情报工程 ISSN: 2095-915X
第1卷 第3期 81-89 2015年6月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN: 2095-915X Vol.1 No.3 81-89 Jun 2015

doi:10.3772/j.issn.2095-915x.2015.03.011

基于隐马尔科夫模型的专利功效词识别

张博培，杜永萍，马文建

（北京工业大学计算机学院，北京 100124）

摘要：随着专利数据规模的不断增长，对专利数据的深入挖掘也变得日益重要，特别是专利数据中所蕴含的技术功效等信息具有较高的价值。本文提出了一种基于隐马尔科夫模型的专利功效词识别方法，通过词法与句法分析筛选出候选功效词，在此基础上，采用隐马尔科夫模型并结合专利发明改进的特征设计了功效词识别算法，对候选功效词进行过滤。在新能源汽车等不同领域的专利数据集上，以准确率与召回率作为评价标准，验证所提出方法的有效性。实验结果表明，此方法有效提高了识别准确率与召回率。

关键词：专利数据，功效词识别，隐马尔科夫模型

Efficacy Word Recognition Method Based on the Hidden Markov Model

ZHANG Bopei, DU Yongping, MA Wenjian

(College Of Computer Science And Technology, Beijing University Of Technology, Beijing 100124)

Abstract: With the development of the patent data, the technique of patent data mining becomes more important, especially the technical efficiency information entailed in the patent data which have the higher value. We put forward the method to recognize the efficacy word based on the hidden Markov model. We select the candidate efficacy word by the use of lexical and syntactic analysis approach firstly. The recognition algorithm is designed by the combination of Hidden Markov model and features in the patent data. We give the experiment in different patent fields and the metric of precision and recall are used for the evaluation. The experimental result shows that our method gets the better performance.

Keywords: Patent Data, Efficacy Word Recognition, Hidden Markov Model

资助项目：国家科技支撑计划子课题（2013BAH21B02-01）；北京市自然科学基金资助项目 (4153058)；上海市智能信息处理重点实验室开放基金（IIPL-2014-004）

作者简介：张博培，（1990-），硕士，主要研究领域为自然语言处理、信息检索，email：bp.zhang@163.com；杜永萍，（1977-），博士，副教授，研究方向自然语言处理，email：ypdu@bjut.edu.cn ；马文建，（1990-），硕士，主要研究领域为信息检索、自然语言处理，email：mawenjian@gmail.com。

1. 引言

专利文献作为科技成果的载体与传统学术论文，著作等文献载体相比，拥有更多的知识技术含量。据世界知识产权组织统计，充分利用专利文献中的信息，可以大幅缩短研发时间，并减少研究经费[1]。目前企业通过专利数据挖掘可以对发现技术空白点，规避技术雷区有很大帮助，因此专利情报挖掘已经成为辅助企业技术创新的重要分析方法。

