。在专利文本中识别组件候选名称 CN 左侧的第一个词语，叠加保存于 SL 中，计算其中所有词语的词频，选取词频大于词频阈值（LSWF）的词语保存到左切分词库 LSL 中。

过程三为组件名称 SCW 的精确识别。利用过程二构建的左切分词库和附图标记，分别作为检索候选名称字符串的左、右边界词，构建组件候选名称集合 $CSCG$，计算 $CSCG_i$ 中的 WFm，以字频差阈值（*CFDV*-II）过滤冗余字符，识别组件名称 SCW。

5.2　阈值选取对识别效果的影响及其优化组合

在基于字频差算法与切分词库构建的机械专利组件名称识别方法实施步骤中，字频差先验阈值（*CFDV*-I）、词频阈值（*LSWF*）和字频差阈值（*CFDV*-II）直接影响了组件名称的识别效果。为了最终确定阈值，对各阈值离散取值，分别进行实验，以精确率（P）、召回率（R）与调和平均数（F1）评判识别效果。

$$P=\frac{EC}{A}\times 100\% \tag{4}$$

$$R=\frac{EC}{S}\times 100\% \tag{5}$$

$$F1=\frac{2\times P\times R}{P+R}\times 100\% \tag{6}$$

其中，EC 为识别正确的个数；A 为识别得到的总个数；S 为标准数据的总个数。

由于字频差先验阈值所得到的组件候选名称直接用于构建左切分词库，因此需要首先确定其优化取值。而词频阈值和字频差阈值对识别效果产生耦合影响，通过交叉实验以确定两者的优化组合。

（1）字频差先验阈值（*CFDV*-I）的确定

逐一从 100 篇具有附图标记说明的机械领域专利文献中识别附图标记、获取组件候选名称集合，对字频差先验阈值（*CFDV*-I）从 0～1 间隔 0.05 离散取值，每一个取值为一组实验，共计 21 组。将每组实验识别的组件名称与原文进行对比，利用式（4）－式（6）计算各项指标，结果如图 6 所示。

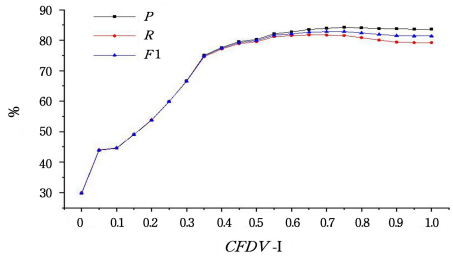

图 6　字频差先验阈值（*CFDV*-Ⅰ）对识别效果的影响

Fig. 6　Influence of *CFDV*-Ⅰ on recognition results

字频差先验阈值（*CFDV*-I）的选取会对组件名称的识别产生较大的影响。随着 *CFDV*-I 的提高，识别的精确率随之提高，但当 *CFDV*-I 介于 0.6～0.75 时精确率趋于平缓，随后略微有所下降。因此，确定字频差先验阈值（*CFDV*-I）为 0.75。

（2）词频阈值（*LSWF*）和字频差阈值（*CFDV*-II）优化组合的确定

逐一从 100 篇专利文本中识别附图标记，获取组件候选名称集合，用字频差先验阈值 0.75 识别组件候选名称，从文本中获取其左侧第一个词语保存至 SL 中并计算每个词语词频。对词频阈值（*LSWF*）从 0.0001～0.0015 中离散取值，构建 11 种不同的左切分词库（LSL）。利用每一组 LSL 中的词汇和附图标记作为左、右边界词，构建组件候选名称集合，再对字频差阈值（*CFDV*-II）在 0～1 范围内间隔 0.05 取值，每一个取值为一组实验，共计 231 组，识别效果如图 7 所示。当 $LSWF=0.0004$，*CFDV*-II＝0.5 时，*F*1 达 91.98%，相比实验 1 的优化结果提高了 9.18%。