随着专利数据量的与日俱增，传统的对专利数据人工进行整理及挖掘的方法已经难以为继，迫切需要一种能够利用计算机自动或半自动的对专利数据进行挖掘的方法。构建专利技术功效矩阵是一种对专利文献数据进行信息提取的重要方式，该矩阵可以很直观的表现出某一技术主题领域下常用的技术手段及所达的功效。构建该矩阵主要需要解决三个问题，即技术词的识别与提取、功效词的识别与提取以及矩阵的构建。目前，在英文专利的技术功效矩阵研究方面 Tseng 等[2]将专利内容分为多个子部分，进一步判断技术词、功效词的分布，通过对技术、功效进行手动分类从而简单建立了技术功效矩阵，该方法工作量大，对于数据规模较大时不适用；Cheng[3]认为可以不通过专家进行技术功效矩阵的构建，然而实验中技术词的来源是通过国际专利分类号的分类来获得相应分类的技术词，因此，无法精确定位到在文本中出现的技术词，从而产生技术词的遗漏；翟东升等[4]提出了一种基于文本挖掘的技术功效提取方案，并由专家对提取结果进行评估，最终结果可以用来直接统计对应包含文献的数量，进而构造功效矩阵，然而并没有构建技术功效矩阵中最为关键的部分技术词与功效词的抽取结果统计；当前，对于功效词与技术词的提取方法，有许多是在参考了传统专利术语的提取方法的基础上进行了相应的修改后设计出来的。韩红旗等[5]提出了一种通过计算词语的术语度来评估候选术语的方法，该方法针对专利中缺少技术关键词的问题，在对主要术语抽取方法的基础上，修改了术语构词规则和术语度计算公式，该方法由于是人工对规则进行总结，因此具有很强的文本领域性；张锋[6]提出了一个中文术语自动抽取系统，该系统基于互信息计算字串内部结合强度，获得术语候选集，并利用搭配前缀、后缀信息进一步过滤，该方法对于文档集规模与质量具有一定的依赖性。目前，在对于英语专利的功效与技术抽取方面已经有了一些比较成熟的方法，国内在对专利技术功效矩阵的构建方面虽然有了一些尝试，但对于技术词与功效词的抽取还主要以人工为主，缺少系统、深入的研究[7]。

中文文本信息的识别与提取主要有基于规则的方法，基于统计的方法和二者相结合的方法[8]，本文提出一种将基于隐马尔科夫模型的中文文本信息识别与提取的方法用于专利文本功效信息的挖掘，系统地实现了一种有效的功效词识别方法，准确率和召回率与传统方法相比有了改进。本文将在第二节介绍隐马尔科夫模型，第三节阐述基于隐马尔科夫模型的功效词识别方法，第四节介绍实验结果以及分析，最后是结论。

2. 隐马尔科夫模型

应用隐马尔科夫模型，主要解决以下三个问题：评估问题，学习问题和解码问题[9]。其中，解码问题主要是指给定观测序列 $O=o_1o_2o_3\cdots o_n$ 以及模型，使得隐状态序列 S 是最具可能的，即求解最有可能的一个隐状态序列[10]。文本信息抽取需要解决隐马尔科夫模型中的解码问题，进行信息抽取时，首先训练样本，采用 Baum-Welch 算法进行学习，得出隐马尔科夫模型参数，然后采用维特比算法将待抽取的输入文本序列标记为最大概率的状态标签序列，状态标签序列中包括事

先定义的待抽取内容标签。

隐马尔科夫模型引入了两条独立性假设：

（1）状态 s_t 仅依赖于 st_{-1} 而与其前序状态 $s_1,s_2,\cdots,s_{t-2}$ 无关；

（2）每个观测值 o_t 仅依赖于其相应的状态 s_t。

一个隐马尔科夫模型由下列参数描述：

隐藏的状态集合：$S=\{s_1,s_2,\cdots,s_N\}$，N 表示所有可能的状态数，并记 t 时刻的状态为 s_t,$s_t \in S$；

观察符号集合 :$O=\{o_1,o_2,\cdots,o_M\}$，M 表示每一个状态对应的可能观察符号数；

状态转移概率矩阵 :$A=\{a_{ij}\}$，其中 $a_{ij}=p(s_{t+1}|s_t)$，$1<=i,j<=N$；

观察概率矩阵，即发射矩阵：$B=\{b_i(k)\}$，$1<=k<=M$，$1<=i<=N$，其中 $b_i(k)=p(o_t|s_t)$，o_t 表示在 t 时刻状态为 s_t 时的观察值；

初始状态分布 :$\pi=\{\pi_i\}$，$1<=i<=N$，其中 $\pi_i=P(q_1=s_i)$。

3. 基于隐马尔科夫模型的功效词识别方法

本文将在新能源汽车领域的专利数据集中进行功效词识别，为了提高识别精确度，通过 LDA 主题模型建模方式将数据集划分成不同技术主题领域的数据，该模型挖掘了潜在的语义信息，获取了主题分布向量。LDA 模型应用广泛，在此不再详述。

对于经 LDA 划分的不同技术主题数据，从中进行功效词提取。功效词抽取经过 3 个步骤：分词，词性标注，边界标注。本文利用 LTP 分词系统 [11] 对专利文本中的发明改进这一属性进行词语切分和词性标注，在此基础上观察发现有一部分功效词会被切分成几个孤立的词，举例如下：

提高 /v　了 /u　运输 /v　效率 /n

在上面的句子中，“运输效率”应该是需要识别出的功效词，但是分词系统将其划分成了 2 个孤立的词语。

耐久性 /n　优良 /a

在上面的句子中，待识别功效词为“耐久性”，其本身作为一个单独的词汇存在。因此，对于功效词的识别，本文将其划分为单词功效词与双词功效词。本文针对两种不同功效词分别设计了不同的识别与提取方法。

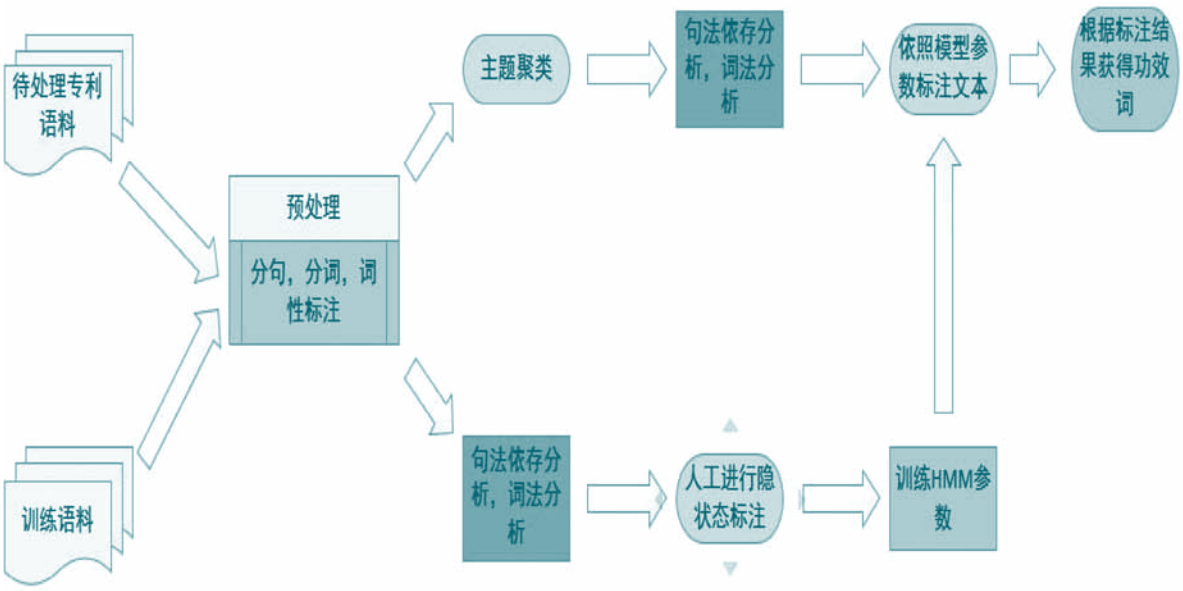

图 1 功效词识别模型

表 1 功效双词句法分析示例 1

序号	词语	词性	关系词语位置	关系
0	改善	v	-1	HED 核心关系
1	了	u	0	RAD 右附加关系
2	膜	n	3	ATT 定中关系
3	电极	n	6	ATT 定中关系
4	的	u	3	RAD 右附加关系
5	电化学	n	6	ATT 定中关系
6	性能	n	0	VOB 动宾关系