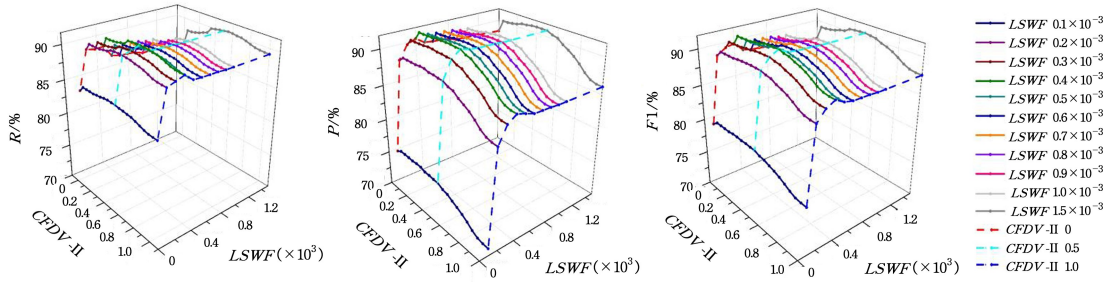

图 7　词频阈值和字频差阈值对识别效果的综合影响

Fig. 7　Influence of LSWF and CFDV-II on recognition results

5.3　组件名称识别方法的实验与结果分析

左切分词库根据词频阈值进行构建，词库中的词汇随着识别的专利数量增加发生变化，为了分析专利数量对组件名称识别效果的影响，通过设置识别不同数量的专利文本，以 *P*，*R*，*F*1 的变化来反映效果。另外，由于智慧芽专利检索网站[1)]推出的智能附图功能选取与本研究内容完全吻合，因此将其作为对比实验的对象之一，并且通过构建目前针对领域内实体名称识别较为常用的 LSTM＋CRF 模型，来进一步验证本文方法的优越性。

（1）实验设计

针对字频差算法与切分词库构建的机械专利组件名称识别方法，设计了两个实验对应两个实验目标。基于 Python 语言编写程序，从大为（INNOJOY）专利检索网站中获取具有附图标记的说明列表并且技术主题为设备、机构的机械领域专利