3.1 句法分析在功效词识别中的应用

针对双词功效词的识别与提取，由于在分词系统中其被分割成 2 个孤立的词语。通过观察发现，2 个独立的词语在句法上有一定必然联系，而通过 LTP 句法分析器对语句进行句法分析也验证了这一点。如：

改善 /v 了 /u 膜 /n 电极 /n 的 /u 电化学 /n 性能 /n

以上这句中“电化学性能”是要提取的功效词，而通过句法分析器对这条语句进行分析，发现“电化学”与“性能”之间构成一种定中关系，而在词性上是一种名词加名词的结构。如表 1 所示。

改善 /v 了 /u 温差 /n 热机 /n 的 /u 工作 / 性能 /n

表 2 功效双词句法分析示例 2

序号	词语	词性	关系词语位置	关系
0	改善	v	6	ATT 定中关系
1	了	u	0	RAD 右附加关系
2	温差	n	3	ATT 定中关系
3	热机	n	0	VOB 动宾关系
4	的	u	0	RAD 右附加关系
5	工作	v	6	ATT 定中关系
6	性能	n	-1	HED 核心关系

延长 /v 电池 /n 的 /u 使用 /v 寿命 /n

以上这句中“使用寿命”也是要提取的功效词，而“使用”与“寿命”之间构成了一种动宾关系，在词性上是动词加名词的结构。如表 3 所示。

表 3 功效双词句法分析示例 3

序号	词语	词性	关系词语位置	关系
0	延长	v	3	SBV 主谓关系
1	电池	n	0	VOB 动宾关系
2	的	u	0	RAD 右附加关系
3	使用	v	-1	HED 核心关系
4	寿命	n	3	VOB 动宾关系

本文提出了如下规则进行功效双词合并：

规则 1：如果物理位置相邻的 2 个独立词语之间构成定中关系，且词性搭配为动词加名词的形式或名词加名词的形式，则将其合并成一个独立双词。

规则 2：如果物理位置相邻的 2 个独立词语之间构成动宾关系，且词性搭配为动词加名词的形式，则同样将 2 个词汇合并成一个独立的双词。

3.2 基于隐马尔科夫模型的功效词识别

在利用句法分析与词法规则匹配将双词合并后，获得了候选功效词集，在这一集合中仍然包含许多噪音词汇，比如在例句“改善了膜电极的电化学性能”中，“膜”与“电极”之间构成定中关系，且词性搭配为名词加名词的形式，符合规则，但“膜电极”并不是要提取的功效词。将合并的单一词汇看成一个整体，从句法结构上看，中文语句中的组织和构成都符合以概率的形式存在的语法规则。由于语言的随机性，这样的语法规则很难通过穷举的方式来形成规则库，而隐马尔科夫模型可以很好的解决上述问题，通过对训练语料的学习可以获取相应语法规则库，从而支持对于后续数据的处理。

3.2.1 参数训练

系统中模型参数训练需要经过以下四个步骤：词语切分；词性标注；双词合并；词性加工标注。本文利用 LTP 语言技术平台语料进行词语切分和词性标注，通过基于规则方法的过滤，将可能的双词功效词进行合并，最后通过人工检查的方法，对经过合并后的可能的功效词进行标注。模型中的观察值集合为 LTP 的各种可能的词性标注结果加上一个特殊词性“hb”，其代表经过合并后的词汇。即 V={‘n’,’v’,’a’,…,’d’,’hb’}。模型中隐状态集合为上述观察状态集合加上一个特殊词性“gx”。即 S={‘n’,’v’,’a’,…,’hb’,’gx’}。在学习过程中，本文参考了 Baum-Welch 算法。具体如下：

$$\pi_t=\frac{\text{句子首字被标成}St\text{的频次}}{\text{训练语料中句子数量}}\quad 1\le t\le N\ (1)$$

$$a_{ij}=\frac{\text{状态转移从}St\text{到}St+1\text{的频次}}{\text{状态转移从}St\text{出现的频次}}\quad 1\le t\le N\ (2)$$

$$b_{i(k)}=\frac{\text{状态为}St\text{时观察值为}Ot\text{的频次}}{\text{状态处于}St\text{的频次}}\quad 1\le t\le N\ (3)$$

通过 LTP 分词和词性标注获得训练样本的观察状态，由人工标注的结果，可以获得训练样本的隐含状态，利用以上三式，获得隐马尔科夫模型的三个参数矩阵，即初始状态概率矩阵 π_i，状态转移概率矩阵 a_{ij}，状态发射概率矩阵 $b_{i(k)}$。

3.2.2 词性标注

对于词性标注问题，设 $O=o_1o_2\cdots o_M$ 是一个句子，其中 oi 是组成的词，S=s1s2…sT 为相应的词性标记序列，λ为词性标注所需的隐马尔科夫模型参数，P(O,S| λ) 表示句子 O 和其词性标记序列 S 正确的概率，词性标注的实质是求解 argmaxP(O,S| λ)，上述问题可以通过维特比 (Viterbi) 算法进行求解：

表 4 利用维特比算法实现功效词自动标注

算法 1 利用维特比算法实现功效词自动标注
输入 : 初始状态概率矩阵π_i，状态转移概率矩阵 a_{ij}，状态发射概率矩阵 $b_{i(k)}$，即模型参数，发明改进数据 输出 : 对发明改进数据的标注结果 1. 根据初始概率矩阵 π_i 以及首词词性利用公式计算 $\delta_1(i)=\pi_i b_i(o_t)$，其中 $\delta_1(i)$ 为首词隐状态。 2. 进行递推计算，根据前一状态的各概率隐状态，状态转移概率矩阵 a_{ij} 以及当前显状态，利用公式 $\delta t(j)=[\max\delta t-1(i)aij]bj(ot)$，$\varphi_t=\text{argmax}(\delta_{t-1}(i)a_{ij})$，$2\le t\le T$；$1\le j\le N$ 计算当前最有可能的隐状态。 3. 计算至最后一个词语时，计算得到路径最大概率，$P=\max\delta_T(i)$，$q_r=\text{argmax}(\delta_T(i))$。 路径回溯，从最后一个词语的状态向前获得所有词语状态 $q_t=\varphi_{t+1}(q_{t+1})$，从而获得句子中所有词语的词性标注序列。

利用维特比算法计算得到最有可能的隐状态序列，实现对语料的自动标注，从而对双词功效

词进行识别与提取。

在利用句法分析与词法规则匹配将双词合并后，从句法上分析，发现合并后的词汇和单词功效词在句法结构上趋于一致，因此对于单词功效词的识别与提取，同样可以利用上述基于隐马尔科夫的方法。在构建训练集时，对经过合并双词之后的语料进行双词功效词和单词功效词的人工标注，可以在构建模型之后同时对单词功效词与双词功效词进行识别与提取，识别算法如表 5。

表 5 功效词识别算法

算法 2 基于隐马尔科夫模型与规则的功效词提取算法
输入：专利发明改进语料，训练语料，聚类簇数目
输出：对发明改进语料的功效词标注结果
1. 对发明改进语料和训练语料进行分句，分词，词性标注。
2. 对经过预处理的发明改进语料利用 LDA 模型进行主题聚类。
3. 对经过预处理的训练语料和聚类后的发明改进语料利用句法分析规则 1,2 进行句法和词法分析。
4. 对经过处理的训练语料进行人工词性标注，并利用 Baum-Welch 算法计算模型参数，其中模型参数包括：状态转移概率矩阵 $A=\{a_{ij}\}$，观察概率矩阵 $B=\{b_i(k)\}$，初始状态分布：$\pi=\{\pi_i\}$。
5. 根据模型参数利用维特比算法对发明改进语料进行自动标注，并获得功效词。