1) https://analytics.zhihuiya.com/search/input#/simple

n 篇，以方便人工获取这些专利中的组件名称并将其作为样本数据。组件名称识别验证实验的具体实施流程如图 8 所示。

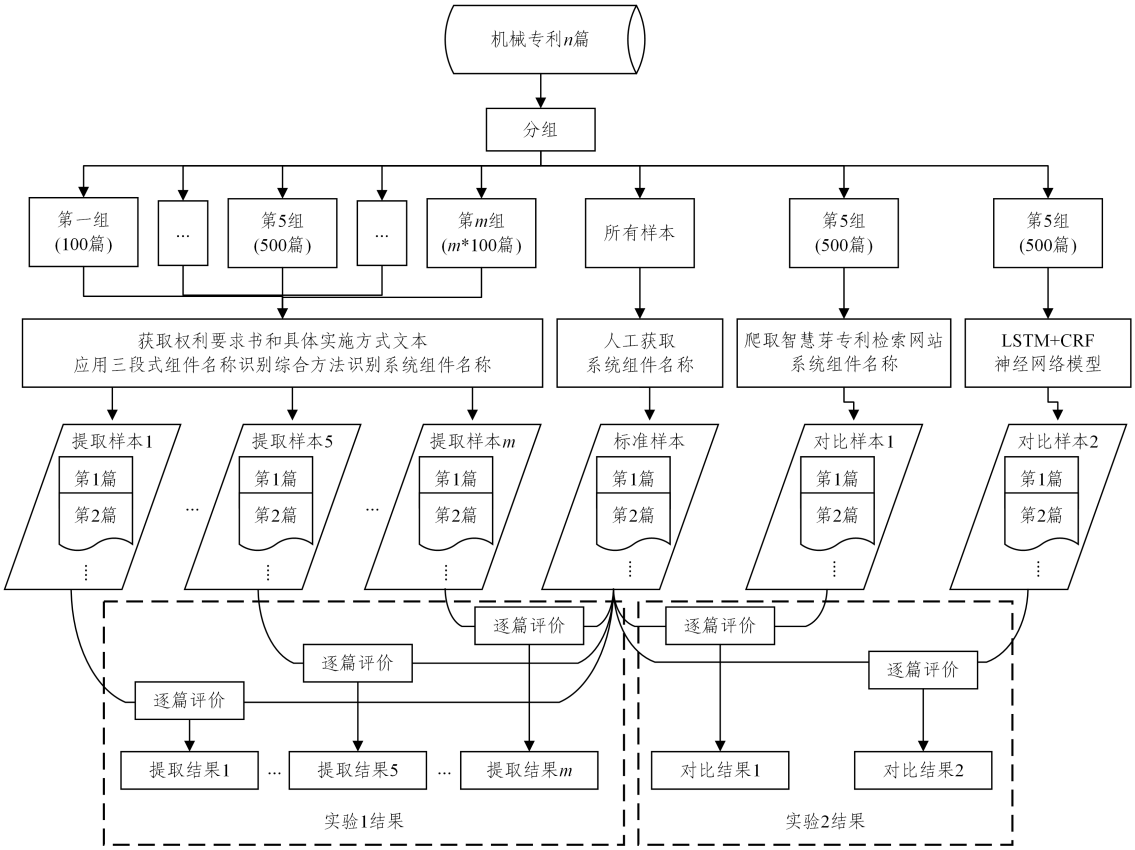

图 8 组件名称识别方法验证实验流程图

Fig. 8 Experimental flow chart of component name recognition method verification

实验 1 的具体步骤如下：

第 1 步 从 n 篇专利中随机选取 m 组专利文本数据，以 100 篇为起点依次增加 100 篇；

第 2 步 采用本文方法分别从 m 组专利的权利要求书和具体实施方式中识别组件名称，与人工获取的组件名称进行逐篇比较；

第 3 步 利用式(4)－式(6)计算各组的 P，R 和 $F1$。

实验 2 的具体步骤如下：

第 1 步 以实验 1 中 500 篇的识别样本的专利号检索、爬取智慧芽专利检索网站智慧附图中的所有组件名称，与人工获取的组件名称进行逐篇比较；

第 2 步 以目前针对领域内文本的命名实体识别方法[22,27]，通过构建 LSTM＋CRF 的神经网络模型，对实验 1 中 500 篇的识别样本进行组件名称识别，与人工获取的组件名称进行逐篇比较；

第 3 步 利用式(4)－式(6)分别计算第 1 步和第 2 步中的各组 P，R 和 $F1$。

(2)实验结果与分析

基于 Python 语言编写程序实现上述实验流程，最终获取了 2000 篇专利文献，设置 10 组实验样本。将实验 1 的各组结果绘制成图，如图 9 所示。当专利论文数量从 100 篇增加到 600 篇时，精确率 P、召回率 R 和调和平均数 $F1$ 都呈现上升趋势。在超过 600 篇的情况下，3 个评判数据都在 0.1% 范围内波动并保持稳定，其中 $F1$ 的最高值达到 94.1%。实验结果表明，随着实验专利数量的增加，各组的识别效果得到一定的提升。

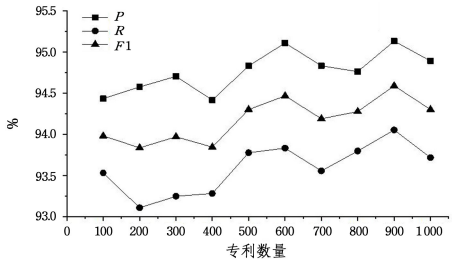

图 9 实验 1 的结果

Fig. 9 Result of experiment 1

经统计，500 篇专利文本的字符总数为 1853508 个，其中含有标准组件名称的共计 12498 个。500 篇专利文本中，采用本方法完全识别正确的专利共计 141 篇，精确率达到 90% 以上的共计 330 篇，总体精确率为 91.97%；而在智慧芽专利检索网站智能附图的爬取结果中，仅有 23 篇专利能够完全正确识别所有组件名称，精确率达 90% 以上的仅有 158 篇，总体精确率较本文方法低 14.69%。如表 3 所列，本文方法所识别的组件名称更为精准，各项评价均在 90% 以上，与智慧芽专利检索网站智能附图以及常规的 LSTM＋CRF 模型方法相比，其识别的结果优势明显。

表 3 实验 2 的结果

Table 3 Results of experiment 2

方法	精确率 P/%	召回率 R/%	调和平均数 $F1$/%	完全正确识别篇数 Rn/篇	精确率达 90% 篇数 Dn/篇	完全错误识别篇数 Wn/篇
三段式组件名称识别综合方法	**91.97**	**91.99**	**91.98**	**141**	**330**	**7**
智慧芽专利检索网站智能附图	77.28	65.84	77.10	23	158	43
LSTM+CRF 神经网络模型	50.01	73.35	59.47	7	84	0

识别结果显示，仍存在少量组件名称被错误识别的情况。为此，摘录识别错误的组件名称，获取其在识别过程中的组件名称集合，进一步探究其识别错误的原因。摘录部分识别错误的情况如表 4 所列，主要有 3 种错误类型。