4. 实验结果与分析

4.1 数据集

本文实验集选取了 3000 条具有代表性的新能源汽车领域专利文本，采用 3000 条专利数据的发明改进这一属性，经过 LDA 进行文本聚类之后选择合适的技术主题文档集共 813 条专利发明改进数据，对其进行分句之后，得到 2456 条句子，对其进行功效词识别与提取。

4.2 评价标准

对于本实验，采用准确率与召回率作为评价指标，其计算公式如下：

$$准确率=\frac{系统正确识别的功效词数量}{系统识别出的功效词数量} \quad (4)$$

$$召回率=\frac{系统正确识别的功效词数量}{文本中功效词数量} \quad (5)$$

4.3 与传统识别功效词方法的对比

传统方法通过维护一个线索词表来实现功效词的提取，其中，线索词为与功效词共现的词语，一般以动词的形式出现，比如提高，改进等词语。在句子中查找线索词，并对句子进行分析，对于线索词是动词的情况，与线索词构成动宾关系的词汇一般认为是功效词，但是由于分词系统会将一个功效词分成 2 个或多个独立的词，因此此方法对于双词功效词很难进行完整识别，而且在进行句法分析时，也会伴随有错误的现象，其抽取的准确率也无法保证。对于线索词是形容词的情况，一般认为与线索词物理位置相邻的名词是功效词，但在线索词与功效词物理位置不相邻的情况以及有多个并列功效词的情况，效果并不理想。当然，传统的方法也极度依赖于所维护的线索词表中线索词的全面性。

本实验对功效词进行分类，并根据不同功效词的类别分别使用不同的方法进行识别与提取。与传统的基于线索词的方法使用同样的实验集提取结果相比，从实验结果上看有了很大的改进，如表 6。

表 6 隐马尔科夫模型标注结果与传统方法标注结果比较

	基于隐马尔科夫的方法		基于线索词的方法	
	准确率	召回率	准确率	召回率
单词功效词	0.76531	0.83908	0.70631	0.80715
双词功效词	0.73636	0.76064	/	/
功效词	0.75084	0.80986	/	/

表 7 所示为功效词抽取的部分结果。

表 7 功效词示例

功效词	共现词	语句
使用寿命	延长	延长电池的使用寿命
放电功率	增加	电池的放电功率增加 14-20%
应用范围	扩大	扩大了自持式水下刨面浮标的应用范围
生产工艺	简单	且生产工艺简单
生产成本	降低	从而大幅度降低生产成本
转换效率	提高	提高能量的转换效率

4.4 句法与词法分析对功效词识别结果的影响

本文对于双词功效词的识别分为两个环节，通过句法词法分析实现双词功效词中孤立词的合并，并利用隐马尔科夫模型进行过滤。为了验证句法与词法分析的必要性，我们同时直接使用隐马尔科夫模型对功效词进行识别，对比结果如表 8。

表 8 句法与词法分析对识别性能的影响

	使用句法与词法分析的方法		只使用隐马尔科夫模型的方法	
	准确率	召回率	准确率	召回率
单词功效词	0.76531	0.83908	0.76531	0.83908
双词功效词	0.73636	0.76064	0.66325	0.68934
功效词	0.75084	0.80986	0.71428	0.76421

从上述结果中看出，使用句法与词法分析对于双词功效词的识别有很大的促进作用。这主要是由于使用句法与词法分析大大降低了双词功效词的识别范围，而在利用隐马尔科夫模型计算时，只有经过合并的词语才有可能被标注为功效词。

4.5 不同领域抽取结果

为了验证所提出的方法对不同领域数据的识别有效性，本实验选择了药品、计算机、半导体等几个领域有代表性的数据分别利用上述方法进行功效词的识别，实验结果如图 2 和图 3。