(1)字符缺漏，“第一上连接轴 13”被错误识别为“上连接轴”、“进水用第一连接管 31”被错误识别为“第一连接管 31”；

(2)完全错误，“C 型梁 1”被错误识别为“内机架 1”；

(3)字符冗余，“插纸入口 17”被错误识别为“作用下被送往插纸入口 17”。

表 4 识别错误的组件名称集合摘录

Table 4 Excerpts from the collection of misidentified component names

标准名称	第一上连接轴 13	C 型梁 1	进水用第一连接管 31	插纸入口 17
识别结果	上连接轴 13	内机架 1	第一连接管 31	作用下被送往插纸入口 17
组件名称集合	13	1	31	17
	第一上连接轴	框体	第一连接管	作用下被送往插纸入口 17
	第一上连接轴	C 型梁	第一连接管	
	第一支链上连接轴	C 型梁	第一连接管	
	上连接轴	内机架	第一连接管	
		内机架	第一连接管	
		内机架		

通过分析表 4 可知，错误识别组件名称的原因主要有以下 3 点：

(1)专利文献撰写不规范。专利中部分同一实体的组件名称撰写不统一，有多字漏字现象；专利文献撰写过程中，多个不同的组件名称采用了相同的附图标记。

(2)左切分词库含有少量构成组件名称的用词。例如，左切分词“用”在组件名称“进水用第一连接管 31”中。

(3)组件在文本中只出现一次。

(4)组件在文本中多次出现，但该组件的左侧用词相同，且该词不属于左切分词。

结束语 针对机械领域专利文本中的组件名称自动化识别需求，提供了一种基于字频差算法与左切分词库构建的机械专利组件名称识别方法。

(1)通过分析国家专利法对专利文献的撰写要求，结合具体的专利书面文本，提炼专利文献中组件名称的构词特征，为自动识别工作奠定基础。

(2)通过附图标记检索候选名称字符串，构建了同一附图标记下的组件候选名称集合，提出了一种利用字频差过滤冗余字符的算法。

(3)通过左切分词的词频统计分析，提出了一种动态构建左切分词库的算法。

(4)探讨了字频差阈值和词频阈值对组件名称识别效果的耦合影响，通过交叉实验确定了两者的优化组合。

实验结果证明，与现有技术相比，组件名称的识别准确率优势明显，为自动挖掘机械领域专利知识提供了准确的组件名称信息。

参 考 文 献

[1] HE M, GONG C C, ZHANG H P, et al. Method of New Word Identification Based on Lager- scale Corpus[J]. Computer Engineering and Applications, 2007, 43(21): 157-159.

[2] ZHAO H, CAI D, HUANG C N, et al. Chinese Word Segmentation: Another Decade Review (2007-2017) [DB/OL]. https://arxiv.org/ftp/arxiv/papers/1901/1901.06079.pdf.

[3] LIU L, WANG D B. A Review on Named Entity Recognition [J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(3): 329-340.

[4] SUN Z, WANG H L. Overview on the Advance of the Research on Named Entity Recognition[J]. Data Analysis and Knowledge Discovery, 2010, 193(6): 42-47.

[5] CHEN Q Y, CHENG G, LI D, et al. Named Entity Recognition for Mechanical Design and Manufacturing Area[J]. Computer Engineering and Applications, 2017, 53(20): 100-104.

[6] VIKAS Y, STEVEN B. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models [C]// Proceedings of the 27th International Conference on Computational Linguistics. 2018: 2145-2158.

[7] PAN Z G. Research on the Recognition of Chinese Named Entity Based on Rulesand Statistics [J]. Information Science, 2012, 30(5): 708-712, 786.

[8] MAO X L, LI F F, WANG H T, et al. Named Entity Recognition of Electronic Medical Record Based on Improved HMM Algorithm[C]// 2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC). IEEE, 2017: 435-438.

[9] JU Z F, WANG J, ZHU F. Named Entity Recognition from Biomedical Text Using SVM[C]// 2011 5th International Conference on Bioinformatics and Biomedical Engineering. IEEE, 2011: 1-4.

[10] SUN A, YU Y X, LUO Y G, et al. Research on Feature Extrac-

tion Scheme of Chinese-character Granularity in Sequence Labeling Model--A Case Study About Clinical Named Entity Recognition of CCKS2017: Task2[J]. Library and Information Service, 2018, 62(11): 103-111.