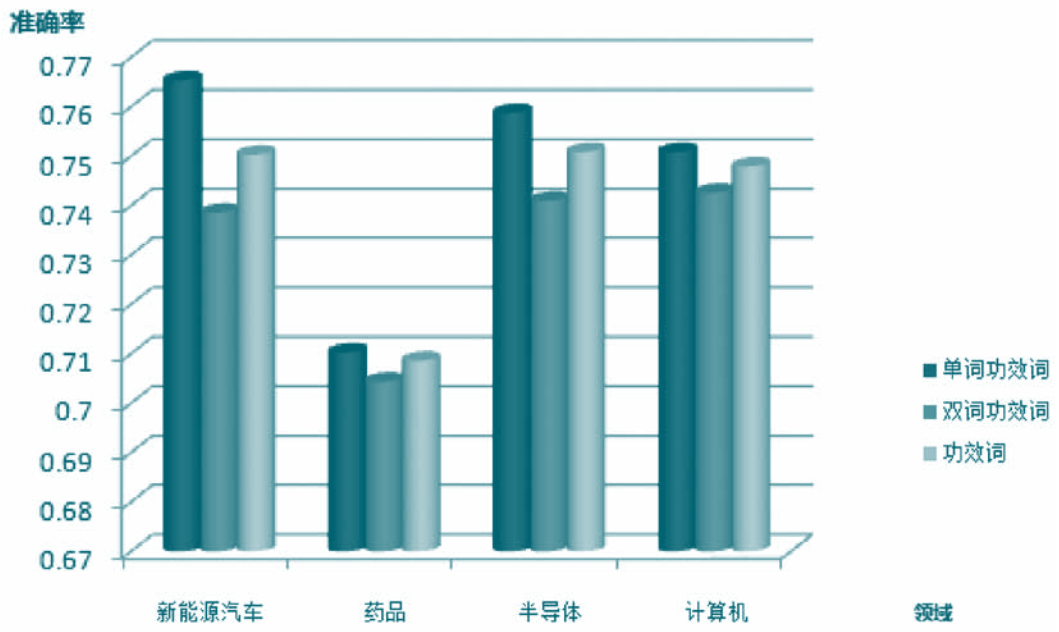

图 2 不同领域功效词标注准确率

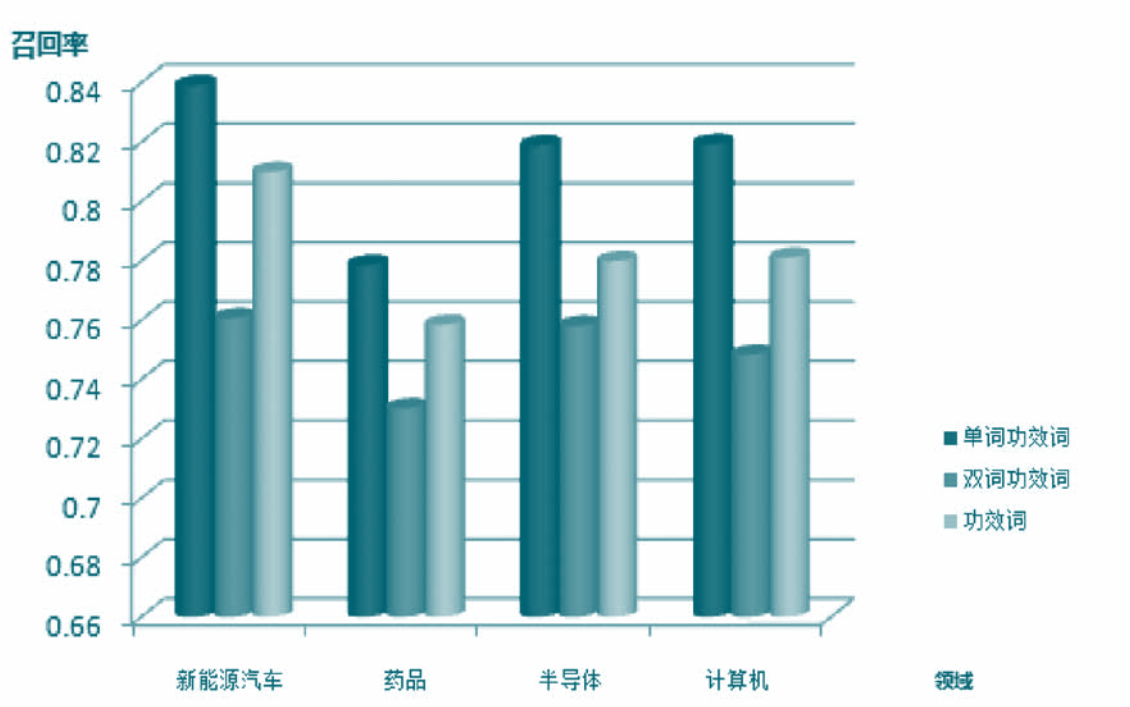

图 3 不同领域功效词标注召回率

由图 2 和图 3 可以看出本方法对不同领域的功效词识别均有效。在词法与句法分析中所提取的规则主要是一些各领域文本功效词的通用规则，因此，在双词功效词进行合并时，所合并的词语对于双词功效词召回率较高，而在利用隐马尔科夫模型时，由于汉语在表达发明改进信息时的一般规律，所以最终在各领域提取的功效词达到了预期的效果。由于训练集采用的是新能源汽车领域的数据集，所以，对于新能源汽车领域的功效词识别提取结果最优。

5. 结论

本文使用基于句法分析与词法规则的方法对候选功效词进行识别，在此基础上利用隐马尔科夫模型对功效词进行提取，实验结果表明，与基于线索词的传统方法相比，本方法有效提高了双词功效词识别的召回率与准确率，因此对于功效词的提取效果具有显著地改进效果，并且对于不同领域的功效词识别具有一定的通用性。

在实验中我们发现本系统对于多个功效词以并列关系出现时的识别效果也不理想，后续工作将着重对这两方面进行改进。另外，在实现了对于功效词的提取的同时，尝试对技术词进行识别与提取，并最终构建技术功效矩阵，该矩阵可以应用于专利预警等方面，值得进一步研究。

参考文献

[1] 陈燕，黄迎燕，万建国 . 专利信息采集与分析 [M]. 北京：清华大学出版社，2006:19-24.

[2] Tseng T H,Wang Y M,Juang D W,et al. Text Mining for Patent Map Analysis [C]//Proceedings of IACIS Pacific 2005 Conference,Taipei,2005.