[11] DONG C H, WU H J, ZHANG J J, et al. Multichannel LSTM-CRF for Named Entity Recognition in Chinese Social Media [C]// China National Conference on Chinese Computational Linguistics International Symposium on Natural Language Processing Based on Naturally Annotated Big Data. 2017: 197-208.

[12] LI Y, MA L, SHAO D G, et al. Chinese Named Entity Recognition for Social Media[J]. Journal of Chinese Information Processing, 2020, 34(8): 61-69.

[13] LI M Y, KONG F. Combined Self-Attention Mechanism for Named Entity Recognition in Social Media[J]. Journal of Tsinghua University(Science and Technology), 2019, 59(6): 461-467.

[14] BATISTA-NAVARRO R, RAK R, ANANIADOU S. Optimizing Chemical Named Entity Recognition with Pre-processing Analytics, Knowledge-Rich Features and Heuristics[J]. Journal of Cheminformatics, 2015, 7(Suppl 1): S6.

[15] YANG P, YANG Z H, LUO, et al. An Attention-Based Approach for Chemical Compound and Drug Named Entity Recognition[J]. Journal of Computer Research and Development, 2018, 55(7): 1548-1556.

[16] LI X, WEI X H, JIA L, et al. Recognition of Crops, Diseases and Pesticides Named Entities in Chinese Based on Conditional Random Fields[J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(S1): 178-185.

[17] FENG Y T, ZHANG H J, HAO W N. Named Entity Recognition for Military Text[J]. Computer Science, 2015, 42(7): 15-18, 47.

[18] SHAN Y D, WANG H J, WANG N. Military Domain Named Entity Recognition Based on Multi-label[J]. Computer Science, 2019, 46(S2): 9-12.

[19] WANG Z X, QIU Q Y, FENG P E, et al. Information Extraction Method of Technical Solution from Mechanical Product Patent [J]. Journal of Mechanical Engineering, 2009, 45(10): 198-206.

[20] FANTONI G, APREDA R, DELL'ORLETTA F, et al. Automatic Extraction of Function-Behaviour-State Information from Patents [J]. Advanced Engineering Informatics, 2013, 27(3): 317-334.

[21] ALEX J, HINRICH S, SOREN B. Unsupervised Training Set Generation for Automatic Acquisition of Technical Terminology in Patents [C]// Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Dublin, Ireland, 2014, Technical Papers. 2014: 290-300.

[22] CHEN L, XU S, ZHU L, et al. A deep Learning Based Method for Extracting Semantic Information from Patent Documents [J]. Scientometrics, 2020, 125: 289-312.

[23] LI S B, WU Y M, XU Y X, et al. A Bayesian Network Based Adaptability Design of Product Structures for Function Evolution [J]. Applied Sciences, 2018, 8(4): 493-509.

[24] WANG M P, WANG H, DENG S H, et al. Extracting Chinese Metallurgy Patent Terms with Conditional Random Fields[J]. Data Analysis and Knowledge Discovery, 2016, 271(6): 28-36.

[25] YU Y, ZHAO N X. Patent Term Extraction Based on Generic Words and Term Components[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(7): 742-752.

[26] CHEN M J, XIE Z P, CHEN X Q, et al. Novel Bidirectional Aggregation Degree Feature Extraction Method for Patent New Word Discovery[J]. Journal of Computer Applications, 2020, 40(3): 631-637.

[27] LI J, JING F Y, LIU J. Study on Patent Entity Extraction Based on Improved Bert Algorithms—A Case Study of Graphene[J]. Journal of University of Electronic Science and Technology of China, 2020, 49(6): 883-890.

[28] GEORGESCU T M, IANCU B, ZAMFIROIU A, et al. A Survey on Named Entity Recognition Solutions Applied for Cybersecurity-Related Text Processing[C]// Proceedings of Fifth International Congress on Information and Communication Technology, ICICT 2020, London, (Volume 2). 2020: 316-325.

KONG Jiabin, born in 1996, postgraduate. His main research interests include mechanical equipment innovation design and patent knowledge mining.

LIU Jiangnan, born in 1965, Ph.D, professor, master supervisor. Her main research interests include innovative design theory and methods, mechanical system optimization methods, patent avoidance and regeneration.

(责任编辑:喻藜)