[3] Cheng T V. A New Method of Creating Technology/Function Matrix for Systematic

Innovation without expert [J].Journal of Technology Management & Innovation, 2012,7(1):18-27.

[4] 翟东升，陈晨，张杰等 . 专利信息的技术功效与应用图挖掘研究 [J]. 情报分析与研究 , 2012(7/8):96-102.

[5] 韩红旗，朱东华，汪雪锋 . 专利技术术语的抽取方法 [J]. 情报学报 , 2011(12).

[6] 张锋，许云，侯艳等 . 基于互信息的中文术语抽取系统 [J]. 计算机应用研究 , 2005(5):72-77.

[7] 陈颖，张晓林 . 专利中技术词和功效词识别方法研究 [J]. 知识组织与知识管理 , 2011(12):24-30.

[8] 张晓艳，王挺，陈火旺 . 命名实体识别研究 [J]. 计算机科学 , 2005(4):44-48.

[9] 岑咏华，韩哲，季培培 . 基于隐马尔科夫模型的中文术语识别与研究 [J]. 情报分析与研究 , 2008(12):54-58.

[10] 林亚平，刘云中，周顺先等 . 基于最大熵的隐马尔科夫模型文本信息抽取 [J]. 电子学报 ,2005(2):236-240.

[11] LTP 语言技术平台 [EB/OL][2015-07-20]. http://www.ltp-cloud.com/.